A DOWNSTREAM METRIC CAN BE NON-DIAGNOSTIC FOR PREDICTIVE-FIT WORLD-MODEL SELECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

World models are often selected by held-out predictive fit and then judged by downstream decision quality. Yet a downstream metric can itself be non-diagnostic of the selector it evaluates: a weak or null comparison cannot distinguish a poor selector from an insensitive evaluator if the metric does not respond to known increases in decision-relevant information. In a controlled finance testbed with fixed model families and policy pools, the inherited realized-path regret endpoint fails this test: a true-dynamics expected-return selector does not separate from the pre-specified within-family random-model calibration. We repair the estimand to expected net return and replicate its information sensitivity on a fresh population block under family-wise controls before examining the selector comparison. Under the gate-passing endpoint, 0/8 pre-specified finance slices show a multiplicity-controlled detectable advantage of held-out one-step fit over uniform model identity after Holm correction; this is non-detection, not equivalence. A post-hoc order-statistic diagnostic finds no slice with positive excess hindsight-best-family headroom relative to references matched to the observed number of distinct decision classes. This exploratory result weakens the explanation that one substantially better candidate was merely hidden among poor models. The contribution is a validity-gated selector-inference procedure that treats information sensitivity as a prerequisite, demonstrated through a preserved failure–repair–replication sequence rather than a new world model or selector.

1 INTRODUCTION

World models are increasingly judged by the decisions they support rather than by one-step prediction alone. This shift addresses a well-established mismatch: one-step dynamics likelihood need not track control quality (Lambert et al., 2020); value-equivalent models can preserve decision-relevant structure without being globally predictive (Grimm et al., 2020; 2021); and decision-focused learning optimizes predictions for the decisions they induce (Elmachtoub and Grigas, 2022). Recent work has made both model learning and evaluation more decision-aware: DT^2 trains digital twins to preserve candidate-policy rankings (Amad and van der Schaar, 2026), while world-model benchmarks emphasize off-policy value, policy rankings, or closed-loop task success (Yin et al., 2025; Quevedo et al., 2026; Zhang et al., 2026), and model-selection protocols have been developed for off-policy evaluation (Liu et al., 2025). Our question is orthogonal to these advances: *before interpreting a nested model-selector comparison, has the downstream estimand itself demonstrated enough information sensitivity to support that inference?*

Consider a nested pipeline in which an upstream criterion selects one model from a fixed candidate family, the selected model ranks a fixed policy pool, and the induced policy determines downstream regret. If that regret measure does not respond when the selector is given known increases in decision-relevant information, a weak or null selector comparison cannot distinguish a poor selector from an insensitive evaluator. Sanity checks for evaluation tools (Adebayo et al., 2018) and assay sensitivity in experimental design (ICH, 2000) motivate the corresponding prerequisite: before attributing a selector result to the selector, the downstream estimand should demonstrate information sensitivity under a controlled information-sensitivity test.

We test this question in a controlled quantitative-finance environment with fixed model families, fixed policy pools, and population seeds as the statistical unit. The initial analysis evaluated selectors by

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

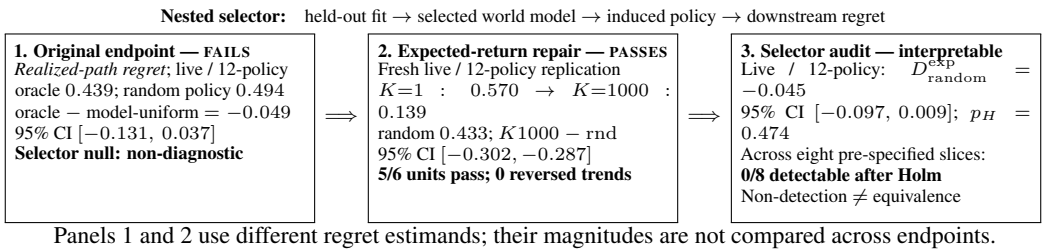


Figure 1: **Why downstream evaluation is not sufficient.** The original realized-path endpoint fails a controlled true-dynamics sensitivity test, so its selector null is non-diagnostic. The expected-return repair passes a fresh sensitivity replication; only then is the selector comparison interpreted, yielding 0/8 Holm-controlled detectable advantages. Values shown are existing live/12-policy examples; the family-quality diagnostic remains post hoc (Fig. 2B).

normalized regret on one realized future path. Because the policy is chosen before future stochastic noise is realized, expected net return is the relevant ex-ante decision target. We challenge the inherited endpoint with an in-domain positive control using true-dynamics expected-return information. The endpoint fails: the oracle remains near the random-policy reference and does not separate from the pre-specified within-family random-model calibration. For example, in the five-model, 12-policy live-feedback slice, oracle mean regret is 0.439 versus a random-policy reference of 0.494, while the paired oracle-versus-model-uniform contrast is -0.049 with 95% CI $[-0.131, 0.037]$. The resulting selector null is therefore treated as non-diagnostic.

