BETTER PREDICTION CAN MAKE A WORSE ALARM: AUDITING PREDICTIVE SELECTION IN LLM-AGENT SAFETY MONITORING

Anonymous authors
Paper under double-blind review

ABSTRACT

Runtime monitors support the safe deployment of LLM agents by warning of impending failures under strict false-alarm budgets. Candidate monitors are commonly benchmarked and selected using offline predictive metrics, but deployment evaluates them through sequential warning policies that turn predictions into alarms. This practice assumes that a monitor that predicts failures better offline will also make the better runtime alarm. However, whether that assumption holds remains largely unexamined.

We formalize this question as *predictive-to-policy ranking transfer*, which measures whether predictive preferences between monitors are preserved under a specified warning rule and false-alarm budget. We audit ranking transfer for five monitors across three warning rules and four false-alarm budgets, with extensions across candidate families, agent models and environments. We find that a consistently better offline predictor can become the worse alarm, even when both rankings are highly stable. These reversals can be costly: in one low-false-alarm setting, stable reversals incur a median 30.4% relative loss in held-out recall. In contrast, selecting monitors by warning performance improves held-out recall by 2.2–5.8 percentage points while reducing achieved false-positive rates.

We further show that this mismatch is structural. Standard predictive metrics summarize prediction quality across observations, while warning performance also depends on how scores behave over an entire execution. Reordering scores within an execution can preserve pointwise predictive summaries while substantially altering warning performance under rules that depend on temporal order. Thus, predictive metrics alone cannot establish which monitor makes the better alarm; ranking-transfer analysis makes the prediction-to-decision handoff explicit and tests whether offline comparative conclusions survive sequential realization.

1 INTRODUCTION

Runtime monitors support safe LLM-agent deployment by detecting impending failure or unsafe behavior before an execution is complete, enabling timely warnings, execution stops, or other interventions. During an execution, a monitor updates its risk score as each new step becomes available. A warning policy processes this sequence of estimates to determine whether and when to issue a warning. The resulting warning outcome is therefore determined over an execution, rather than by evaluating each prediction in isolation. The policy must identify failures in time for intervention while controlling false alarms at the execution level. Forecasting from partial executions has a long history in predictive process monitoring, where incomplete traces are used to anticipate future events or outcomes (Tax et al., 2017; Teinemaa et al., 2019; Ceravolo et al., 2024). Recent LLM-agent monitors extend this idea to long interactive trajectories, predicting failure, task success, or safety risk from the execution observed so far (Wang et al., 2026; Li et al., 2026a; Pham et al., 2026; Huang et al., 2026; Baidya et al., 2026; Ruan et al., 2026; Lin et al., 2026; Zhang et al., 2026).

Predictive performance is an established basis for comparing monitors (Teinemaa et al., 2019), and recent LLM-agent work reports predictive metrics alongside warning performance (Huang et al., 2026). Offline predictive metrics offer a common basis for comparing candidates, including when

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.

the warning rule and false-alarm budget have yet to be specified. This suggests a natural selection principle: monitor a's score predicts failure better than monitor b, a should also be preferred for warning. Yet deployment introduces an additional step: each monitor's scores must be converted into alerts using a warning rule, such as triggering an alarm when a score first crosses a threshold, under a specified false-alarm budget, such as a target rate of 2% of non-failure executions (Teinemaa et al., 2018; Kubrak et al., 2022; Fahrenkrog-Petersen et al., 2022). Whether the predictive preference between a and b is preserved under this sequential warning process remains unclear.

We study this gap through *predictive-to-policy ranking transfer*: whether preferences established by predictive evaluation are preserved when the same candidates are compared by warning performance under a specified warning rule and operating point. For a set of candidate monitors, we quantify transfer as the fraction of candidate pairs for which predictive and warning evaluations prefer the same monitor. Our ranking-transfer audit compares these preferences across warning policies, assesses their stability separately, and evaluates the held-out consequences when predictive and policy-based selections disagree.

We audit five closely matched monitors under three warning rules and four false-positive-rate (FPR) budgets, with candidate-specific thresholds calibrated to the same target. We assess stability by grouped resampling, evaluate selection on unseen groups, and extend the analysis across mixed monitor constructions, a recursive warning rule, additional agent models, and environments.

Our central finding is that *better prediction can make a worse alarm*. Predictive and policy rankings can diverge even when the predictive preference is strong and reproducible. In the most extreme condition, over 90% of the strongest predictive preferences disagree with the policy ranking, including stable reversals. These disagreements are consequential: in one low-false-alarm setting, every retained stable reversal incurs held-out recall loss, with a median relative loss of 30.4%; predictive-only selection can also reduce deployment utility.

We further ask why such reversals are possible. Pointwise predictive metrics summarize score–label relationships, whereas sequential warning depends on additional properties of the score process. First crossing depends on episode-level score extremes relative to calibrated boundaries, while order-sensitive rules also depend on temporal organization. Controlled rearrangements change warning performance while preserving pointwise predictive summaries, establishing that predictive quality alone does not determine downstream warning performance.

More broadly, these findings have implications for both monitor evaluation and monitor design. Predictive benchmarks establish forecasting quality, whereas deployment depends on how a monitor's score process interacts with a warning rule at a given operating point. Evaluation should therefore make this prediction-to-decision handoff explicit and assess the coupled monitor–policy system. For monitor design, this suggests that the objective is not only to produce accurate risk estimates, but also to produce score processes whose deployment-relevant behavior is compatible with the intended warning policy.

Our contributions are threefold:

- **Empirical characterization of predictive–policy reversals.** We show that predictive and policy rankings can diverge sharply across warning policies and operating points, including stable reversals that lead to substantial held-out selection consequences.
- **Predictive-to-policy ranking-transfer analysis.** We formalize ranking transfer and develop an evaluation procedure for quantifying whether predictive preferences remain supported after sequential warning realization, including stability and held-out analyses.
- **Structural insufficiency of predictive metrics.** We establish that standard pointwise predictive summaries do not determine sequential warning performance. First crossing depends on episode-level score extremes relative to calibrated boundaries, while order-sensitive rules additionally depend on temporal structure.

2 Related Work

Predictive and prescriptive runtime monitoring. Predicting outcomes from incomplete executions is established in predictive process monitoring and longitudinal prediction (Tax et al., 2017; Teinemaa

et al., 2019; Lee et al., 2018; 2020; Ceravolo et al., 2024). Predictive metrics assess forecasting quality, whereas alarm-based and prescriptive monitoring study when to act under timing, cost, and resource constraints (Gneiting & Raftery, 2007; Teinemaa et al., 2018; Fahrenkrog-Petersen et al., 2022; Shoush & Dumas, 2022). Recent LLM-agent monitoring work similarly studies early prediction of failure or safety risk to support warning, stopping, or intervention (Wang et al., 2026; Li et al., 2026a; Pham et al., 2026; Huang et al., 2026; Baidya et al., 2026; Ruan et al., 2026; Lin et al., 2026; Zhang et al., 2026).

Predictive evaluation and downstream decision-making. Decision-focused learning incorporates downstream objectives into model training (Elmachtoub & Grigas, 2022; Mandi et al., 2022), while evaluation alignment studies whether predictive evaluation reflects downstream value (Shahroudi et al., 2025). In LLM-agent monitoring, PrefixGuard explicitly reports that strong predictive ranking need not translate into effective low-false-alarm warnings (Huang et al., 2026). These lines of work connect prediction with downstream use, but leave the comparative selection handoff insufficiently characterized: whether an ordering induced by predictive evaluation is preserved after candidate monitors are realized through a specified warning rule and operating constraint, and whether disagreement reflects instability or stable reversal.

Sequential decision rules and score-process structure. Sequential testing, evidence accumulation, and change-point detection operate on ordered observations (Wald, 1947; Page, 1954; Adams & MacKay, 2007; Tartakovsky et al., 2014). Early-classification and prescriptive-monitoring methods further study stopping and intervention decisions over evolving observations (Schäfer & Leser, 2020; Mori et al., 2018; Shoush & Dumas, 2022; Fahrenkrog-Petersen et al., 2022). These works establish the importance of sequential structure, while leaving open whether pointwise predictive comparisons retain sufficient information to preserve candidate preferences after scores are consumed by sequential warning policies.

3 A Ranking-Transfer Audit for Monitor Selection

To assess whether predictive preferences remain valid as selection evidence under deployment warning policies, we introduce a ranking-transfer audit for monitor selection. The audit makes the correspondence between predictive and policy-induced preferences explicit and quantifies the extent to which predictive preferences are preserved after policy realization. Given a set of candidate monitors, it first constructs predictive and policy rankings and measures their agreement. It then assesses the stability of the resulting preferences and evaluates whether development-stage selection evidence remains supported on unseen executions. Figure 1 provides an overview of this audit.

Predictive selection preferences. For candidate monitor j, let $s_{j,e,t}$ denote its risk score at step t of execution e, computed only from information available up to that step. All candidates are evaluated against the same prespecified future-event query. A predictive evaluation functional assigns each candidate a score $M_{\mathrm{pred}}(j)$, which induces a predictive ranking and corresponding pairwise preferences. For candidates a and b, this preference represents the comparative selection evidence supplied by predictive evaluation. The specific predictive metric and event query are part of the audit specification rather than properties of the framework itself.

Policy selection preferences. A warning policy $c = (\pi, \alpha)$ pairs a warning rule π with an episode-level FPR target α, which defines the policy's operating point. Each candidate is calibrated separately to the same target, after which a policy-specific evaluation functional $M_c(j)$ induces the corresponding policy ranking. Predictive and policy evaluation therefore compare the same candidates and target, but through different evaluation functionals.

Predictive-to-policy ranking transfer. For comparison instance r and candidates a, b, let $w^r_{\mathrm{pred}}(a, b)$ and $w^r_c(a, b)$ denote the predictive and policy winners, respectively. Over a prespecified comparison set S_c,

$$\widehat{\mathcal{I}}_c = \frac{1}{|S_c|} \sum_{(r,a,b)\in S_c} \mathbf{1}\left[w^r_{\mathrm{pred}}(a, b) = w^r_c(a, b)\right].$$

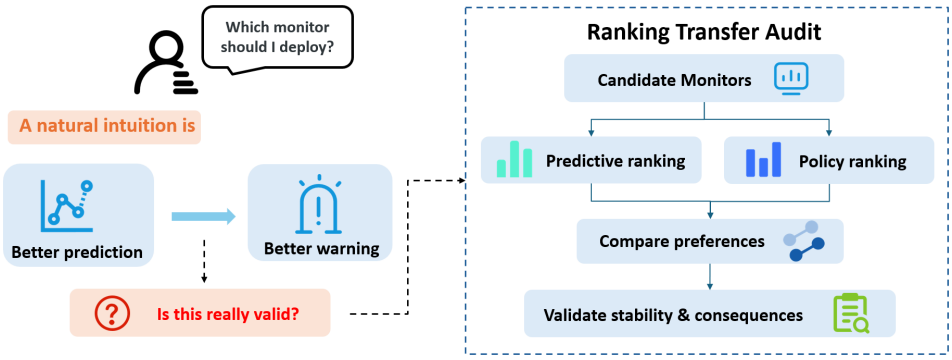


Figure 1: **From predictive intuition to ranking-transfer audit.** Better prediction is often assumed to imply better warning, but this assumption requires validation. Our audit compares predictive and warning rankings, then examines their agreement, stability, and held-out consequences.

We refer to this quantity as *predictive-to-policy ranking transfer*. It quantifies how often predictive preferences remain supported under the specified warning policy. The audit can be applied to the full candidate family or to prespecified subsets of comparisons.

Validation and selection consequences. Observed agreement alone does not distinguish a systematic transfer pattern from uncertainty in either ranking. We therefore validate predictive and policy preferences in two ways. First, grouped resampling at the natural dependence unit assesses the stability of each preference. Second, development-stage selection decisions, calibrated policy parameters, and candidate orientations are frozen before evaluation on unseen groups, testing whether the resulting selection evidence remains supported under the downstream policy. We then quantify the held-out consequences when predictive and policy-based selections disagree.

4 Empirical Evaluation of Ranking Transfer

4.1 Experimental Setup

Benchmarks and agent models. Our primary evaluation uses Agent Security Bench (ASB) with Qwen2.5-7B, comprising 14,280 executions from ten scenario groups. ASB/Qwen2.5-14B provides an aligned model-size extension (Zhang et al., 2025; Yang et al., 2024). We further evaluate additional agent models on AgentLab and an additional environment on AgentCanary

Monitor candidates. The primary candidate family contains five hidden-state telemetry monitors, d1, d2, d4, d8, and d16, with representation dimensions 1, 2, 4, 8, and 16. We use matched variants of the same monitoring construction because ranking transfer compares candidate preferences under a common prediction target and warning interface. This design holds the information source, supervision target, and alarm interface fixed while varying representation capacity, avoiding confounds that would arise from comparing heterogeneous monitor systems. Hidden-state telemetry provides a common prefix-causal signal from which such matched candidates can be constructed. We additionally evaluate a mixed family that adds a prefix-causal behavior-feature monitor to test whether the findings extend beyond this controlled setting (Appendix C.6).

Prediction and warning settings. All candidates predict whether failure will occur within the next two eligible steps (`failure_within2`). Predictive rankings use scenario-equal negative log-likelihood (NLL) on held-out development data; alternative criteria are reported in Appendix C.7. We evaluate three warning rules—first crossing, two consecutive threshold exceedances, and cumulative evidence over a three-step window—covering instantaneous, persistent, and accumulative behavior (Fahrenkrog-Petersen et al., 2022; Page, 1954). We use episode-level FPR targets of $1\%, 2\%, 5\%, 10\%$, with each candidate calibrated separately to the same target. The rule set and operating-point grid were fixed before inspecting transfer outcomes; exact ranking and tie-breaking conventions appear in Appendix A.

Evaluation protocol. We partition the ten ASB scenario groups into five grouped outer folds, each holding out two groups for final evaluation. Within the remaining development groups, disjoint episode sets are used for monitor fitting, selection calibration, selection evaluation, and warning-threshold calibration. Predictive and policy rankings and all warning thresholds are fixed before the outer groups are evaluated. With five candidates, $\binom{5}{2} = 10$ candidate pairs are compared per fold, yielding 50 fold–pair comparisons for each warning rule–operating-point condition. Held-out evaluation tests whether development-stage selection evidence remains supported on unseen groups and quantifies the held-out recall, FPR, and cost-sensitive utility consequences of predictive versus policy-based selection.

4.2 Predictive Winners Can Fail as Alarms

Predictive and warning evaluation can disagree strongly enough to change which monitor is selected. Across the 60 fold–policy comparisons, the predictive and warning top-ranked candidates differ in 48 cases. Under the unified predictive criterion, d16 ranks first in all five folds, yet at the 5% and 10% FPR targets the warning top-1 differs in all 30 comparisons across three warning rules and five folds. Thus, the mismatch can change the development-stage monitor choice itself.

A predictive preference can be robust to both sampling variation and query specification. In the controlled d8–d1 comparison, predictive evaluation consistently favors d8: all six independently fitted model–environment settings prefer d8, and an exact-query control using `failure_within2` preserves this direction in all 100 predictive splits (Appendix C.1.1). Figure 2A shows the corresponding $\mathrm{NLL}(d1) - \mathrm{NLL}(d8)$ differences, where positive values favor d8 because lower NLL is better.

Such predictive robustness is nevertheless insufficient to determine the preferred monitor under a warning policy. Under first crossing at the 5% FPR target, d1 achieves higher failure recall than d8 (0.0866 versus 0.0711) at nearly identical empirical FPRs (0.0484 versus 0.0487), and the policy preference favors d1 in all 100 locked policy splits (Appendix Table C1). Figure 2B shows the corresponding episode-level difference: a larger fraction of d1 failure episodes crosses its calibrated warning boundary, even though d8 is the stronger predictor. This episode-level property is not determined by pointwise predictive performance, as examined in Section 5.

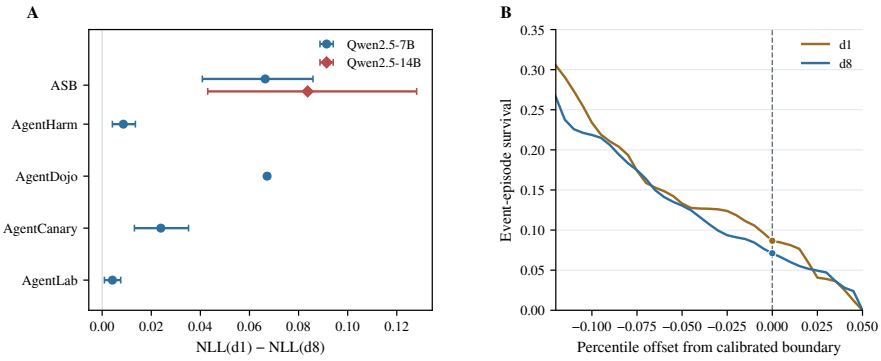


Figure 2: **Predictive advantage and warning reversal. (A)** Cross-setting predictive comparison of d8 and d1. The horizontal axis reports $\mathrm{NLL}(d1) - \mathrm{NLL}(d8)$; because lower NLL is better, positive values favor d8. **(B)** Failure-execution score maxima relative to each monitor's calibrated warning boundary. The survival fraction at zero equals first crossing recall, with d1 exceeding d8 despite d8's predictive advantage.

4.3 Ranking Transfer Depends on Warning Policies

The mismatch established above is not confined to a single pair: its prevalence varies systematically with the warning policy, especially the operating point. In the primary transfer surface, predictive-to-policy agreement decreases as the FPR target increases under all three warning rules (Fig. 3A). Aggregating the strongest-margin comparisons across the three rules, agreement falls from $36/45$ pair–rule instances at the 1% target to $5/45$ at 10%. The same endpoint contrast persists across

alternative predictive criteria and after restricting comparisons to candidates with closely matched development-stage FPR (Appendix C.4, C.7).

This decline is not explained by ranking instability. Among the 20 pair–rule instances whose predictive and frozen-policy preferences are each at least 0.9 stable at both endpoint budgets, $17/20$ are aligned at 1%, but only $2/20$ remain aligned at 10% (Fig. 3B). Fifteen of these same instances switch from agreement to disagreement. Thus, stable predictive and policy preferences can still reverse their relative ordering as the operating point changes.

Held-out evaluation shows that development-stage transfer evidence is informative but not definitive. Holding development predictive preferences fixed, prospective agreement is $38/45$, $28/45$, $19/45$, and $24/45$ across the four FPR targets (Fig. 3C), so the operating-point dependence persists on unseen groups but does not transfer monotonically. At the same time, combining development-stage alignment with preference stability matches or improves the held-out reliability of a stability-only predictive baseline at matched coverage in all 12 rule–target conditions, with strict improvement in nine (Appendix C.9). Thus, development alignment and stability do not guarantee policy support on unseen groups, but together provide useful evidence for identifying predictive preferences that are more likely to remain supported.

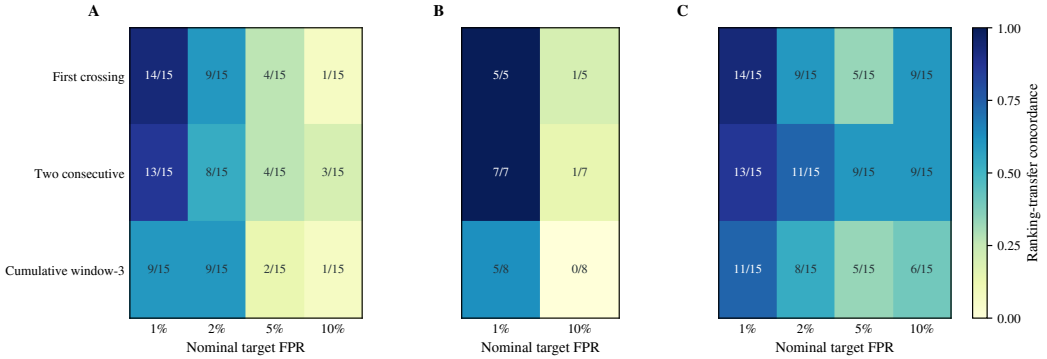


Figure 3: **Predictive selection evidence across warning policies. (A)** Development concordance among the three largest predictive-margin pairs per fold. **(B)** Concordance for the same pairs within each warning rule whose predictive and frozen-policy preferences are at least 0.9 stable at both endpoint FPR targets. **(C)** Agreement between development predictive preferences and held-out policy preferences.

4.4 Consequences for Monitor Selection

Ranking-transfer failures are decision-relevant: they can change which monitor is selected and degrade performance on unseen executions. We examine these consequences at two levels: pairwise regret when predictive and policy preferences stably disagree, and top-1 outcomes when a single monitor is selected by predictive or policy evaluation.

Stable ranking reversals can translate into substantial performance costs on unseen data. Under first crossing at the 2% FPR target, all nine retained stable oppositions incur positive held-out recall regret, and selecting the predictive winner yields a median relative recall loss of 30.4% (Fig. 4A; Appendix Table C19). Thus, disagreement between stable predictive and policy preferences can remain consequential beyond development data.

These pairwise consequences can also propagate to the monitor ultimately selected for deployment. In the mixed candidate family, policy-based selection improves mean held-out recall by 2.2–5.8 percentage points across all six rule–target combinations at the 5% and 10% FPR targets, while yielding lower mean held-out FPRs in all six (Fig. 4B; Appendix Table C20). At stricter operating points, the trade-off is more mixed (Appendix C.8). Ranking-transfer failure can therefore affect both pairwise selection costs and the performance of the final selected monitor.

The same conclusion holds when recall and FPR are combined into a cost-sensitive deployment objective. Using $U_\lambda = \text{Recall} - \lambda\text{FPR}$, policy-based selection in the mixed candidate family yields positive mean utility gains across all evaluated penalty values at the 5% target. At $\lambda = 1$, the mean gains are 0.0611, 0.0737, and 0.0665 for the three warning rules, respectively, and are positive in all five folds for each rule (Fig. 4C).

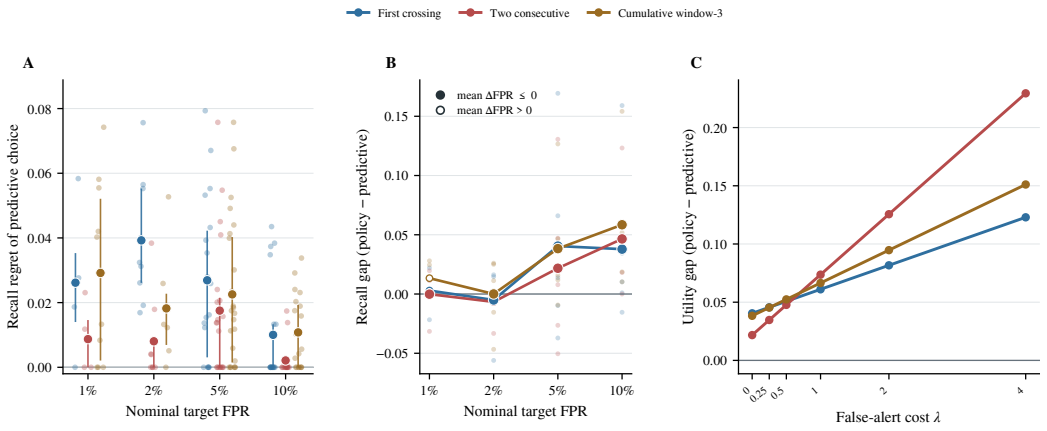


Figure 4: **Consequences of ranking-transfer failure. (A)** Held-out recall regret for stable, opposed pairs. **(B)** Held-out recall gain from policy-based top-1 selection; filled markers indicate nonpositive mean FPR gaps. **(C)** Mean utility gain from warning-based selection under different FPR penalties.

4.5 Breadth Beyond the Primary Setting

Incomplete ranking transfer is not specific to the primary candidate family, warning rule, agent model, or environment. Across all evaluated extensions, predictive and policy rankings remain imperfectly aligned, although the magnitude and operating-point profile vary by setting.

Broadening the candidate and policy spaces preserves incomplete ranking transfer. In the mixed candidate family, first crossing agreement decreases from 0.78 at the 1% target to 0.22 at 10%, with the other two warning rules showing the same endpoint decline. A recursive evidence-accumulation policy likewise shows agreement decreasing from 0.66 to 0.20 across the same endpoint targets (Appendix C.6). Thus, the primary finding is not an artifact of closely matched telemetry candidates or finite-window warning rules.

Incomplete transfer also recurs across agent models and environments, while its exact profile remains setting-specific. The aligned ASB/Qwen2.5-14B execution exhibits a non-uniform transfer surface across all three warning rules and four operating points (Appendix C.5). On AgentLab, all-pair agreement at the 10% target ranges from $11/50$ to $34/50$ across model–policy settings, while the AgentCanary/Qwen2.5-14B extension yields $27/50$–$30/50$ across its three supported warning rules (Appendix C.6). Thus, the existence of incomplete transfer generalizes across settings, but its magnitude and operating-point dependence do not follow a universal pattern.

5 Why Predictive Metrics Do Not Determine Warning Performance

Section 4 shows that even strong and stable predictive preferences need not identify the monitor preferred under sequential warning. We now show why such mismatches are structurally possible: warning performance depends on properties of the episode-level score process that standard pointwise predictive metrics do not determine. This gap appears at two levels. Even first crossing depends on episode-level score extremes relative to calibrated boundaries; order-sensitive rules depend additionally on how those scores are organized over time. We examine these two levels in turn.

5.1 First Crossing Depends on Episode-Level Score Extremes

First crossing performance depends on episode-level score extremes relative to candidate-specific calibrated boundaries, rather than on pointwise predictive quality alone. An episode triggers an alarm if any eligible risk score reaches the monitor's threshold. For monitor j and episode e, define the maximum eligible score as

$$Z_{j,e} = \max_{t \in \mathcal{T}_e^{\text{elig}}} s_{j,e,t},$$

where $\mathcal{T}_e^{\text{elig}}$ contains the steps eligible for the warning query. At FPR target α, the threshold $\theta_j(\alpha)$ is calibrated from the maxima of non-failure executions, and first crossing recall is the fraction of failure executions satisfying $Z_{j,e} \geq \theta_j(\alpha)$.

Monitor selection under first crossing therefore depends on the distribution of failure-execution maxima relative to a boundary determined by non-failure maxima. Standard pointwise predictive metrics do not determine this relationship. The controlled reversal in Section 4.2 illustrates the distinction: although d8 has lower query-matched predictive loss, a larger fraction of d1 failure executions crosses its calibrated warning boundary (Fig. 2B). This dependence is episode-level but not order-sensitive: rearranging scores within an episode preserves its maximum and therefore leaves first-crossing warnings unchanged at a fixed threshold.