We then repaired the estimand to expected-return regret while keeping the simulator, candidate models, policy pools, and selector question fixed. A first fresh-block sensitivity analysis showed information sensitivity but remained formally MIXED under its pre-specified global rule, so its selector outcome was not examined. We next used a fresh replication block disjoint from the earlier diagnostic seeds and changed only the family-wise handling of the negative control. Numerical precision passed in all six sensitivity units, five of six units met all unchanged sensitivity criteria, no information trend reversed, and both negative controls behaved as specified. Thus the repaired endpoint passes the pre-specified information-sensitivity gate for this controlled ladder; this does not establish universal evaluator validity.

We examined the primary selector comparison only after the replication passed. Across eight pre-specified finance slices, held-out one-step fit shows a multiplicity-controlled detectable advantage over exact uniform model identity in **0/8** slices after Holm correction. All eight point estimates favor fit, but the sign pattern was not a pre-specified aggregate test and the slices are dependent, so we do not treat it as a separate positive result. We also do not claim equivalence because no equivalence margin was specified. The result is conditional on these fixed candidate families and policy pools.

The family average can obscure a high-performing individual candidate, so we examine this possibility *post hoc*. The hindsight best policy induced by each fixed family has substantial raw headroom over a single random policy, but no slice shows positive excess headroom over exact order-statistic references matched to the observed number of distinct decision classes (Appendix 9). Taking a minimum over several choices creates headroom by construction, so raw best-in-family improvement does not by itself establish useful family enrichment. The diagnostic is consistent with limited decision enrichment of the fixed families, but it neither identifies a family-capacity mechanism nor changes the confirmatory selector result.

Contributions. The methodological contribution is an information-sensitivity prerequisite for nested selector inference. In the controlled study, preserving the endpoint failure, repairing the estimand, and replicating the sensitivity test changes the status of the same selector question from non-diagnostic to interpretable without changing the candidate models, policy pools, or selector under study. Under the gate-passing endpoint, we then report a multiplicity-controlled audit of held-out predictive fit over eight pre-specified finance slices, with decision-equivalence and order-statistic diagnostics that bound the interpretation of the null. Objective mismatch, decision-focused learning, downstream world-model evaluation, and model-selection protocols are prior foundations rather than novelty claims.

2 Related Work

Decision-aware world-model learning. Prediction–decision mismatch is established rather than our novelty claim. One-step dynamics likelihood need not track control quality (Lambert et al., 2020); value-equivalent models can preserve decision-relevant structure without globally faithful dynamics (Grimm et al., 2020; 2021); and value-targeted regression fits models through value predictions (Ayoub et al., 2020). Recent methods act directly on this mismatch: transition-occupancy matching and policy-shaped prediction emphasize policy-relevant dynamics (Ma et al., 2023; Hutson et al., 2024), DT^2 trains digital twins to preserve candidate-policy rankings (Amad and van der Schaar, 2026), and RLVR-World optimizes world models for task-facing metrics rather than likelihood alone (Wu et al., 2025). Decision-focused learning makes the analogous move in predict-then-optimize pipelines (Elmachtoub and Grigas, 2022). These methods modify model learning or its objective; we instead hold trained candidates fixed and ask whether an *external predictive-fit selector* can be interpreted through its downstream estimand.

Downstream world-model evaluation. Recent benchmarks increasingly evaluate what a world model enables rather than transition fidelity alone. TrajWorld uses off-policy evaluation and model-predictive control (Yin et al., 2025); WorldGym evaluates real-robot policies through world-model rollouts and relative rankings (Quevedo et al., 2026); World-In-World emphasizes closed-loop task success (Zhang et al., 2026); and SIMPLER validates simulated policy evaluation against paired real-world outcomes (Li et al., 2025a). WorldModelBench further shows that generic video-quality metrics can miss application-relevant world-model violations (Li et al., 2025b). These works motivate decision-facing or application-facing evaluation. Our question is narrower: even after choosing a downstream target, has that target demonstrated enough information sensitivity to license a conclusion about the upstream model selector?

Model and policy selection under uncertainty. Model selection itself is also prior art. Liu et al. select candidate value functions or dynamics models for accurate off-policy evaluation and provide a protocol for comparing selectors (Liu et al., 2025). In related settings, unbiased or point-estimate evaluators can fail when reused for hyperparameter search (Saito and Nomura, 2024); pessimistic concentration bounds support more reliable off-policy evaluation and selection (Sakhi et al., 2024); and adaptive Learn-then-Test provides finite-sample-valid selection under multiple testing (Zecchin et al., 2025). These methods improve selection or its statistical guarantees. They do not address the distinct failure studied here: a selector comparison can remain non-diagnostic when the *downstream estimand itself* is insensitive to known decision-relevant information.

Evaluator validity and stochastic decision objectives. Testing whether a metric responds to variables it should measure is a general methodology. Randomization sanity checks expose explanation metrics that remain insensitive to model or data changes (Adebayo et al., 2018; Oikarinen et al., 2025), while negative-control methodology asks whether favorable evaluation can arise under chance behavior (Petersen, 2025). In experimental design, ICH E10 calls the ability to distinguish effective from ineffective interventions *assay sensitivity* (ICH, 2000). Separately, decision-focused optimization already distinguishes empirical or sample-average objectives from expected decision quality: empirical regret can be a poor surrogate under uncertainty (Schutte et al., 2024), and DF^2 identifies sample-average-approximation error as a bottleneck in probabilistic decision-focused learning (Kong et al., 2025); classical SAA theory formalizes the broader approximation problem (Kleywegt et al., 2002). We therefore do not claim novelty for sanity checks, expected-regret objectives, or multiplicity control (Holm, 1979). The residual contribution is their prospective use inside a nested world-model selector analysis: retain the failed endpoint, repair the estimand, replicate its information sensitivity on fresh data, and inspect the selector result only after that gate passes.