5.2 Order-Sensitive Policies Also Depend on Temporal Structure

Order-sensitive rules introduce an additional dependence on temporal structure. Episode-level extremes can characterize first-crossing alerts at a fixed threshold, but not rules that depend on relationships across successive scores. For example, the score sequences $(0.9, 0.9, 0.1, 0.1)$ and $(0.9, 0.1, 0.9, 0.1)$ contain the same scores and have the same maximum, but only the first contains two consecutive high scores. A rule requiring consecutive threshold exceedances can therefore distinguish them. We test this temporal dependence by rearranging existing scores while keeping each monitor's calibrated warning threshold fixed.

Temporal order can change warning performance even when standard pointwise predictive metrics are unchanged. We first permute score–target pairs together within each execution. This preserves every score–label pairing and therefore leaves pointwise predictive summaries unchanged, while altering their temporal order. At the 5% FPR target, mean recall changes by -0.109 for two consecutive and -0.092 for cumulative window-3, while first crossing recall and FPR remain unchanged (Fig. 5A). The unchanged first crossing result is expected because each execution retains the same score maximum. This control isolates temporal ordering, although the permutation can also change the sequence of failure labels.

The effect persists after preserving both failure labels and event timing. We therefore shuffle scores only among positions with the same label within each execution, preserving the same pointwise predictive summaries as well as the first-failure timestamp. At the 5% target, mean recall changes by 0 for first crossing, -0.027 for two consecutive, and -0.060 for cumulative window-3 (Fig. 5B). Thus, order-sensitive warning performance can change even when predictive summaries, stepwise failure labels, and event timing remain fixed.

By contrast, preserving the within-execution temporal structure eliminates the effect. As a negative control, we exchange complete score–target paths only between executions with the same failure status and path length, without changing the order within each path. Recall and FPR remain unchanged for all three warning rules and four FPR targets (Fig. 5C). Together, these controls isolate within-execution temporal structure as information used by the order-sensitive rules but not captured by pointwise predictive summaries. Additional construction and timing checks appear in Appendix D.

Structural implication. The two analyses expose the same structural limitation: pointwise predictive summaries do not, in general, determine warning performance. Let $M(x)$ denote a collection of pointwise predictive metrics evaluated on scored episodes, and let $P_c(x)$ denote warning performance under policy c. If these summaries were sufficient to determine warning performance, any transformation g preserving them would also preserve P_c:

$$M(x) = M(g(x)) \quad \Longrightarrow \quad P_c(x) = P_c(g(x)).$$

The temporal controls directly violate this implication for order-sensitive rules: identical predictive summaries can correspond to different warning outcomes. First crossing exhibits a complementary form of underdetermination: its performance depends on the distribution of episode-level extrema relative to calibrated boundaries, which pointwise predictive metrics do not determine. This establishes underdetermination, rather than showing that temporal structure explains every naturally occurring reversal; the rearranged score sequences also need not be realizable outputs of a causal monitor.

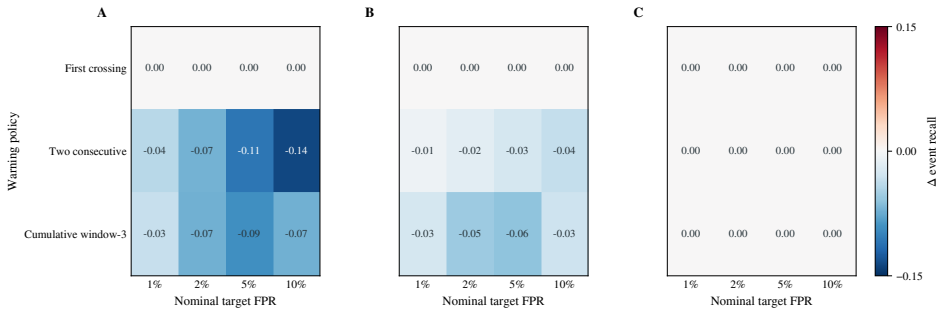


Figure 5: **Temporal structure changes warning performance without changing predictive metrics.** **(A)** Paired score–target permutation. **(B)** Label-preserving score shuffle. **(C)** Intact-path reassignment, which leaves recall and FPR unchanged.

6 Discussion

Predictive and warning evaluation support different selection claims. Predictive evaluation compares monitors as forecasters, whereas warning evaluation compares them as components of a sequential decision system. Their rankings therefore answer different questions, and predictive superiority need not imply downstream selection superiority once a warning rule and operating point are introduced.

Ranking stability does not imply ranking transfer. A reproducible predictive preference can still disagree with an equally stable policy preference. This is structurally possible because sequential warning depends on episode-level score properties—including extremes and temporal organization—that pointwise predictive summaries do not determine. Stability therefore strengthens confidence in each evaluation without guaranteeing agreement between them.

Implications for monitor evaluation and design. Our findings motivate moving beyond predictor-centric benchmarking toward evaluation of the coupled monitor–policy system. Predictive quality alone does not determine how a monitor's score process behaves under a warning rule and operating point, so evaluation should test whether comparative conclusions survive sequential realization. The same distinction matters for monitor design, where optimizing pointwise prediction may be insufficient for the intended warning policy. Ranking-transfer analysis makes this prediction-to-decision gap explicit.

7 Conclusion

We introduced predictive-to-policy ranking transfer to study whether offline predictive preferences remain valid for sequential monitor selection. Predictive and policy rankings can diverge even when both are stable, with measurable consequences for held-out recall and deployment utility. We further show that this mismatch is structurally possible because sequential warning depends on episode-level score extremes and temporal organization not determined by pointwise predictive metrics. These findings motivate making the prediction-to-decision handoff explicit in runtime-monitor evaluation and assessing monitors as part of the downstream warning system they ultimately serve.

AI USE STATEMENT

In this work, we used generative AI tools to assist in method implementation (specifically, drafting and debugging experimental scripts for evaluation workflows). Additionally, we used generative AI tools for tasks under recommended disclosure, including editing the manuscript to improve readability and grammatical flow, querying and summarizing literature to identify potentially relevant related work. All AI-assisted outputs were rigorously verified and audited by the authors. We take full responsibility for the final content of this work, including all text, claims, empirical findings, and code artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We provide the information needed to reproduce our ranking-transfer audit and empirical analyses. Section 3 defines the predictive and policy rankings, transfer measure, and evaluation procedure. Section 4.1 describes the candidate monitors, warning policies, operating points, and grouped evaluation design. The appendix documents data provenance, execution-group splits, data roles, target definitions and eligibility rules, score and threshold calibration, and the full comparison and stability results. It also specifies the controlled score rearrangements used to examine the structural limits of predictive metrics. Monitor selection and calibration are performed on development data, with decisions and thresholds fixed before held-out evaluation.Core analysis code is available at `https://anonymous.4open.science/r/ranking-transfer-3B6C/`.

REFERENCES

Ryan Prescott Adams and David J. C. MacKay. Bayesian online changepoint detection, 2007. URL `https://arxiv.org/abs/0710.3742`.

Maksym Andriushchenko, Alexandra Souly, Mateusz Dziemian, Derek Duenas, Maxwell Lin, Justin Wang, Dan Hendrycks, Andy Zou, J. Zico Kolter, Matt Fredrikson, Yarin Gal, and Xander Davies. Agentharm: A benchmark for measuring harmfulness of LLM agents. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL `https://openreview.net/forum?id=AC5n7xHuR1`.

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. *CoRR*, abs/2606.05414, 2026. doi: 10.48550/ARXIV.2606.05414. URL `https://doi.org/10.48550/arXiv.2606.05414`.

Andrew P. Bradley. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognit.*, 30(7):1145–1159, 1997. doi: 10.1016/S0031-3203(96)00142-2. URL `https://doi.org/10.1016/S0031-3203(96)00142-2`.

Paolo Ceravolo, Marco Comuzzi, Jochen De Weerdt, Chiara Di Francescomarino, and Fabrizio Maria Maggi. Predictive process monitoring: concepts, challenges, and future research directions. *Process Sci.*, 1(1), 2024. doi: 10.1007/S44311-024-00002-4. URL `https://doi.org/10.1007/s44311-024-00002-4`.

Edoardo Debenedetti, Jie Zhang, Mislav Balunovic, Luca Beurer-Kellner, Marc Fischer, and Florian Tramèr. Agentdojo: A dynamic environment to evaluate prompt injection attacks and defenses for LLM agents. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang (eds.), *Advances in Neural Information Processing Systems 37: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. URL `http://papers.nips.cc/paper_files/paper/2024/hash/97091a5177d8dc64b1da8bf3e1f6fb54-Abstract-Datasets_and_Benchmarks_Track.html`.

DeepSeek-AI. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *CoRR*, abs/2501.12948, 2025. doi: 10.48550/ARXIV.2501.12948. URL `https://doi.org/10.48550/arXiv.2501.12948`.

Adam N. Elmachtoub and Paul Grigas. Smart "predict, then optimize". *Manag. Sci.*, 68(1):9–26, 2022. doi: 10.1287/MNSC.2020.3922. URL https://doi.org/10.1287/mnsc.2020.3922.

Stephan A. Fahrenkrog-Petersen, Niek Tax, Irene Teinemaa, Marlon Dumas, Massimiliano de Leoni, Fabrizio Maria Maggi, and Matthias Weidlich. Fire now, fire later: alarm-based systems for prescriptive process monitoring. *Knowl. Inf. Syst.*, 64(2):559–587, 2022. doi: 10.1007/S10115-021-01633-W. URL https://doi.org/10.1007/s10115-021-01633-w.

Tilmann Gneiting and Adrian E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007. ISSN 01621459. URL http://www.jstor.org/stable/27639845.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, pp. 1321–1330, 2017.

Xinmiao Huang, Jinwei Hu, Rajarshi Roy, Changshun Wu, Yi Dong, and Xiaowei Huang. Prefixguard: From llm-agent traces to online failure-warning monitors. *CoRR*, abs/2605.06455, 2026. doi: 10.48550/ARXIV.2605.06455. URL https://doi.org/10.48550/arXiv.2605.06455.

Tanqiu Jiang, Yuhui Wang, Jiacheng Liang, and Ting Wang. Agentlab: Benchmarking LLM agents against long-horizon attacks. *CoRR*, abs/2602.16901, 2026. doi: 10.48550/ARXIV.2602.16901. URL https://doi.org/10.48550/arXiv.2602.16901.

Kateryna Kubrak, Fredrik Milani, Alexander Nolte, and Marlon Dumas. Prescriptive process monitoring: *Quo vadis*? *PeerJ Comput. Sci.*, 8:e1097, 2022. doi: 10.7717/PEERJ-CS.1097. URL https://doi.org/10.7717/peerj-cs.1097.

Changhee Lee, William R. Zame, Jinsung Yoon, and Mihaela van der Schaar. Deephit: A deep learning approach to survival analysis with competing risks. In Sheila A. McIlraith and Kilian Q. Weinberger (eds.), *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pp. 2314–2321. AAAI Press, 2018. doi: 10.1609/AAAI.V32I1.11842. URL https://doi.org/10.1609/aaai.v32i1.11842.

Changhee Lee, Jinsung Yoon, and Mihaela van der Schaar. Dynamic-deephit: A deep learning approach for dynamic survival analysis with competing risks based on longitudinal data. *IEEE Trans. Biomed. Eng.*, 67(1):122–133, 2020. doi: 10.1109/TBME.2019.2909027. URL https://doi.org/10.1109/TBME.2019.2909027.

Jiaqian Li, Yanshu Li, Boxuan Zhang, Ruixiang Tang, and Kuan-Hao Huang. TRACES: proactive safety auditing for multi-turn LLM agents via trajectory-state modeling. *CoRR*, abs/2605.27690, 2026a. doi: 10.48550/ARXIV.2605.27690. URL https://doi.org/10.48550/arXiv.2605.27690.

Peiyang Li, Songping Wang, Yi Huang, Yanhua Shi, Chenhao Zhang, Qi Li, Yueming Lyu, Caifeng Shan, Fengting Li, Chao Feng, Chuanqun Zhu, and Liang Chen. Agentcanary: A security evaluation framework for autonomous AI agents in real executable environments. *CoRR*, abs/2606.10484, 2026b. doi: 10.48550/ARXIV.2606.10484. URL https://doi.org/10.48550/arXiv.2606.10484.

Shi Lin, Peng Qian, Dinghao Liu, Renjie Sun, Sifan Wu, Dezhang Kong, Chenpei Wang, and Xun Wang. Forecasting trajectory-level safety risks in black-box multi-turn interactions. *CoRR*, abs/2607.26820, 2026. doi: 10.48550/ARXIV.2607.26820. URL https://doi.org/10.48550/arXiv.2607.26820.

Jayanta Mandi, Víctor Bucarey, Maxime Mulamba Ke Tchomba, and Tias Guns. Decision-focused learning: Through the lens of learning to rank. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvári, Gang Niu, and Sivan Sabato (eds.), *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pp. 14935–14947. PMLR, 2022. URL https://proceedings.mlr.press/v162/mandi22a.html.

Usue Mori, Alexander Mendiburu, Sanjoy Dasgupta, and José Antonio Lozano. Early classification of time series by simultaneously optimizing the accuracy and earliness. *IEEE Trans. Neural Networks Learn. Syst.*, 29(10):4569–4578, 2018. doi: 10.1109/TNNLS.2017.2764939. URL https://doi.org/10.1109/TNNLS.2017.2764939.

E. S. Page. Continuous inspection schemes. *Biometrika*, 41(1/2):100–115, 1954. ISSN 00063444. URL http://www.jstor.org/stable/2333009.

Dzung Pham, Kleomenis Katevas, Ali Shahin Shamsabadi, and Hamed Haddadi. Agentstop: Terminating local AI agents early to save energy in consumer devices. In *Proceedings of the ACM Conference on AI and Agentic Systems, CAIS 2026, San Jose, CA, USA, May 26-29, 2026*, pp. 1051–1069. ACM, 2026. doi: 10.1145/3786335.3813163. URL https://doi.org/10.1145/3786335.3813163.

Kai Ruan, Zihe Huang, Ziqi Zhou, Qianshan Wei, Jinghao Lin, Xuan Wang, and Hao Sun. Doomed from the start: Early abort of LLM agent episodes via a recall-controlled probe cascade. *CoRR*, abs/2607.06503, 2026. doi: 10.48550/ARXIV.2607.06503. URL https://doi.org/10.48550/arXiv.2607.06503.

Takaya Saito and Mark Rehmsmeier. The precision-recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets. *PLoS ONE*, 10(3):e0118432, 2015. doi: 10.1371/journal.pone.0118432.

Patrick Schäfer and Ulf Leser. TEASER: early and accurate time series classification. *Data Min. Knowl. Discov.*, 34(5):1336–1362, 2020. doi: 10.1007/S10618-020-00690-Z. URL https://doi.org/10.1007/s10618-020-00690-z.

Novin Shahroudi, Viacheslav Komisarenko, and Meelis Kull. Aligning the evaluation of probabilistic predictions with downstream value. In Inês Lynce, Nello Murano, Mauro Vallati, Serena Villata, Federico Chesani, Michela Milano, Andrea Omicini, and Mehdi Dastani (eds.), *ECAI 2025 - 28th European Conference on Artificial Intelligence, 25-30 October 2025, Bologna, Italy - Including 14th Conference on Prestigious Applications of Intelligent Systems (PAIS 2025)*, volume 413 of *Frontiers in Artificial Intelligence and Applications*, pp. 1969–1976. IOS Press, 2025. doi: 10.3233/FAIA251032. URL https://doi.org/10.3233/FAIA251032.

Mahmoud Shoush and Marlon Dumas. When to intervene? prescriptive process monitoring under uncertainty and resource constraints. In Claudio Di Ciccio, Remco M. Dijkman, Adela del-Río-Ortega, and Stefanie Rinderle-Ma (eds.), *Business Process Management Forum - BPM 2022 Forum, Münster, Germany, September 11-16, 2022, Proceedings*, volume 458 of *Lecture Notes in Business Information Processing*, pp. 207–223. Springer, 2022. doi: 10.1007/978-3-031-16171-1_13. URL https://doi.org/10.1007/978-3-031-16171-1_13.

Alexander G. Tartakovsky, Igor V. Nikiforov, and Michele Basseville. *Sequential Analysis: Hypothesis Testing and Changepoint Detection*. Chapman and Hall/CRC, 2014. doi: 10.1201/b17279. URL https://doi.org/10.1201/b17279.

Niek Tax, Ilya Verenich, Marcello La Rosa, and Marlon Dumas. Predictive business process monitoring with LSTM neural networks. In Eric Dubois and Klaus Pohl (eds.), *Advanced Information Systems Engineering - 29th International Conference, CAiSE 2017, Essen, Germany, June 12-16, 2017, Proceedings*, volume 10253 of *Lecture Notes in Computer Science*, pp. 477–492. Springer, 2017. doi: 10.1007/978-3-319-59536-8_30. URL https://doi.org/10.1007/978-3-319-59536-8_30.

Llama Team. The llama 3 herd of models. *CoRR*, abs/2407.21783, 2024. doi: 10.48550/ARXIV.2407.21783. URL https://doi.org/10.48550/arXiv.2407.21783.

Irene Teinemaa, Niek Tax, Massimiliano de Leoni, Marlon Dumas, and Fabrizio Maria Maggi. Alarm-based prescriptive process monitoring. In Mathias Weske, Marco Montali, Ingo Weber, and Jan vom Brocke (eds.), *Business Process Management Forum - BPM Forum 2018, Sydney, NSW, Australia, September 9-14, 2018, Proceedings*, volume 329 of *Lecture Notes in Business Information Processing*, pp. 91–107. Springer, 2018. doi: 10.1007/978-3-319-98651-7_6. URL https://doi.org/10.1007/978-3-319-98651-7_6.

Irene Teinemaa, Marlon Dumas, Marcello La Rosa, and Fabrizio Maria Maggi. Outcome-oriented predictive process monitoring: Review and benchmark. *ACM Trans. Knowl. Discov. Data*, 13(2): 17:1–17:57, 2019. doi: 10.1145/3301300. URL https://doi.org/10.1145/3301300.

Abraham Wald. *Sequential Analysis*. John Wiley and Sons, 1947.

Haoyu Wang, Christopher M. Poskitt, Jiali Wei, and Jun Sun. Probguard: Proactive runtime monitoring for llm agent safety via probabilistic prediction. In *Proceedings of the 41st IEEE/ACM International Conference on Automated Software Engineering (ASE 2026)*, 2026. Accepted; to appear.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report. *CoRR*, abs/2412.15115, 2024. doi: 10.48550/ARXIV.2412.15115. URL https://doi.org/10.48550/arXiv.2412.15115.

Boxuan Zhang, Jianing Zhu, Zeru Shi, Dongfang Liu, and Ruixiang Tang. Agentforesight: Online auditing for early failure prediction in multi-agent systems. *CoRR*, abs/2605.08715, 2026. doi: 10.48550/ARXIV.2605.08715. URL https://doi.org/10.48550/arXiv.2605.08715.

Hanrong Zhang, Jingyuan Huang, Kai Mei, Yifei Yao, Zhenting Wang, Chenlu Zhan, Hongwei Wang, and Yongfeng Zhang. Agent security bench (ASB): formalizing and benchmarking attacks and defenses in llm-based agents. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL https://openreview.net/forum?id=V4y0CpX4hK.

APPENDIX

The appendix follows the evidence dependencies of the main paper rather than the chronological experiment history. Appendix A specifies the evaluation contract and data provenance; Appendix B expands the predictive reference; Appendix C reports the ranking-transfer evidence; and Appendix D gives the score-process structural analysis. Unless stated otherwise, candidate preferences are determined without outer-test outcomes, and uncertainty intervals are descriptive.

A EVALUATION CONTRACT AND DATA PROVENANCE

Table A1: Candidate, query, policy, and operating-point manifest.

Component	Specification	Role / source
Candidates	d1, d2, d4, d8, d16	Matched telemetry family
Mixed candidates	prefix-causal behavior-feature monitor, d1, d4, d8, d16	Representation-family extension
Predictive query	`failure_within2`	Primary predictive ordering
Warning query	`failure_within2`	Sequential event recall
Primary policies	first crossing; two consecutive; cumulative window-3	Candidate-specific calibration
Policy extension	CUSUM-style	Recursive accumulation audit
Episode-level FPR targets	1%, 2%, 5%, 10%	Candidate-specific calibration
Predictive-margin strata	top-75%, top-50%, top-25% within fold	8, 5, 3 pairs per fold
Feasibility rule	selection-stage FPR $\leq \alpha + 0.02$	Feasibility-first policy ranking
Tie rule	exact primary-family ties use smaller dimension	No unresolved primary comparisons

The candidate family, predictive criteria, warning rules, operating-point grid, and predictive-margin strata were fixed before inspecting the corresponding transfer outcomes.

Table A2: Evaluation collections and execution settings.

Collection	Execution	Role
ASB	Qwen2.5-7B / Qwen2.5-14B	Primary and model-size evaluation
AgentHarm, AgentDojo, AgentCanary, AgentLab	Qwen2.5-7B	Cross-setting predictive reference
AgentLab	Qwen / Llama / DeepSeek	Cross-model transfer extension (C.6)
AgentCanary	Qwen2.5-14B	Cross-environment transfer extension (C.6)

A.1 ESTIMAND AND RANKING RULES

We consider a finite family of candidate runtime signals evaluated on multi-step agent executions. For episode i, candidate j, future-risk query q, and prefix step t, let $s_{i,t}^{(j,q)}$ denote the scalar risk score produced using only information available up to t. The sequence of these scores over an episode forms the candidate's **score process**. In the primary telemetry family, candidates share the same monitoring construction and differ in the dimensionality of the exported representation; we denote the one-, two-, four-, eight-, and sixteen-dimensional variants by $d1, d2, d4, d8$, and $d16$, respectively. The exact candidate and query manifest is reported in Table A1.

A **predictive preference** compares two candidates using a prespecified predictive criterion on selection-stage data. The primary analysis uses scenario-equal-weighted binary NLL for `failure_within2`, with lower loss preferred; alternative criteria are examined separately in Appendix C.7. A **policy preference** compares the same candidates after their score processes are consumed by a declared sequential warning policy. For each candidate, the policy is calibrated independently to the same false-positive operating point, after which candidates are compared by event recall on the corresponding selection data. Predictive and policy preferences are therefore defined on the same candidate pair but under different evaluation stages.

For candidates a and b, let $\Delta_{\mathrm{pred}}^{f}(a, b)$ denote their signed predictive separation in outer fold f, with its sign determining the predictive preference and $|\Delta_{\mathrm{pred}}^{f}|$ defining the predictive margin. Ranking-transfer agreement records whether the policy preference has the same direction as the predictive preference. For downstream policy c and margin threshold τ, let $S_c(\tau)$ be the eligible fold–candidate-pair comparisons satisfying the margin criterion. The primary estimand is the finite-family empirical concordance

$$\widehat{\mathcal{I}}_c(\tau) = \frac{1}{|S_c(\tau)|} \sum_{(f,a,b)\in S_c(\tau)} \mathbf{1}\left[r^f_{\text{pred}}(a,b) = r^f_c(a,b)\right].$$

In the primary five-candidate audit, each fold contributes ten candidate pairs; the top-75%, top-50%, and top-25% strata retain eight, five, and three pairs per fold, respectively. Margin strata are formed using the locked empirical rule reported in Table A1. Exact predictive and policy-recall ties are resolved by smaller dimension, and the tie count is recorded. The feasibility-first policy order places feasible candidates first, with feasibility defined by selection-stage FPR $\leq \alpha + 0.02$; this prespecified tolerance accommodates finite-sample and discrete calibration error without requiring exact empirical attainment of the FPR target. Appendix C.4 tests the corresponding FPR-matching sensitivity. Infeasible candidates remain in the total order. The primary audit has no unresolved pairwise comparisons, so all fixed pair slots contribute to the concordance denominator. All ranking preferences used in the transfer audit are determined before outer-test evaluation. Outer-test data are used only to report held-out predictive and operating behavior, including achieved false-positive rate, recall, and timing. The estimand is conditional on the declared candidate family, queries, policies, folds, and operating points.

A.2 Collections and Provenance

The empirical study centers on Agent Security Bench (ASB), with Qwen2.5-7B providing the primary execution and Qwen2.5-14B providing an aligned model-size extension (Table A2). The primary ASB evaluation supports the predictive candidate comparison, query-matched ranking reversal, policy-by-operating-point transfer audit, deployment-consequence analysis, and temporal-structure controls. The aligned Qwen2.5-14B execution evaluates whether the transfer phenomenon recurs under a larger model.

Additional analyses broaden the candidate and policy spaces. A mixed representation family combines a prefix-causal behavior-feature monitor with d1, d4, d8, and d16 telemetry monitors, while a one-sided CUSUM-style accumulator extends the sequential policy family. AgentLab provides cross-model transfer evaluation across Qwen, Llama, and DeepSeek executions (Team, 2024; DeepSeek-AI, 2025), and AgentCanary provides a cross-environment security-event evaluation. AgentHarm (Andriushchenko et al., 2025), AgentDojo (Debenedetti et al., 2024), AgentCanary(Li et al., 2026b), and AgentLab (Jiang et al., 2026) additionally provide cross-setting predictive references for Qwen2.5-7B.

Detailed data-role allocations for the primary ASB/Qwen2.5-7B execution are reported in Table A3, with outer-fold support and policy-calibration resolution in Table A4.

Table A3: Primary ASB/Qwen2.5-7B role allocations by outer fold. Each slash-separated entry is ordered as folds 0/1/2/3/4. Outer-test splits are grouped by scenario identity; within the eight development groups of each fold, episodes are partitioned into the four roles. Prefix counts use the locked `failure_within2` eligibility contract. Role rows are fold-specific, so an episode may appear in the development roles of multiple outer folds; the collection-total row reports unique collection counts, whereas fold-specific role counts can recur across folds.