3 Problem Formulation and Validity-Gated Audit

3.1 Nested model-to-policy selection

Let s denote a population instance, $\mathcal{M}$ a finite family of already-trained world models, and Π a fixed candidate-policy pool. A policy $\pi \in \Pi$ executed in the true stochastic environment receives return

$J_s(\pi;\epsilon)$ under future noise ϵ. The decision target is its expected return

$$V_s(\pi) = \mathbb{E}_\epsilon[J_s(\pi;\epsilon) \mid s], \qquad \pi_s^\star \in \arg\max_{\pi\in\Pi} V_s(\pi). \tag{1}$$

The policy is chosen before the future noise ϵ is realized, so V_s is the intended ex-ante decision target; a single realized path is evaluation noise rather than an alternative deployment objective. Monte Carlo trajectories approximate V_s, rather than define a different pathwise objective (Kleywegt et al., 2002).

Each world model $m \in \mathcal{M}$ recursively rolls out the same policy pool and produces model-predicted returns $\widehat{V}_{s,m}(\pi)$. The model therefore induces a downstream decision

$$d_s(m) \in \arg\max_{\pi\in\Pi} \widehat{V}_{s,m}(\pi). \tag{2}$$

An upstream selector S chooses one model $S(s) \in \mathcal{M}$; the deployed decision is then $d_s(S(s))$. Our audited selector is held-out one-step fit,

$$m_{\mathrm{fit},s} \in \arg\min_{m\in\mathcal{M}} L_{\mathrm{obs}}(s,m), \tag{3}$$

where L_{obs} is one-step MSE on policy-agnostic held-out data. The scientific object is thus the composition

$$S \longrightarrow m \longrightarrow d_s(m) \longrightarrow V_s(d_s(m)),$$

not model identity or one-step prediction error in isolation.

3.2 Expected downstream regret and calibration

We score an induced policy by normalized expected regret,

$$R_{\mathrm{exp}}(s,m) = \frac{V_s(\pi_s^\star) - V_s(d_s(m))}{V_s(\pi_s^\star) - \min_{\pi\in\Pi} V_s(\pi) + 10^{-9}}, \tag{4}$$

with lower values better. We also retain the unnormalized quantity $AR_{\mathrm{exp}}(s,m) = V_s(\pi_s^\star) - V_s(d_s(m))$ as a secondary scale check.

The primary calibration asks whether held-out fit carries detectable decision-relevant information beyond uniform model identity within the *same fixed family*. Rather than drawing one random model, we integrate that randomization exactly:

$$R^{\mathrm{exp}}_{\mathrm{fit}}(s) = R_{\mathrm{exp}}(s, m_{\mathrm{fit},s}), \tag{5}$$

$$R^{\mathrm{exp}}_{\mathrm{unif}}(s) = \frac{1}{|\mathcal{M}|}\sum_{m\in\mathcal{M}} R_{\mathrm{exp}}(s,m), \tag{6}$$

$$D^{\mathrm{exp}}_{\mathrm{random}}(s) = R^{\mathrm{exp}}_{\mathrm{fit}}(s) - R^{\mathrm{exp}}_{\mathrm{unif}}(s). \tag{7}$$

Negative $D^{\mathrm{exp}}_{\mathrm{random}}$ favors fit selection. This comparator is a *calibration null*, not a proposed deployment baseline: it represents the pre-specified intervention "draw one model identity uniformly from the deployed candidate list" and tests whether the proxy selector demonstrates incremental decision value relative to that intervention. Because it is conditional on the fixed candidate family, failure to reject this calibration cannot establish that held-out fit is intrinsically useless or equivalent to random selection.

3.3 Decision equivalence and calibration multiplicity

Different model identities may induce the same downstream policy. Define

$$m \sim_s m' \quad \Longleftrightarrow \quad d_s(m) = d_s(m'), \tag{8}$$

and let $\mathcal{C}_s = \mathcal{M}/\!\sim_s$ denote the resulting decision classes. Since Eq. (4) depends on m only through $d_s(m)$,

$$m \sim_s m' \implies R_{\mathrm{exp}}(s,m) = R_{\mathrm{exp}}(s,m'). \tag{9}$$

Consequently, a model-uniform calibration can change if a decision-equivalent model is duplicated, even though the available downstream decisions do not change. We also report the duplication-invariant class-uniform robustness

$$R^{\mathrm{exp}}_{\mathrm{class}}(s) = \frac{1}{|\mathcal{C}_s|}\sum_{C\in\mathcal{C}_s} R_{\mathrm{exp}}(s,C), \qquad D^{\mathrm{exp}}_{\mathrm{class}}(s) = R^{\mathrm{exp}}_{\mathrm{fit}}(s) - R^{\mathrm{exp}}_{\mathrm{class}}(s). \tag{10}$$

Equation (9) follows directly from the definitions and exposes a calibration dependence that model identity alone can hide. Class-uniform calibration corresponds to a different, duplication-invariant intervention—uniform sampling over distinct downstream decisions—rather than the pre-specified model-identity intervention. We report $D^{\mathrm{exp}}_{\mathrm{class}}$ only as a secondary robustness analysis and retain Eq. (7) as the primary comparison.