Role	Groups/fold	Episodes	Events	Non-events	Eligible prefixes (steps)
Monitor fitting	8	6316/6162/6316/6315/6315	4105/4440/4150/4105/4100	2211/1722/2166/2210/2215	24949/25306/25293/25431/25379
Selection calibration	8	1724/1682/1725/1724/1725	1121/1212/1134/1121/1120	603/470/591/603/605	6807/6915/6944/6938/6877
Selection evaluation	8	1722/1680/1722/1722/1722	1120/1212/1133/1120/1119	602/468/589/602/603	6767/6818/6975/6926/6916
Policy calibration	8	1718/1676/1717/1719/1718	1117/1208/1129/1118/1116	601/468/588/601/602	6723/6802/6950/6921/6895
Outer test	2	2800/3080/2800/2800/2800	2037/1428/1954/2036/2045	763/1652/846/764/755	12017/11751/11224/11221/11251
Collection total	10	14280 unique	9500 unique	4780 unique	57464 unique

We distinguish provenance explicitly because aligned or separately instrumented executions can share an evaluation protocol without being identical data collections. Claims in the main text therefore refer to the primary ASB/Qwen2.5-7B runtime unless an extension is named. The appendix tables provide the exact provenance needed to audit each result without repeating these qualifications throughout the empirical narrative.

Table A4: Primary ASB/Qwen2.5-7B outer-fold support and operating-point resolution. Here N_- is the eligible non-event episode count in the policy-calibration role; $1/N_-$ is the smallest nonzero empirical FPR increment supported by that role.

Fold	Outer groups	Outer episodes	Outer events	Outer non-events	Policy-calibration non-events	$1/N_-$; supported targets
0	2	2800	2037	763	601	0.166%; 1%, 2%, 5%, 10%
1	2	3080	1428	1652	468	0.214%; 1%, 2%, 5%, 10%
2	2	2800	1954	846	588	0.170%; 1%, 2%, 5%, 10%
3	2	2800	2036	764	601	0.166%; 1%, 2%, 5%, 10%
4	2	2800	2045	755	602	0.166%; 1%, 2%, 5%, 10%

Table A5: Sequential warning rules and calibration.

Warning rule	Triggering criterion	Calibration
First crossing	Any eligible score $\geq$ threshold	Candidate-specific threshold for target FPR α
Two consecutive	Two consecutive eligible scores $\geq$ threshold	Candidate-specific threshold at nominal FPR α
Cumulative window-3	Accumulated evidence over three consecutive steps	Candidate-specific threshold at nominal FPR α
CUSUM-style	Recursive clipped-logit accumulation	Candidate-specific h; μ_0 estimated from non-event episodes

A.3 Grouped Folds and Data Roles

Outer-test splits are grouped by scenario or task identity, so no group crosses the development–test boundary. Within the development groups, episodes are partitioned into distinct roles for **monitor fitting**, **selection calibration**, **selection evaluation**, and **policy calibration**. The primary protocol uses grouped outer folds, with each fold reserving a held-out group set for final reporting. The role rows in Table A3 cover the eight development groups, with episode allocations partitioned across these roles within each fold. Primary ASB/Qwen2.5-7B group, episode, event-support, and prefix counts are reported by role and fold in Tables A3–A4.

These roles keep selection separate from final evaluation. Monitor parameters are fit only on the fitting role, and any calibration required to place predictive scores on the declared scale is performed on the selection-calibration role. Predictive candidate preferences are computed on the selection-evaluation role. Sequential-policy thresholds are determined from a separate policy-calibration population using only eligible non-event episodes for the false-positive constraint. Once these quantities are fixed, the outer fold is used only to report held-out predictive and warning behavior.

Consequently, outer-test recall, FPR, and lead-time values describe how a development-selected configuration behaves on unseen groups, while candidate preferences for the primary estimand are fixed on development roles. Because candidate comparisons share candidates and underlying trajectories, uncertainty is assessed by resampling scenario groups within the development roles and recomputing the relevant rankings and transfer quantities while conditioning on fitted out-of-fold score processes. Appendix C reports the replicate counts, membership distributions, and stability summaries.

A.4 Queries, Censoring, and Event Construction

Each risk score is associated with a prespecified future-event query. At a prefix ending at step t, the query asks whether the relevant event occurs within an observable future horizon. Prefixes whose query outcome cannot be determined from the available execution are handled by the locked censoring rule rather than labeled as negative. Table A1 summarizes the primary query and its evaluation roles; collection- and fold-level support is reported in Tables A2–A4.

The primary predictive ordering uses scenario-equal-weighted performance on the same `failure_within2` query consumed by the warning policies. The cross-setting multi-query reference and alternative predictive criteria are reported separately in Appendix B and Appendix C.7, respectively, and do not define the primary ranking-transfer estimand.

For event episodes, timing is anchored to the first occurrence specified by the dataset-specific event contract. Score construction is prefix-causal: the score at step t uses only information available through that step. Event eligibility and censoring are fixed before candidate comparison using the same locked rules within each collection.

A.5 Sequential Policies and Calibration

We evaluate three sequential warning rules that differ in how they consume the score process. **First crossing** raises an alert when the calibrated score first exceeds its threshold,

$$T_{\mathrm{FC}} = \inf\{t : s_t \geq \theta\}.$$

Two consecutive requires threshold exceedance on two adjacent steps,

$$T_{\mathrm{2C}} = \inf\{t : s_{t-1} \geq \theta,\ s_t \geq \theta\}.$$

The **three-step cumulative-window** policy aggregates evidence over each chronological three-step window and alerts at the first window whose statistic reaches its candidate-specific calibrated threshold. We additionally evaluate a one-sided CUSUM-style evidence accumulator as a recursive-policy extension. With calibrated probability score p_t, define

$$z_t = \mathrm{logit}(\mathrm{clip}(p_t, \epsilon, 1-\epsilon)), \qquad G_t = \max\{0, G_{t-1} + z_t - \mu_0\}, \quad G_0 = 0,$$

where μ_0 is the mean of z_t on non-event policy-calibration episodes. The policy alerts at the first t such that $G_t \geq h$, with h calibrated separately for each candidate and target operating point.

For every candidate, policy, fold, and nominal operating point α, the alert threshold is calibrated using the designated non-event policy-calibration episodes. The primary targets are 1%, 2%, 5%, and 10% false-positive rate. Candidate thresholds are calibrated separately so that policy preferences compare event recall under a common nominal false-positive constraint rather than under a shared numerical threshold. Calibration thresholds are then frozen before selection evaluation and outer-test reporting.

Because empirical false-positive rates are discrete in finite calibration banks, not every nominal target is identifiable in every fold or extension. We apply the locked conservative feasibility rule reported in A.1. Conditions that cannot support the requested operating point are marked unresolved and excluded from transfer summaries requiring that condition. Nominal target FPR and achieved held-out FPR are reported separately throughout the paper.

A.6 Metrics, Uncertainty, and Implementation

Predictive evaluation uses NLL, Brier score, AUROC, and AUPRC (Gneiting & Raftery, 2007; Guo et al., 2017; Bradley, 1997; Saito & Rehmsmeier, 2015) over the eligible records defined by each query contract; lower NLL/Brier and higher AUROC/AUPRC indicate stronger predictive performance. Appendix C.7 examines whether the transfer observations depend on the predictive criterion.

Sequential-policy performance is reported at the episode level. **Event recall** is the fraction of eligible event episodes in which the declared policy produces a qualifying warning under the event-timing contract. **False-positive rate** is the fraction of eligible non-event episodes producing an alert. **Lead time** measures warning timing relative to the first relevant event under the locked event contract. Ranking-transfer agreement is computed only from candidate pairs for which the corresponding predictive and policy preferences are defined under the locked tie and resolution rules.

Uncertainty summaries preserve the grouped structure of the data. We use two grouped resampling analyses with different targets. **Frozen-realization directional stability** holds the original candidate-pair identity and candidate-specific warning thresholds fixed while resampling selection-evaluation scenario groups; the resampled frequencies with which predictive and policy directions are preserved are denoted p_{pred} and $p_{\mathrm{pol,frozen}}$. **Calibration-aware transfer stability** separately resamples selection and policy-calibration groups, recalibrates thresholds, and reapplies the locked predictive-margin selection rule. Pair identities may therefore change although the top-k count is fixed. We report the resulting distributions as descriptive reselected-stratum stability summaries. Resampling is performed at the scenario or task-group level, and permutation procedures preserve the unit required by the corresponding control.

Full implementation provenance, seeds, checksums, feature-dimension checks, and runner identifiers are provided in the artifact manifest.

B PREDICTIVE REFERENCE

The cross-setting d8–d1 reference retains the original multi-query predictive objective; the primary ASB candidate ranking and the query-matched control use `failure_within2`.

Table B1: Cross-setting d8 versus d1 predictive reductions.

Setting	Candidate	NLL reduction	Interval	Positive groups
ASB / Qwen2.5-7B	d8	0.0664	[0.0408, 0.0859]	9/10
ASB / Qwen2.5-14B	d8	0.0837	[0.0430, 0.1281]	9/10
AgentHarm / Qwen2.5-7B	d8	0.0086	[0.0042, 0.0135]	8/8
AgentDojo / Qwen2.5-7B	d8	0.0672	[0.0667, 0.0682]	2/2
AgentCanary / Qwen2.5-7B	d8	0.0239	[0.0131, 0.0352]	137/250
AgentLab / Qwen2.5-7B	d8	0.0042	[0.0009, 0.0075]	55/88

Table B2: ASB fold-level predictive candidate sweep. Each cell reports scenario-equal-weighted `failure_within2` NLL with the within-fold predictive rank in parentheses. Calibration uses `selection_calibration` and evaluation uses disjoint `selection_evaluation` records. Lower NLL is better; bold entries denote the predictive winner.

Candidate	Fold 0	Fold 1	Fold 2	Fold 3	Fold 4
d1	0.063451 (3)	0.081428 (5)	0.068190 (4)	0.069817 (5)	0.078049 (5)
d2	0.064609 (4)	0.070451 (4)	0.068482 (5)	0.065149 (4)	0.068806 (3)
d4	0.059677 (2)	0.063421 (3)	0.065299 (2)	0.063472 (3)	0.066601 (2)
d8	0.065311 (5)	0.062854 (2)	0.065610 (3)	0.059781 (2)	0.071523 (4)
d16	**0.055102 (1)**	**0.059956 (1)**	**0.060508 (1)**	**0.057586 (1)**	**0.063532 (1)**

This section expands the predictive evidence used as the reference ordering in the ranking-transfer analysis. We first test whether the d8-over-d1 preference recurs across independently fitted execution settings, then examine the complete d1–d16 candidate family on the primary ASB runtime. These analyses establish the predictive reference independently of sequential warning outcomes; the matched trajectory-only baseline remains available in the artifact.

B.1 CROSS-SETTING D8 VERSUS D1

We compare the one- and eight-dimensional telemetry exports while holding the monitoring protocol fixed within each execution setting. The ASB/Qwen2.5-7B primary result and aligned ASB/Qwen2.5-14B model-size check both prefer d8; the four additional Qwen2.5-7B settings retain the same direction (Table B1), matching the hierarchy in Fig. 2A.

The strongest characterization comes from the primary ASB/Qwen2.5-7B runtime within this cross-setting reference. Averaged over its declared three-query predictive objective, the d8-over-d1 reduction in held-out NLL is

$$0.0664\ [0.0408,\ 0.0859],$$

where positive values denote lower loss for d8. The same directional preference appears in the additional execution settings. Across settings, the evidence supports directional reproducibility, with the primary ASB/Qwen2.5-7B runtime providing the strongest grouped characterization.

The aligned ASB/Qwen2.5-14B execution provides a model-size extension under the corresponding locked protocol. Its d8-over-d1 NLL reduction is

$$0.0837\ [0.0430,\ 0.1281],$$

with a positive difference in 9 of 10 grouped comparisons. Additional predictive extensions also retain the same sign where the corresponding execution bank supports the comparison. Table B1 records the effect estimates and grouped support.

These results establish a reproducible local preference between two members of the telemetry family and provide a clear predictive reference for the downstream transfer audit.

B.2 Full ASB Candidate Family

We next examine whether the predictive evidence on the primary ASB/Qwen2.5-7B runtime extends beyond the d8-versus-d1 comparison. The candidate family consists of otherwise matched telemetry exports with dimensions

$$d \in \{1, 2, 4, 8, 16\}.$$

Table B2 reports NLL, predictive rank, and predictive selection for each fold and candidate.

The full family exhibits a clear top-level predictive preference: d16 is selected by the primary predictive criterion in all five outer folds, while the remaining candidates are not monotonically ordered by export dimension in every fold. The result is therefore stronger than a single pairwise comparison: predictive evaluation supplies a stable top candidate and nontrivial separation among multiple alternatives before any sequential-policy criterion is applied.

The ASB sweep therefore supplies a clear predictive reference for the ranking-transfer audit. The aligned Qwen2.5-14B results extend this characterization under the Qwen2.5-7B-defined protocol with separate provenance.

C Ranking-Transfer Evidence

This section provides the complete evidence underlying the predictive-to-policy ranking-transfer results in the main text. For the primary transfer estimand, predictive and policy preferences are defined on the development roles; outer-test outcomes do not define those labels. We separately report a frozen prospective audit in which the development predictive preference is held fixed and compared with the policy preference realized on unseen outer-test groups.

We proceed from a controlled two-candidate separation to the primary full-family transfer surface, then report stable-pair evidence, dependence-aware resampling, generalization across executions and groups, breadth across candidate and policy families, sensitivity analyses, deployment consequences, and an illustrative use case.

C.1 Controlled Predictive–Policy Separation

Table C1: Controlled d8-versus-d1 reversal at 5% target FPR.