3.4 Information-sensitivity gate before selector inference

A downstream selector comparison is informative only if its evaluator reacts to known increases in decision-relevant information. We operationalize this requirement with a true-dynamics information ladder. For K inner trajectories,

$$\widetilde{\pi}_{s,K} \in \arg\max_{\pi\in\Pi} \frac{1}{K}\sum_{k=1}^{K} J_s(\pi;\epsilon_k), \qquad K \in \{1, 10, 50, 200, 1000\}, \tag{11}$$

using an inner noise stream independent of the high-precision outer estimator of V_s. Increasing K supplies more information about the *same* true dynamics; it does not change the candidate policies or scientific target.

For each pre-specified sensitivity unit, the endpoint passes only if three conditions hold: (i) the paired 95% interval for $R_{\mathrm{exp}}(\widetilde{\pi}_{s,1000}) - R_{\mathrm{exp}}(\text{random policy})$ lies below zero; (ii) its mean regret is at most half the random-policy mean; and (iii) regret decreases with information, requiring a negative median per-instance Spearman correlation with $\log K$ and a paired interval for $K{=}1000$ minus $K{=}1$ below zero. The global gate additionally requires at least five of six units to pass, at most one reversed information trend, a family-wise Holm test with no rejection for a label-permuted null control, and an anti-oracle control that is worse than random in every unit (Holm, 1979; ICH, 2000).

Before inspecting selector outcomes, we fix the fresh block, selector analysis, and sensitivity criteria. A mixed or failed sensitivity analysis stops the procedure; only a replicated pass permits inspection of Eq. (7). This separation prevents a favorable selector outcome from motivating a retrospective change to the endpoint or its sensitivity criterion. The experimental section specifies the simulator, seeds, Monte Carlo precision design, and statistical inference.

4 Experimental Setup

Controlled finance testbed. We use a synthetic trading environment designed to expose prediction-to-decision mismatch without making a real-market claim. The prediction target is the midprice increment over a horizon of $H = 50$. Candidate policies produce signed trades, while the simulator combines stochastic background price moves with optional permanent impact and an aggregate policy-conditioned response. The four regimes are: (a) static/no impact; (b) fixed impact from the evaluated policy; (c) replayed aggregate flow generated once from a reference policy and then held fixed; and (d) live aggregate feedback recomputed from the evaluated policy's recent trades. The downstream objective is self-financing expected net return: mark-to-horizon trading gains minus a quadratic execution-cost term. Exact simulator constants, feature definitions, and implementation property tests are deferred to the appendix.

Policies and candidate world models. The original pool contains 12 fixed trading policies spanning signed TWAP, front/back-loaded execution, momentum, and mean-reversion behavior; the expanded pool contains 24 policies by adding volume, execution-window, and lookback variants. The five-model family used in the four-regime sensitivity analysis and selector audit contains linear and MLP dynamics models; a pooled eight-model family additionally includes GRU and Transformer dynamics. All models are trained on the same policy-agnostic background episodes using the specifications in Appendix H, and held-out one-step MSE defines m_{fit}. Model-based policy values are obtained by recursive rollouts: after initialization, both state history and policy actions depend only on previous model predictions and chosen actions. Full architectures, feature maps, training hyperparameters, and policy definitions are appendix material.

Sensitivity units and confirmatory population. The confirmatory study uses 100 fresh seeds, 4001–4100, chosen before simulation after screening against previously used seed ranges; the

population seed is the statistical unit. The sensitivity analysis uses six pre-specified sensitivity units:

$$(\text{static}, 12\text{-policy}), \quad (\text{static}, 24\text{-policy}), \quad (\text{fixed}, 12\text{-policy}),$$
$$(\text{replay}, 12\text{-policy}), \quad (\text{live}, 12\text{-policy}), \quad (\text{live}, 24\text{-policy}).$$

These units probe the same expected-return estimand while varying policy-pool size and impact/feedback regime. The half-random criterion in Section 3 is a study-specific practical threshold, not a theoretical constant; it was specified before observing the confirmatory outcomes and was not relaxed after one unit missed it.

Independent inner and outer Monte Carlo. For each seed and sensitivity unit, the outer value $V_s(\pi)$ is estimated from $K_{\text{eval}} = 80{,}000$ common-random-number trajectories shared across policies. This outer estimator defines the scoring surface used by both the sensitivity analysis and the selector analysis. The information ladder uses an independent random stream with $K \in \{1, 10, 50, 200, 1000\}$ trajectories and 20 inner repetitions per seed. Thus increasing K changes only how much true-dynamics information the selector receives; it does not change the policy pool, simulator, or evaluation target. Numerical precision is itself gated before scientific interpretation; no K_{eval} escalation or outcome-dependent seed replacement is allowed.