Candidate	Event recall	Achieved FPR	Preference
d1	0.0866	0.0484	Policy preferred
d8	0.0711	0.0487	Predictive preferred

We first examine the d8-versus-d1 comparison using their common out-of-fold score processes. Predictive evaluation prefers d8 over d1 under the primary forecasting criterion. We then realize the same candidate score processes through a first crossing warning rule, calibrating each candidate independently to the same nominal 5% false-positive operating point.

The sequential preference reverses. Event recall is **0.0866 for d1** and **0.0711 for d8**, while the achieved false-positive rates are nearly identical at **0.0484** and **0.0487**, respectively. Expressed as d8 minus d1, the recall difference is negative across all policy splits and ranges from

$$-0.0217 \text{ to } -0.0029.$$

Table C1 summarizes recall and achieved FPR; complete split-level threshold and timing rows are provided in the artifact. The nearly matched achieved FPRs isolate the opposite preference under a common operating regime.

Predictive evaluation and sequential realization answer different selection questions, and this controlled comparison shows that they can induce opposite preferences when applied to the same underlying candidate score processes.

First crossing alignment audit. We verified the episode-maximum representation used in Fig. 2B against the original controlled first crossing implementation over the 100 locked policy splits and two candidates. Across all 200 split–candidate rows, direct first crossing alerts and eligible-episode maximum thresholding produced zero alert mismatches. Recall and FPR were identical under the two computations, and the maximum absolute discrepancy from the original split-level artifact was 9.71×10^{-17}. The resulting boundary values recover the reported controlled-reversal operating points: $S_{d1}(0) = 0.0866$, FPR $= 0.0484$, and $S_{d8}(0) = 0.0711$, FPR $= 0.0487$. Because the finite calibration banks do not place every frozen threshold at an exact empirical 95th percentile, the boundary-relative coordinate centers each curve on its realized threshold; the threshold percentile averages 0.951640 for d1 (range 0.950628–0.958159) and 0.950628 for d8.

C.1.1 Exact-Query Control

The cross-setting predictive reference in Appendix B uses its original multi-query objective, whereas the controlled warning comparison in Section C.1 uses `failure_within2`. We therefore repeat the predictive comparison using only this exact warning query.

Table C2: Exact-query predictive control using `failure_within2`. Each entry reports the number of post-fit role splits in which predictive evaluation prefers d8 to d1.

Collection	Execution	d8 preferred / repeats
ASB	Qwen2.5-7B	100/100
AgentDojo	Qwen2.5-7B	100/100
AgentDojo	Qwen2.5-14B	100/100
AgentHarm	Qwen2.5-7B	100/100
AgentCanary	Qwen2.5-7B	100/100
AgentLab MiniWoB	Qwen2.5-7B	100/100

The predictive preference remains unchanged in all six evaluated settings and in all 100 post-fit role splits within each setting. For the controlled ASB comparison, query matching therefore leaves the predictive direction unchanged, showing that the reversal in Section C.1 is not an artifact of a mismatch between the predictive and warning targets.

C.2 Primary Full-Family Transfer Surface

We quantify this relationship through pairwise ranking-transfer agreement. For every candidate pair, outer fold, warning rule, operating point, and predictive-margin stratum, we record whether the policy preference agrees with the predictive preference defined on the corresponding selection-stage data. The full audit contains the Cartesian product of five candidates, five outer folds, three policy families, and four target operating points for the primary execution, subject to the resolution rules defined in Appendix A. The subsections below separate the transfer surface, stable-pair evidence, dependence-aware resampling, generalization, breadth, sensitivities, deployment consequences, and illustrative use.

Figure C1 reports the resulting policy-by-operating-point transfer surface for the primary execution. The accompanying tables report raw agreement counts, fold-level values, margin strata, selected-pair count distributions under group resampling, and held-out operating quantities. Unresolved cells are marked explicitly rather than assigned zero agreement.

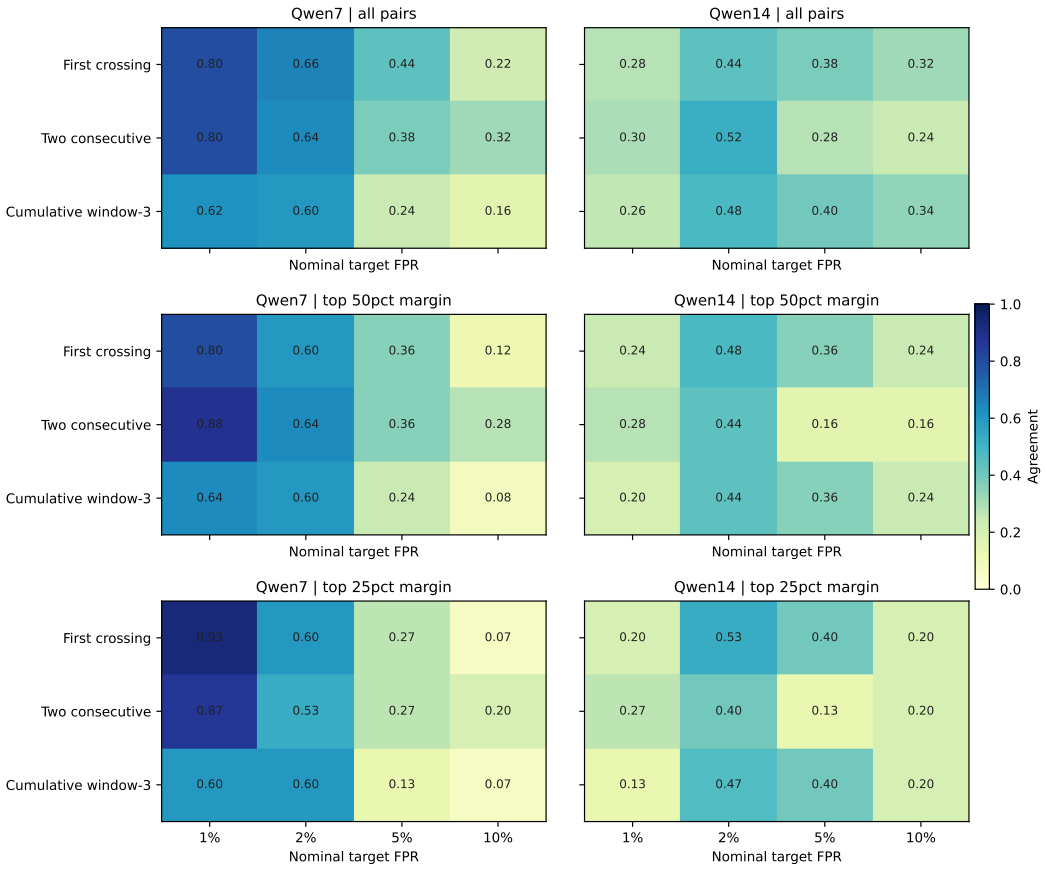


Figure C1: **Complete ranking-transfer surfaces for the primary and aligned executions.** Concordance across warning rules, operating points, and predictive-margin strata for the primary Qwen2.5-7B audit and aligned Qwen2.5-14B execution.

C.3 Transfer Among Stable Candidate Pairs

For the frozen-realization audit, the original pair identity and candidate-specific warning thresholds are fixed, selection-evaluation scenario groups are resampled, and exact ties are retained. The marginal directional-stability and tie-rate summaries are provided in the artifact.

Table C3: Transfer among jointly stable candidate pairs under the unified scenario-equal-weighted predictive criterion; joint stability requires $p_{\mathrm{pred}} \geq .9$ and $p_{\mathrm{pol,frozen}} \geq .9$ at both 1% and 10%.

Policy	Jointly stable	Aligned @1%	Aligned @10%	Aligned → opposed
cumulative window-3	8	5/8	0/8	5/8
first crossing	5	5/5	1/5	4/5
two consecutive	7	7/7	1/7	6/7

Table C3 restricts to the same pairs satisfying

$$p_{\mathrm{pred}} \geq .9, \qquad p^{1\%}_{\mathrm{pol,frozen}} \geq .9, \qquad p^{10\%}_{\mathrm{pol,frozen}} \geq .9.$$

For first crossing, the common intersection changes from **5/5 aligned at 1% to 1/5 at 10%**, with **4/5** aligned-to-opposed transitions. For two-consecutive, the corresponding values are **7/7 to 1/7**, with **6/7** transitions; for cumulative-window-3 they are **5/8 to 0/8**, with **5/8** transitions. Holding pair identity and both directional-stability criteria fixed therefore preserves the operating-point effect.

C.4 Dependence-Aware Resampling and Calibration Sensitivity

Table C4: Canonical-pair overlap for the Qwen2.5-7B top-25% predictive-margin stratum under the unified ranking. The 15 selected slots are three pairs per fold, selected by absolute margin.

Summary	Value	Canonical pairs by occurrence	Fold-pair Jaccard overlap
Selected slots	15	d1–d16: 5; d2–d16: 3; d8–d16: 2	mean 0.40
Unique pairs	6	d1–d4: 2; d1–d8: 2; d1–d2: 1	range [0.20, 1.00]

Table C5: Raw finite-family transfer surface for the unified primary audit. Each fold contributes the three largest absolute predictive margins, yielding 15 comparisons per policy–operating-point cell. Each cell reports aggregate agreement among 15 comparisons (three pairs per fold across five folds).

Policy	1%	2%	5%	10%
First crossing	14/15	9/15	4/15	1/15
Two consecutive	13/15	8/15	4/15	3/15
Cumulative window-3	9/15	9/15	2/15	1/15

Table C6: Paired grouped-resampling contrast between the 1% and 10% operating points.

Policy	Stage	Observed Δ	Median	2.5–97.5%	Positive-resample fraction
first crossing	development	0.867	0.733	[0.467, 0.867]	1.000
two consecutive	development	0.667	0.600	[0.333, 0.802]	1.000
cumulative window-3	development	0.533	0.467	[0.267, 0.600]	1.000
first crossing	prospective	0.333	0.267	[0.200, 0.333]	1.000
two consecutive	prospective	0.267	0.200	[0.067, 0.267]	0.997
cumulative window-3	prospective	0.333	0.333	[0.200, 0.467]	1.000

Finite-family dependence. Here $\Delta = \mathcal{I}(1\%) - \mathcal{I}(10\%)$, with the same scenario-group draw paired across the two operating points. Development and prospective contrasts are both positive for all three policies; the prospective contrast is smaller than the development contrast. Ranges are empirical 2.5–97.5 percentiles over 2,000 paired scenario-group resamples and are reported descriptively rather than as conventional confidence intervals. The positive fraction is the fraction of resamples with $\Delta > 0$, not a posterior probability or significance test.

The raw strongest-margin Qwen2.5-7B first crossing counts are shown in Fig. 3A and the complete policy surfaces in Fig. C1. The 15 top-margin slots comprise six canonical pairs, with mean fold-pair Jaccard overlap 0.40 (Table C4). The grouped analyses therefore preserve scenario dependence rather than treating the pair slots as independent observations.

The primary transfer, stability, intersection, and prospective values in this section are regenerated from the final unified artifacts ("ranking_transfer_strongest_margin.csv", "stability_audit.csv", "stable_pair_intersections.csv", and "prospective_transfer.csv"). The calibration-aware reselected-stratum decomposition remains a separate resampling diagnostic and does not define the primary ranking.

C.4.1 Calibration-Aware Reselected-Stratum Sensitivity

Table C7: Calibration-aware sensitivity for Qwen2.5-7B first crossing. Cells report median [2.5–97.5%] agreement over grouped resamples.

Target	Frozen realization	Recalibrated / reselected
1%	0.867 [0.667, 1.000]	0.600 [0.333, 0.867]
2%	0.600 [0.400, 0.800]	0.667 [0.400, 0.933]
5%	0.400 [0.200, 0.602]	0.333 [0.133, 0.600]
10%	0.267 [0.067, 0.533]	0.333 [0.067, 0.600]

The top-k rule fixes 15 comparisons per replicate while recalibration can change pair membership. Table C7 shows that the operating-point profile persists under both frozen-realization and recalibrated/reselected analyses; the complete factorial decomposition and random-correspondence reference are provided in the artifact.

Frozen-realization directional stability. For the frozen-realization audit, the original pair identity and candidate-specific warning thresholds are fixed, selection-evaluation scenario groups are resampled, and exact ties are retained. The marginal directional-stability and tie-rate summaries are provided in the artifact; the jointly stable same-pair result used by the main argument follows.

Selection-stage FPR matching. Sharing the same FPR target does not ensure that two candidates have the same empirical FPR on selection data. We therefore repeat the transfer audit after retaining only pairs whose selection-stage FPRs satisfy

$$\Delta^{f}_{\mathrm{FPR,dev}}(a,b) = \left|\widehat{\mathrm{FPR}}^{f}_{\mathrm{selection},a} - \widehat{\mathrm{FPR}}^{f}_{\mathrm{selection},b}\right| \le \epsilon.$$

The candidate orientation, candidate-specific thresholds, and development rankings are fixed before filtering; no outer-test quantity enters the pair definition. Table C8 reports the strongest-margin first crossing slice that directly corresponds to the headline transfer contrast. The all-pair and outer-test transport diagnostics give the same qualitative conclusion and are provided in the artifact.

Table C8: Selection-stage FPR matching for the strongest predictive-margin first crossing pairs. Each cell reports (N;) development agreement; prospective agreement.