Controls and multiplicity. The random-policy reference is the exact mean of R_{exp} across policies. The label-permutation control is null by construction; for each sensitivity unit we use a two-sided sign-flip randomization test with $10{,}000$ repetitions and Holm correction across the six units. The anti-oracle selects the lowest-value policy and must be worse than random in every unit. The sensitivity analysis returns PASS, MIXED, or FAIL under the pre-specified rules in Section 3; only a replicated PASS permits inspection of the selector analysis.

Confirmatory selector analysis. The selector analysis uses the same fresh seeds and unchanged fixed model families. The eight primary slices are the five-model family with the 12-policy pool in all four regimes, plus the pooled eight-model family in static and live-feedback regimes under both the 12- and 24-policy pools. Primary inference is the mean $D^{\text{exp}}_{\text{random}}$ from Eq. (7), using a paired nonparametric bootstrap over population seeds ($B = 10{,}000$, seed 90001) and Holm correction across all eight slices. Medians, $D^{\text{exp}}_{\text{class}}$, absolute expected regret, and decision-class diagnostics are secondary. The seed list, sensitivity analysis, and selector analysis were fixed before the sensitivity results were inspected; a MIXED or FAIL outcome would have stopped the analysis before any selector result was examined. The sensitivity and selector analyses nevertheless share the same population seeds and outer expected-value surface. Pre-specification prevents outcome-driven changes to the endpoint or analysis, but it does *not* make the two analyses statistically independent.

5 Main Results

Figure 1 summarizes the evidence chain from a non-diagnostic downstream endpoint to a gate-passing selector comparison, and Fig. 2 shows the confirmatory selector result and the post-hoc family diagnostic.

The original downstream endpoint is non-diagnostic. The initial diagnostic asks a simpler question than model selection: does the inherited endpoint improve when policy choice uses expected-return information from the *true* dynamics rather than a learned-model proxy? Under realized-path normalized regret, it does not. The expected-return oracle attains mean normalized regret between 0.439 and 0.517 across the eight diagnostic slices, close to the random-policy reference (0.493–0.500), and every pre-specified paired oracle-versus-within-family-random-selector 95% interval includes zero. The endpoint is therefore classified as INSENSITIVE. This does not imply that true dynamics are unhelpful: the selector optimizes expected return but is scored by a pathwise normalized quantity whose numerator and denominator depend on the realized future.

The expected-return repair shows reproducible information sensitivity. The repair sensitivity analysis replaces the endpoint with Eq. (4) while keeping the selector question fixed. All six sensitivity units show decreasing regret as true-dynamics information increases, and the $K{=}1000$ selector beats random policy with a paired interval below zero in every unit. Five of six units satisfy the pre-specified

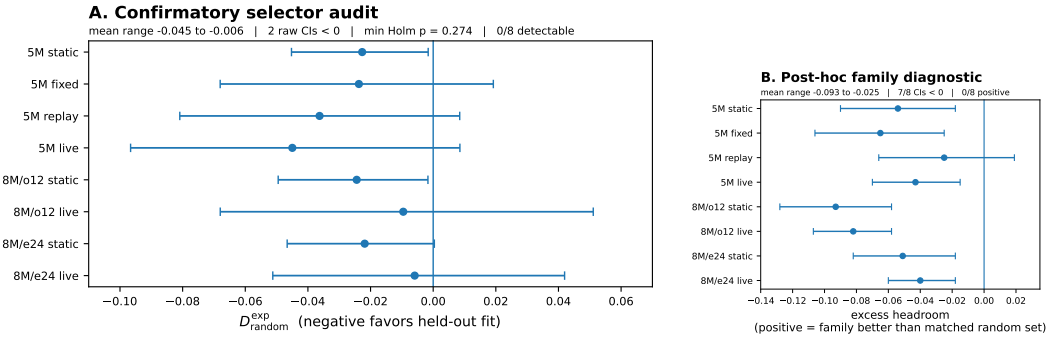


Figure 2: **Confirmatory selector result and post-hoc family diagnostic. A – confirmatory:** primary effects $D^{\text{exp}}_{\text{random}}$ for the eight pre-specified finance slices. Negative values favor held-out fit. Error bars are unadjusted paired-bootstrap 95% intervals, whereas the inferential decision uses Holm-adjusted p-values: two raw intervals exclude zero, the smallest adjusted p-value is 0.274, and **0/8** slices show a multiplicity-controlled detectable advantage. Non-rejection is not equivalence. **B – exploratory:** excess headroom of the hindsight best family-induced policy relative to the exact best-of-C distinct-policy order-statistic reference, where C is the observed number of decision classes. Positive values would indicate family enrichment beyond matched random choice; all eight point estimates are non-positive and seven raw intervals lie below zero. This post-hoc analysis has no multiplicity-controlled confirmatory interpretation and does not alter Panel A.

half-random criterion. The global outcome nevertheless remains MIXED: one label-permutation negative-control interval narrowly excludes zero under the original "all six 95% intervals must include zero" rule. The outcome remains MIXED, and the selector result for this block is not examined.

The sensitivity analysis is then repeated on an independent block of 100 fresh population seeds, disjoint from all seeds used in prior diagnostics, with a prospectively specified family-wise procedure for the label-permutation control. The sensitivity pattern replicates closely (Appendix 5). Numerical precision passes in all six units; five of six satisfy all three sensitivity criteria; no unit has a reversed trend; no label-permutation control rejects after Holm correction; and the anti-oracle control is worse than random in every unit. The only sensitivity miss is again the 24-policy static unit, where $R_{\text{exp}}(K{=}1000) = 0.262$ versus the pre-specified half-random threshold 0.250. Because the global rule permits one failed unit, the replication passes. The selector comparison is examined only after this point.

Held-out one-step fit has no multiplicity-controlled detectable advantage in the tested families. After the replicated pass, we examine the selector comparison on the same fresh population. Across the eight pre-specified slices, all point estimates of $D^{\text{exp}}_{\text{random}}$ favor held-out fit, ranging from -0.045 to -0.006. Two unadjusted 95% intervals lie below zero, but the confirmatory procedure is the pre-specified Holm correction across all eight primary means. Under that procedure, $0/8$ slices show a detectable advantage; the smallest adjusted p-value is 0.274. Held-out one-step fit therefore shows *no multiplicity-controlled detectable advantage over uniform model identity within the tested candidate families*. The all-negative sign pattern is descriptive because the slices share population seeds and overlapping model families and no aggregate sign test was pre-specified. The result is not an equivalence claim because no equivalence margin was specified.

We next examine decision-equivalent models, candidate-family quality, and transport beyond the controlled finance testbed.

6 Analysis and Robustness

Decision classes expose model-identity multiplicity. The candidate families induce only 2.5–3.4 distinct policy classes per population seed on average, so several model identities are behaviorally redundant even though the models genuinely disagree about the final policy. As required by Eq. (9), regret is numerically identical within every induced class. Reweighting uniformly over unique

decision classes makes the secondary contrast $D^{\mathrm{exp}}_{\mathrm{class}}$ negative in all eight slices; four unadjusted 95% intervals lie below zero. We report this direction transparently but do not promote it to a positive claim: no multiplicity-controlled confirmatory rule was pre-specified for this secondary family, and the primary model-uniform calibration remains the pre-specified analysis. The result instead shows why the reference distribution over model identities is scientifically consequential when multiple models induce the same action.

Post-hoc family diagnostic. The selector result remains conditional on the fixed candidate family. A hindsight best-in-family analysis shows raw headroom over a single random policy, but no slice shows positive excess headroom over exact order-statistic references matched to the observed number of distinct decision classes (Fig. 2B; Appendix 9). This analysis is exploratory: it is consistent with limited decision enrichment of the fixed family, not evidence that held-out fit is intrinsically poor or that family capacity is the unique bottleneck.

Continuous control provides a positive transport check, not replication. The earlier nested selector audit on HalfCheetah-v5, Hopper-v5, and Walker2d-v5 yields negative mean selector-regret differences with 95% intervals entirely below zero: -0.052 $[-0.072, -0.031]$, -0.276 $[-0.299, -0.253]$, and -0.161 $[-0.220, -0.101]$, respectively; median intervals are also below zero. These three task-level selections show that the audit can produce positive fit effects. They do *not* replicate the finance validity chain: the continuous-control study did not use the same failure–repair–replication design, shares one reward model per task whose accuracy is validated on true rather than imagined transitions, and covers only one continuous-control ecosystem. MuJoCo is therefore treated as secondary transport evidence.

7 Discussion and Limitations

Why downstream evaluation is not sufficient. A downstream metric can still be non-diagnostic of the selector it evaluates. In the finance experiment, the realized-path endpoint supports a null comparison even though a true-dynamics expected-return selector does not clear the sensitivity diagnostic. The repaired endpoint passes the pre-specified information-sensitivity gate on a fresh block disjoint from the earlier diagnostic seeds, which makes the selector comparison interpretable under this controlled information-sensitivity test. The gate establishes sensitivity to the tested ladder, not universal validity or guaranteed power for arbitrary learned-model effects.

Scope of the selector conclusion. The selector result is conditional on the fixed candidate model families, policy pools, simulator, and expected-return estimand. Held-out one-step fit has no multiplicity-controlled detectable advantage over uniform model identity in this setting, but modest benefits remain compatible with the confidence intervals because no equivalence margin was specified. The result also does not show that predictive fit is intrinsically uninformative for decisions. The post-hoc family analysis leaves open whether the tested families contain enough decision-useful variation for any selector to obtain a large gain; resolving that mechanism would require a prospectively specified best-in-family or known-good-candidate experiment.

Limits of the sensitivity test. The sensitivity ladder is available because the simulator exposes the true data-generating mechanism. Real deployments rarely provide such an oracle. Other applications would need a domain-appropriate positive control whose expected direction is known independently of the selector. The sensitivity and selector analyses also share the same confirmatory population block; pre-specification prevents adaptive changes but does not provide sample independence, so a fully independent selector block would be a stronger design. The half-random threshold is likewise study-specific rather than universal; it prevents a monotone but negligible response from being treated as sufficient.

External validity. The finance environment is synthetic and should not be read as evidence about live trading performance, market efficiency, or deployable profitability. The MuJoCo results provide only a secondary transport check: three task-level selector events show that the audit can yield a positive fit effect, but they do not reproduce the full prospective validity chain and remain within one continuous-control ecosystem. Both finance and control experiments use finite model and policy

sets; transfer to different architectures, richer policy classes, or adaptive model-generation procedures remains untested.