Target	$\epsilon = .005$	$\epsilon = .010$	$\epsilon = .020$	Test-gap median[†]
1%	12; 11/12; 11/12	15; 14/15; 14/15	15; 14/15; 14/15	0.010
2%	8; 6/8; 6/8	10; 8/10; 7/10	15; 9/15; 9/15	0.021
5%	5; 1/5; 2/5	9; 3/9; 3/9	12; 4/12; 4/12	0.022
10%	6; 1/6; 5/6	11; 1/11; 9/11	15; 1/15; 9/15	0.020

[†]Median held-out FPR gap among pairs retained at $\epsilon = .010$. The margin stratum is defined from development predictive separation.

The strongest-margin first crossing slice gives the same interpretation: at $\epsilon = .010$, development agreement is **14/15, 8/10, 3/9, and 1/11**, while prospective agreement is **14/15, 7/10, 3/9, and 9/11** across the four targets (Table C8). As a descriptive post-hoc check, applying the same numerical filter to outer-test achieved FPRs yields much smaller within-pair gaps, but the retained set then depends on outer-test outcomes. We therefore use the development-filtered tables as the matched-realized-FPR sensitivity and retain the outer-filtered calculation only as a transport diagnostic.

Development-stage achieved-FPR proximity therefore leaves the transfer profile non-uniform. Full all-pair and post-hoc outer-test matching tables are available in the machine-readable results.

C.5 Generalization Across Executions and Unseen Groups

Aligned Qwen2.5-14B execution. The aligned Qwen2.5-14B ASB execution bank contains 14,280 episodes across 10 scenario groups and 92,065 step rows. Its five scenario-grouped folds contain 284–320 policy-calibration non-event episodes, corresponding to minimum empirical FPR increments of approximately 0.31%–0.35%. The d1, d2, d4, d8, and d16 candidate family is evaluated using the matched `failure_within2` predictive NLL criterion, the three warning rules, the 1%, 2%, 5%, and 10% operating points, and the same margin construction. The corresponding Qwen2.5-7B matched-query comparator is contained in the same `predictive_policy_ranking_transfer/` artifact.

Qwen2.5-14B strongest-margin surface. Each cell reports agreement among the 15 development-defined top-25% pair slots.

Table C9: Qwen2.5-14B strongest-margin ranking-transfer surface. ...

Policy	1%	2%	5%	10%
First crossing	3/15	8/15	6/15	3/15
Two consecutive	4/15	6/15	2/15	3/15
Cumulative window-3	2/15	7/15	6/15	3/15

Qwen2.5-14B all-pair surface. Each cell reports agreement among 50 fold–candidate-pair comparisons.

Policy	1%	2%	5%	10%
First crossing	14/50	22/50	19/50	16/50
Two consecutive	15/50	26/50	14/50	12/50
Cumulative window-3	13/50	24/50	20/50	17/50

Both strata peak near 2% rather than following the sharp Qwen2.5-7B 1%-to-10% decline. The aligned execution also exhibits a strongly non-uniform transfer surface, although its numerical operating-point profile differs from the primary audit. Because the complete surface preceded the final full-surface protocol lock, it is reported as an aligned extension with separate provenance.

Table C10: Prospective outer-test ranking validation in the top-25% predictive-margin stratum. *Prospective transfer* compares the development predictive preference with the outer-test policy preference; *policy generalization* compares the development policy preference with the outer-test policy preference. The development predictive preference generalizes to the outer-test predictive ranking in $12/15$ retained comparisons.

	Prospective transfer				Policy generalization			
Policy	1%	2%	5%	10%	1%	2%	5%	10%
First crossing	14/15	9/15	5/15	9/15	13/15	13/15	12/15	6/15
Two consecutive	13/15	11/15	9/15	9/15	11/15	10/15	9/15	7/15
Cumulative window-3	11/15	8/15	5/15	6/15	13/15	12/15	12/15	10/15

Unseen groups. The prospective audit treats the predictive-to-policy relation itself as a held-out object of evaluation. The development predictive preference is held fixed and compared with the policy preference on unseen outer-test groups. Table C10 reports prospective transfer and policy-ranking generalization. The development predictive preference agrees with the outer-test predictive ranking in $12/15$ retained comparisons. The held-out surface has a different operating-point profile, with first crossing prospective counts of **14/15, 9/15, 5/15, and 9/15** across the four operating points.

Across the 60 fold–policy–operating-point conditions, the two top-1 selectors disagree in 48 cases. This counts selector-level disagreement and is distinct from the pairwise transfer denominator.

The equal mismatch totals do not imply identical policy rankings: the two policy procedures select different policy top-1 candidates in six conditions. The recall-only result is a sensitivity analysis; the primary estimand keeps the protocol-defined feasibility-first ordering.

Table C11: Top-1 selection disagreement under the unified scenario-equal-weighted predictive ranking. Entries count folds in which the predictive top-1 and feasibility-first policy top-1 differ.

Warning rule	1%	2%	5%	10%
First crossing	1/5	4/5	5/5	5/5
Two consecutive	3/5	3/5	5/5	5/5
Cumulative window-3	3/5	4/5	5/5	5/5

Table C12: Policy-ranking sensitivity. The primary ranking is feasibility-first, then selection-stage event recall, then smaller dimension. The sensitivity ranking ignores feasibility and uses recall followed by smaller dimension.

Statistic	Result
Feasibility-first top-1 mismatch	48/60
Recall-only top-1 mismatch	48/60
Conditions with different policy top-1	6/60
Exact policy recall ties	20
Unresolved comparisons	0

C.6 Breadth Across Candidates, Policies, Models, and Environments

Table C13: Ranking-transfer concordance for the mixed representation family. Each cell contains 50 fold–candidate-pair comparisons.

Policy	1%	2%	5%	10%
First crossing	0.78	0.74	0.40	0.22
Two consecutive	0.62	0.60	0.24	0.26
Cumulative window-3	0.66	0.56	0.28	0.16

Mixed candidate family. The primary d1–d16 audit deliberately varies export dimensionality within one monitoring construction. We therefore repeat the same query, grouped folds, role separation, policy calibration, and operating-point grid on a mixed representation family containing a prefix-causal behavior-feature monitor together with d1, d4, d8, and d16 telemetry candidates. The behavior-feature predictor uses only prefix-available features; `episode_len`, `prefix_frac`, and actual `remaining_steps` are excluded from predictor inputs. `remaining_steps` is retained only for eligibility and censoring checks.

The complete all-pair transfer surface is reported in Table C13. Across 1% and 10% target FPR, agreement changes from **0.78 to 0.22** for first crossing, **0.62 to 0.26** for two consecutive, and **0.66 to 0.16** for cumulative-window-3. This audit extends the transfer characterization to a mixed representation family.

Table C14: Recursive CUSUM-style transfer summary.

Policy	Target FPR	Agree / total	Agreement
CUSUM-style	1%	33/50	0.66
CUSUM-style	2%	28/50	0.56
CUSUM-style	5%	17/50	0.34
CUSUM-style	10%	10/50	0.20

Recursive policy family. To test policy breadth beyond finite-window threshold rules, we apply the one-sided CUSUM-style evidence accumulator defined in Appendix A.5 to the primary ASB/Qwen2.5-7B score processes. Candidate-specific accumulation thresholds are calibrated under the same grouped policy-calibration contract and the same four nominal operating points; outer-test outcomes do not define the rankings.

All-pair agreement is **33/50 (0.66), 28/50 (0.56), 17/50 (0.34), and 10/50 (0.20)** at 1%, 2%, 5%, and 10%, respectively (Table C14). The recursive detector consequently has its own transfer profile, extending the characterization beyond first crossing and finite-window persistence heuristics.

Table C15: Cross-model ranking-transfer results at 5% and 10% target FPR. Entries report the number of agreeing candidate pairs over the total number of comparisons for all pairs and the top predictive-margin strata.

Warning rule	Target FPR = 5%			Target FPR = 10%		
	All	Top-50%	Top-25%	All	Top-50%	Top-25%
Qwen2.5-7B						
First crossing	20/50	8/25	4/15	23/50	10/25	5/15
Two consecutive	36/50	18/25	10/15	33/50	20/25	12/15
Cumulative window-3	20/50	8/25	4/15	31/50	17/25	10/15
Qwen2.5-14B						
First crossing	20/50	9/25	6/15	24/50	13/25	8/15
Two consecutive	24/50	11/25	7/15	34/50	17/25	10/15
Cumulative window-3	22/50	13/25	8/15	27/50	13/25	8/15
Llama 3.1 8B						
First crossing	29/50	14/25	8/15	29/50	15/25	9/15
Two consecutive	29/50	14/25	8/15	34/50	20/25	11/15
Cumulative window-3	28/50	12/25	6/15	34/50	18/25	11/15
DeepSeek-R1/Llama 8B						
First crossing	—	—	—	21/50	10/25	4/15
Two consecutive	—	—	—	22/50	11/25	8/15
Cumulative window-3	—	—	—	11/50	4/25	3/15

Additional model and environment settings. We further evaluate ranking transfer using the separately generated AgentLab and AgentCanary execution banks. Both extensions provide pairwise transfer records and held-out utility evaluations. Their operating-point support differs from the primary ASB audit: AgentLab reports resolved conditions at 5% and 10% for three agent models and at 10% for DeepSeek-R1/Llama 8B, while the AgentCanary condition summary covers the 10% target. Unresolved lower-FPR conditions are not included in these summaries.

Each reported condition contains 50 fold–candidate-pair comparisons. The top-50% and top-25% predictive-margin strata retain five and three pairs per fold, respectively. These counts describe finite-family agreement rather than independent experimental repetitions.

The AgentLab results show incomplete ranking transfer across the evaluated agent models and warning rules. At the 10% target, first crossing all-pair agreement ranges from 21/50 to 29/50 across models, while cumulative window-3 ranges from 11/50 to 34/50. The results vary across model–policy settings and do not establish a common operating-point profile.

Table C16: AgentCanary/Qwen2.5-14B cross-environment transfer condition summaries from the independent extension bank. The available condition summary is at the 10% target; entries are agreement counts over all 50 pairs, the top-50% margin stratum (25 pairs), and the top-25% margin stratum (15 pairs).

Policy	All pairs	Top-50%	Top-25%
First crossing	27/50	15/25	9/15
Two consecutive	29/50	15/25	11/15
Cumulative window-3	30/50	18/25	12/15

The AgentCanary/Qwen2.5-14B extension provides further evidence of incomplete transfer in an additional environment. At its supported 10% target, all-pair agreement is 27/50, 29/50, and 30/50 for first crossing, two consecutive, and cumulative window-3, respectively. These results extend the audit beyond the primary ASB execution without implying that transfer follows the same pattern across environments. Complete fold-level transfer and held-out utility records are provided in the supplementary artifacts.

C.7 Predictive-Criterion and Operating-Point Sensitivities

Table C17: Alternative predictive criteria by candidate dimension.

Candidate	NLL	Brier	AUROC	AUPRC	Near-event NLL
d1	0.072	0.024	0.982	0.618	0.162
d2	0.068	0.023	0.981	0.646	0.147
d4	0.064	0.022	0.985	0.708	0.137
d8	0.065	0.022	0.985	0.679	0.140
d16	0.060	0.020	0.991	0.844	0.127

Table C18: first crossing all-pair transfer under alternative predictive criteria. Each entry reports agreement among 50 fold–candidate-pair comparisons; the predictive direction is induced separately by each criterion.

Predictive criterion	1%	10%
NLL	40/50 (0.80)	11/50 (0.22)
Brier score	40/50 (0.80)	13/50 (0.26)
Prefix AUROC	41/50 (0.82)	14/50 (0.28)
Prefix AUPRC	43/50 (0.86)	14/50 (0.28)
Near-event weighted NLL	41/50 (0.82)	12/50 (0.24)

The primary analysis uses predictive loss to induce the reference ordering. The alternative-criteria audit repeats the candidate-ranking comparison with Brier score, prefix-level AUROC, prefix-level AUPRC, episode-level discrimination summaries, and near-event-weighted criteria. The higher-dimensional candidate remains favored (Table C17), while first crossing correspondence remains condition-dependent for each reported criterion (Table C18). Across these standard predictive criteria, the top-level predictive preference is preserved and the main transfer pattern is not specific to NLL.

Development achieved-FPR matching. The matched-achieved-FPR analysis is reported in Table C8.

C.8 Deployment Consequence

This deployment-consequence audit was added after the primary ranking-transfer analyses. It does not retrain monitors or redefine transfer labels: candidate orientation, stability strata, and selector choices are fixed from development data, and consequence is evaluated using the existing frozen outer-test utilities.

C.8.1 Conditional regret among development-stable oppositions

The consequence audit separates how often predictive selection loses from how much it loses when it does. Candidate orientation and the stable-opposition stratum are fixed from development; outer-test utilities are then evaluated without reselection. Pairwise regret includes all utility-defined pairs, including policy ties.

Table C19: Conditional outer-test regret among development-stable opposed pairs. Mean absolute regret and relative-loss quantiles are computed over the positive-regret pairs; the retained and positive-regret counts make the conditioning denominator explicit.

Policy	Target	Stable pairs	Positive regret	Mean regret	Median relative	Q90 relative
First crossing	1%	4	3	0.0349	70.1%	80.6%
First crossing	2%	9	9	0.0393	30.4%	42.7%
First crossing	5%	18	13	0.0373	18.1%	39.8%
Two consecutive	5%	20	12	0.0292	12.2%	33.8%
Cumulative window-3	1%	10	7	0.0417	42.1%	61.8%
Cumulative window-3	5%	24	18	0.0301	11.4%	34.3%

At first crossing 2%, all nine development-stable oppositions incur positive outer-test regret, with a mean absolute recall regret of 0.0393 and a median relative loss of 30.4%. The complete policy-by-

operating-point summary, including regret frequency, quantiles, and lead-time utilities, is provided in the machine-readable audit.