Within these limits, information sensitivity changes how the selector result can be interpreted: the initial null cannot be attributed to the selector, whereas the later non-detection is reported under an endpoint that passes the sensitivity gate, without being treated as equivalence.

8 Conclusion

A downstream decision metric can support selector inference only if it responds to decision-relevant information. In the controlled finance experiment, the realized-path endpoint fails this requirement, whereas an expected-return repair passes the pre-specified sensitivity gate on a fresh replication block. Under that endpoint, held-out one-step fit shows no multiplicity-controlled detectable advantage over uniform model identity across the eight pre-specified slices. The finding is a scoped non-detection, not equivalence or a universal failure of predictive fit. Information sensitivity should therefore be checked before a negative selector result is attributed to the selection rule.

Reproducibility Statement

The main text specifies the nested-selector estimands, sensitivity gate, confirmatory population, and multiplicity rules. Appendix B–I reports the endpoint diagnostics, full sensitivity and selector results, simulator equations, model/policy definitions, RNG streams, and statistical procedures. Appendix J summarizes pre-specification and execution integrity, and Appendix K describes the supplementary files needed to reproduce the reported summaries.

AI Use Statement

Generative AI tools were used only for language polishing and assisting with experiment execution under human-defined protocols. All research ideas, hypotheses, methodological choices, experimental designs, interpretations, and conclusions were developed by the human authors. All AI-assisted outputs were reviewed and verified by the authors, who take full responsibility for the final submission.

References

Nathan Lambert, Brandon Amos, Omry Yadan, and Roberto Calandra. Objective mismatch in model-based reinforcement learning. In *Proceedings of the 2nd Conference on Learning for Dynamics and Control*, 2020.

Christopher Grimm, Andre Barreto, Satinder P. Singh, and David Silver. The value equivalence principle for model-based reinforcement learning. In *Advances in Neural Information Processing Systems*, 2020.

Alex Ayoub, Zeyu Jia, Csaba Szepesvari, Mengdi Wang, and Lin Yang. Model-based reinforcement learning with value-targeted regression. In *Proceedings of the 37th International Conference on Machine Learning*, 2020.

Yecheng Jason Ma, Kausik Sivakumar, Jason Yan, Osbert Bastani, and Dinesh Jayaraman. Learning policy-aware models for model-based reinforcement learning via transition occupancy matching. In *Proceedings of the 5th Annual Learning for Dynamics and Control Conference*, 2023.

Miles Hutson, Isaac Kauvar, and Nick Haber. Policy-shaped prediction: avoiding distractions in model-based reinforcement learning. In *Advances in Neural Information Processing Systems*, 2024.

Harry Amad and Mihaela van der Schaar. DT^2: Decision-targeted digital twins. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026.

Adam N. Elmachtoub and Paul Grigas. Smart "Predict, then Optimize". *Management Science*, 68(1):9–26, 2022.

Shaofeng Yin, Jialong Wu, Siqiao Huang, Xingjian Su, Xu He, Jianye Hao, and Mingsheng Long. Trajectory world models for heterogeneous environments. In *Proceedings of the 42nd International Conference on Machine Learning*, 2025.

Julian Quevedo, Ansh Kumar Sharma, Yixiang Sun, Varad Suryavanshi, Percy Liang, and Sherry Yang. WorldGym: World model as an environment for policy evaluation. In *International Conference on Learning Representations*, 2026.

Jiahan Zhang, Muqing Jiang, Nanru Dai, Taiming Lu, Arda Uzunoglu, Shunchi Zhang, Yana Wei, Jiahao Wang, Vishal M. Patel, Paul Pu Liang, Daniel Khashabi, Cheng Peng, Rama Chellappa, Tianmin Shu, Alan Yuille, Yilun Du, and Jieneng Chen. World-In-World: World models in a closed-loop world. In *International Conference on Learning Representations*, 2026.

Pai Liu, Lingfeng Zhao, Shivangi Agarwal, Jinghan Liu, Audrey Huang, Philip Amortila, and Nan Jiang. Model selection for off-policy evaluation: New algorithms and experimental protocol. In *Advances in Neural Information Processing Systems*, 2025.

Jialong Wu, Shaofeng Yin, Ningya Feng, and Mingsheng Long. RLVR-World: Training world models with reinforcement learning. In *Advances in Neural Information Processing Systems*, 2025.

Xuanlin Li, Kyle Hsu, Jiayuan Gu, Oier Mees, Karl Pertsch, Homer Rich Walke, Chuyuan Fu, Ishikaa Lunawat, Isabel Sieh, Sean Kirmani, Sergey Levine, Jiajun Wu, Chelsea Finn, Hao Su, Quan Vuong, and Ted Xiao. Evaluating real-world robot manipulation policies in simulation. In *Proceedings of the 8th Conference on Robot Learning*, 2025.

Dacheng Li, Yunhao Fang, Yukang Chen, Shuo Yang, Shiyi Cao, Justin Wong, Michael Luo, Xiaolong Wang, Hongxu Yin, Joseph E. Gonzalez, Ion Stoica, Song Han, and Yao Lu. WorldModelBench: Judging video generation models as world models. In *Advances in Neural Information Processing Systems*, 2025.