C.8.2 MIXED-FAMILY TOP-1 SELECTOR CONSEQUENCE

The heterogeneous candidate pool provides a set-level deployment view. In each fold, the predictive top-1 candidate and the development-policy top-1 candidate are frozen before outer-test evaluation. The resulting recall and achieved-FPR gaps are summarized below; full fold-level records are in the artifact.

Table C20: Mixed-family top-1 selector consequence on outer-test folds. Gaps are development-policy-selected minus predictive-selected; entries give mean recall gap and win/tie/loss counts.

Policy	Target	Mean recall gap	Wins / ties / losses	Mean FPR gap
First crossing	5%	0.0404	3 / 0 / 2	-0.0206
First crossing	10%	0.0379	4 / 0 / 1	-0.0410
Two consecutive	5%	0.0217	3 / 0 / 2	-0.0519
Two consecutive	10%	0.0465	4 / 1 / 0	-0.0452
Cumulative window-3	5%	0.0383	4 / 0 / 1	-0.0282
Cumulative window-3	10%	0.0584	5 / 0 / 0	-0.0321

Across all six reported 5% and 10% mixed-family cells, the development-policy-informed selector has a positive mean outer-test recall gap and a lower mean held-out FPR gap relative to predictive selection. The largest recall improvement is 0.0584 for cumulative-window-3 at 10%, with wins in all five folds.

Comparison with direct development-policy selection. For each fold, policy, and operating point, we separately freeze the development predictive winner and the development policy-utility winner. Both choices are then evaluated without reselection on held-out event recall and FPR. Table C21 summarizes the twelve policy–operating-point cells. Because the two frozen selectors can realize different outer-test FPRs, these signed recall gaps are descriptive operating differences.

Table C21: Outer-test comparison between development-policy and predictive selectors. In Panel A, each entry reports the mean recall gap followed by wins/ties/losses across the five outer folds. Positive recall gaps favor the development-policy selector. In Panel B, negative FPR gaps indicate lower achieved FPR for the development-policy selector.

Policy	1%	2%	5%	10%
Panel A: Mean recall gap (wins/ties/losses)				
First crossing	−0.0118 (0/4/1)	−0.0152 (1/1/3)	0.0154 (4/0/1)	−0.0005 (1/2/2)
Two consecutive	−0.0043 (1/2/2)	0.0022 (1/2/2)	0.0060 (4/0/1)	0.0001 (1/3/1)
Cumulative window-3	0.0065 (2/2/1)	0.0133 (4/1/0)	0.0155 (4/0/1)	0.0123 (4/0/1)
Panel B: Mean held-out FPR gap				
First crossing	−0.0079	−0.0235	−0.0217	−0.0196
Two consecutive	−0.0085	−0.0018	−0.0226	0.0001
Cumulative window-3	0.0023	−0.0104	0.0027	−0.0042

C.8.3 COST-SENSITIVE DEPLOYMENT UTILITY

We evaluate the fixed utility family

$$U_\lambda = \text{Recall} - \lambda\,\text{FPR}$$

over the fixed grid $\lambda \in \{0, .25, .5, 1, 2, 4\}$ at the 5% operating point. At $\lambda = 1$, the mean utility gaps are 0.0611, 0.0737, and 0.0665 for first crossing, two consecutive, and cumulative-window-3, respectively; each is positive in all five folds. The complete cost surface is reported in the artifact and visualized in Fig. 4 of the main paper.

C.9 ILLUSTRATIVE SELECTIVE CORROBORATION

As an exploratory use case, we examine whether high-stability predictive–policy alignment can be used to selectively corroborate predictive preferences.

The matched-coverage values below are regenerated from the unified "matched_coverage_records.csv" and "matched_coverage_summary.csv" artifacts. The all-pair values are unchanged; the accompanying top-50% margin sensitivity is recomputed with the unified fold-3 membership ("2–8" replaced by "4–16"). This remains an illustrative application of the transfer audit rather than a replacement for direct policy evaluation.

Table C22: Matched-coverage comparison at stability threshold $\tau = 0.9$. Entries report decisive-correct / decisive-total. "Stable pairs" gives the number of policy-stable comparisons retained from the 50 fold–candidate-pair comparisons.

Policy	Comparison rule	1%	2%	5%	10%
First crossing	*Stable pairs*	25/50	20/50	14/50	4/50
	Predictive only	36/50	32/50	23/49	27/44
	Stable predictive	19/25	13/20	2/14	4/4
	Selective corroboration	22/25	17/20	12/14	4/4
	Policy-direction reference	23/25	18/20	11/14	2/4
Two consecutive	*Stable pairs*	23/50	14/50	12/50	5/50
	Predictive only	33/49	31/50	30/49	28/45
	Stable predictive	18/23	7/14	7/12	5/5
	Selective corroboration	18/23	10/14	12/12	5/5
	Policy-direction reference	17/23	12/14	11/12	3/5
Cumulative window-3	*Stable pairs*	22/50	12/50	8/50	1/50
	Predictive only	33/50	28/50	20/50	21/46
	Stable predictive	13/22	4/12	0/8	0/1
	Selective corroboration	20/22	9/12	8/8	1/1
	Policy-direction reference	20/22	11/12	8/8	1/1

Margin-stratum sensitivity follows the same protocol and is provided in the artifact.

For a declared threshold τ, a pair is accepted when its development predictive direction and frozen-policy direction are non-tied, aligned, and each has grouped stability at least τ:

$$G_\tau(a, b, c) = \mathbf{1}[p_{\text{pred}} \geq \tau,\ p_{\text{pol,frozen}} \geq \tau,\ r_{\text{pred,dev}} = r_{\text{pol,dev}} \neq 0]\,.$$

Selective corroboration. Accepted comparisons use the development predictive preference as selection evidence; rejected comparisons abstain from treating predictive ranking as sufficient evidence for policy preference. The primary application audit fixes $\tau = 0.9$ and evaluates coverage, selective error, and error capture on the frozen outer-test policy preferences. Table C22 reports all 12 policy–operating-point cells for selective corroboration and two matched-coverage references. The policy-stability-only predictive baseline takes the same number of pairs selected using only $p_{\text{pol,frozen}}$, with cutoff ties resolved by the outcome-independent key (outer fold, canonical pair identifier), and still uses the predictive direction. This is the incremental ablation for the alignment and joint-stability requirements. The matched-coverage policy-direction reference uses that same policy-stability-selected set and applies the development policy direction, providing a strong policy-specific reference. Outer-test policy ties are excluded from accuracy/error denominators and reported separately. Margin-stratum sensitivity is included in the artifact. We treat this analysis as an illustration of one possible use of transfer evidence rather than as a general monitor-selection procedure.

Together, these controls show that a clear and reproducible predictive candidate ordering can retain strongly condition-dependent concordance after sequential realization. Appendix D next isolates one form of score-process structure that affects sequential policies while remaining invisible to standard pointwise predictive summaries.

D Score-Process Structural Analysis

The temporal audit asks whether sequential warning depends on score-process properties left unchanged by standard pointwise evaluation. For each fitted candidate and episode, the unrestricted control permutes paired records $\{(s_t, y_t)\}_{t=1}^{T}$ jointly within the episode. It therefore preserves the score–target multiset exactly while changing chronological organization. A second, target-sequence-preserving transformation permutes scores only among positions with the same target label, leaving

the complete `failure_within2` target sequence and first-event timestamp fixed. Both controls use 25 fold-by-dimension units, 100 permutations per unit, and the original candidate-specific policy thresholds.

D.1 First Crossing Episode-Tail Structure

For each candidate and outer fold, we compute episode maxima on the policy-calibration non-event bank and express selection-stage event maxima relative to the corresponding candidate-specific empirical non-event percentile. Thresholds, feasibility decisions, and score processes remain frozen from the primary audit. Table D1 gives the fold-level tail masses and achieved selection FPRs for the controlled d8–d1 pair.These fold-level selection-stage diagnostics use a different aggregation from the 100-split controlled-reversal summary in Table C1, so their absolute values are not expected to coincide.

Table D1: Fold-level episode-tail diagnostics for the controlled d8–d1 first crossing comparison. Event-tail mass is the fraction of event episodes whose eligible maximum meets or exceeds the candidate-specific calibrated threshold. Thresholds are frozen from policy calibration. The final column reports d1 minus d8 event-tail mass.

		Event-tail mass		Achieved FPR		
Target FPR	Fold	d1	d8	d1	d8	d1–d8 tail gap
1%	0	0.0973	0.0143	0.0133	0.0000	+0.0830
	1	0.0347	0.0660	0.0150	0.0107	-0.0314
	2	0.0291	0.0485	0.0017	0.0017	-0.0194
	3	0.0455	0.0670	0.0066	0.0033	-0.0214
	4	0.0858	0.0527	0.0133	0.0182	+0.0331
2%	0	0.1062	0.0241	0.0166	0.0050	+0.0821
	1	0.0784	0.0866	0.0256	0.0192	-0.0083
	2	0.0865	0.0680	0.0187	0.0034	+0.0185
	3	0.0625	0.0813	0.0116	0.0100	-0.0188
	4	0.1412	0.0769	0.0216	0.0232	+0.0643
5%	0	0.1723	0.0598	0.0449	0.0183	+0.1125
	1	0.1444	0.1444	0.0491	0.0299	+0.0000
	2	0.1668	0.1862	0.0407	0.0424	-0.0194
	3	0.1982	0.1411	0.0449	0.0498	+0.0571
	4	0.2011	0.2055	0.0680	0.0746	-0.0045
10%	0	0.2223	0.2107	0.0814	0.0781	+0.0116
	1	0.1906	0.1955	0.0791	0.0812	-0.0050
	2	0.2215	0.2251	0.0747	0.0764	-0.0035
	3	0.2313	0.2304	0.1013	0.1063	+0.0009
	4	0.2243	0.2225	0.1161	0.1177	+0.0018

At the 5% operating point, aggregate event-tail mass is **0.1766** for d1 and **0.1474** for d8, with corresponding selection-stage achieved FPRs of **0.0495** and **0.0430**. The fold-level d1-minus-d8 tail differences are (+0.1125), (0.0000), (-0.0194), (+0.0571), and (-0.0045). The aggregate tail advantage is therefore not fold-uniform. We use it as a distributional diagnostic illustrating the episode-level structure consumed by first crossing, rather than as a general explanation of candidate ordering.

D.2 Pointwise-Invariant Temporal Transformations

Table D2: Pointwise invariance under the two temporal transformations.

Transformation	Max ΔNLL	Max ΔBrier	Max ΔAUROC	Max ΔAUPRC	Target sequence	Event time
Paired-record permutation	0	0	0	0	permuted	permuted
Target-sequence-preserving shuffle	1.11×10^{-16}	2.78×10^{-17}	0	0	fixed	fixed

D.3 Target-Sequence-Preserving Transformation

The first construction is an exact algebraic control, although its permuted target sequence need not represent a realizable event trajectory. The second retains the target sequence and first-event

timestamp and verifies pointwise invariance to numerical precision. In both cases, pointwise sample counts and target prevalence are unchanged.

Table D3: Changes in warning performance under the two temporal controls. Entries report transformed minus original performance. First crossing recall and FPR are unchanged under both controls at all four operating points.

Policy	Temporal control	1%	2%	5%	10%
Panel A: ΔRecall					
Two consecutive	Paired-record permutation	-0.0421	-0.0714	-0.1090	-0.1370
	Target-preserving shuffle	-0.0080	-0.0157	-0.0268	-0.0354
Cumulative window-3	Paired-record permutation	-0.0344	-0.0719	-0.0922	-0.0720
	Target-preserving shuffle	-0.0259	-0.0546	-0.0601	-0.0309
Panel B: ΔFPR					
Two consecutive	Paired-record permutation	-0.0108	-0.0229	-0.0494	-0.0798
	Target-preserving shuffle	-0.0107	-0.0227	-0.0491	-0.0794
Cumulative window-3	Paired-record permutation	-0.0055	-0.0165	-0.0395	-0.0392
	Target-preserving shuffle	-0.0055	-0.0165	-0.0395	-0.0390

First crossing is invariant in alert occurrence because within-episode permutation preserves the episode maximum. The adjacency-sensitive two-consecutive rule and the cumulative-window functional respond to both transformations (Table D3); preserving the target sequence and first-event timestamp attenuates the recall changes while retaining the pointwise–sequential separation.

D.4 Path-Reassignment Control

This control isolates path reassignment from within-path temporal organization.

Complete score paths are exchanged only between eligible episodes of equal length, preserving each path internally. Across all three policies and 10,000 rows per policy, the maximum absolute changes in recall and FPR are zero.

These controls establish an invariance mismatch. If a predictive functional M is invariant to a transformation g, while a sequential functional P is not, then M alone cannot determine P on that domain. Equivalently, such determination would require a mapping h satisfying

$$P = h \circ M.$$

The constructed score processes violate this factorization for persistence- and accumulation-based policies. Complete permutation quantiles, fold-level responses, timing outputs, and negative-control rows are provided in the artifact. The temporal controls establish an invariance mismatch; they are not intended as a causal reconstruction of the naturally occurring score processes.