Yuta Saito and Masahiro Nomura. Hyperparameter optimization can even be harmful in off-policy learning and how to deal with it. In *Proceedings of the 33rd International Joint Conference on Artificial Intelligence*, pages 4860–4867, 2024.

Otmane Sakhi, Imad Aouali, Pierre Alquier, and Nicolas Chopin. Logarithmic smoothing for pessimistic off-policy evaluation, selection and learning. In *Advances in Neural Information Processing Systems*, 2024.

Matteo Zecchin, Sangwoo Park, and Osvaldo Simeone. Adaptive Learn-then-Test: Statistically valid and efficient hyperparameter selection. In *Proceedings of the 42nd International Conference on Machine Learning*, 2025.

Noah J. Schutte, Krzysztof S. Postek, and Neil Yorke-Smith. Robust losses for decision-focused learning. In *Proceedings of the 33rd International Joint Conference on Artificial Intelligence*, pages 4868–4875, 2024.

Lingkai Kong, Wenhao Mu, Jiaming Cui, Yuchen Zhuang, B. Aditya Prakash, Bo Dai, and Chao Zhang. DF^2: Distribution-free decision-focused learning. In *Proceedings of the 41st Conference on Uncertainty in Artificial Intelligence*, pages 2269–2290, 2025.

Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. In *Advances in Neural Information Processing Systems*, 2018.

Tuomas Oikarinen, Ge Yan, and Tsui-Wei Weng. Evaluating neuron explanations: A unified framework with sanity checks. In *Proceedings of the 42nd International Conference on Machine Learning*, 2025.

Anne Helby Petersen. Are you doing better than random guessing? A call for using negative controls when evaluating causal discovery algorithms. In *Proceedings of the 41st Conference on Uncertainty in Artificial Intelligence*, 2025.

International Council for Harmonisation. E10: Choice of control group and related issues in clinical trials. Harmonised Guideline, 2000.

Anton J. Kleywegt, Alexander Shapiro, and Tito Homem-de-Mello. The sample average approximation method for stochastic discrete optimization. *SIAM Journal on Optimization*, 12(2):479–502, 2002.

Christopher Grimm, Andre Barreto, Greg Farquhar, David Silver, and Satinder P. Singh. Proper value equivalence. In *Advances in Neural Information Processing Systems*, 2021.

Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979.

A Evidence Chain and Reporting Conventions

This appendix records the evidence underlying the main-text claims. The scientific statistical unit is the *population seed*. Unless stated otherwise, lower regret is better. "Raw CI" denotes the unadjusted paired-bootstrap 95% interval from the pre-specified analysis; multiplicity decisions use the prospectively specified Holm procedure. An interval containing zero is never interpreted as equivalence.

Table 1: Study-level evidence chain. The first repair sensitivity analysis remains formally MIXED; the later replication does not retroactively change that outcome.

Study phase	Question	Pre-specified outcome	Consequence
Original endpoint diagnostic	Can realized-path regret reward a true-dynamics expected-return selector?	INSENSITIVE	Initial selector null is non-diagnostic.
Repair sensitivity analysis	Does expected-return regret respond to increasing true-dynamics information?	MIXED	Five of six unit criteria pass, but the global rule is not met.
Fresh sensitivity replication	Does sensitivity replicate on fresh seeds under family-wise control?	PASS	Sensitivity gate clears; selector analysis proceeds.
Confirmatory selector audit	Does held-out one-step fit beat model-uniform calibration?	0/8 after Holm	No multiplicity-controlled detectable advantage within tested families.

B Original Endpoint Diagnostic

The original diagnostic scores a selector that uses true-dynamics expected-return information with the inherited *realized-path* normalized regret. The implementation invariant passed exactly (maximum absolute difference 0 over 800 seed-slices), but none of the eight slices passed the sensitivity criterion. Table 2 shows that the oracle remains close to random-policy regret, while every pre-specified oracle-versus-within-family-random-model-selector paired interval includes zero.

Table 2: Original endpoint diagnostic. The random-policy column is a descriptive reference. D_{orc} is the pre-specified oracle-minus-within-family-random-model-selector contrast under the old endpoint. "Policy share" is the median fraction of return variance attributable to policy rather than the realized future path.

Slice	Oracle mean regret	Random-policy ref.	D_{orc} vs model-uniform	95% CI	Policy share
5M / 12-pol / static	0.517	0.500	0.034	$[-0.047, 0.118]$	0.00265
5M / 12-pol / fixed	0.504	0.493	-0.051	$[-0.135, 0.030]$	0.00269
5M / 12-pol / replay	0.467	0.493	-0.028	$[-0.112, 0.053]$	0.00268
5M / 12-pol / live	0.439	0.494	-0.049	$[-0.131, 0.037]$	0.00358
8M / 12-pol / static	0.517	0.500	0.049	$[-0.034, 0.133]$	0.00265
8M / 12-pol / live	0.439	0.494	-0.040	$[-0.119, 0.043]$	0.00358
8M / 24-pol / static	0.493	0.500	0.011	$[-0.076, 0.104]$	0.00202
8M / 24-pol / live	0.442	0.500	-0.043	$[-0.133, 0.049]$	0.00257

The result is a property of the evaluator/model-objective pairing, not a theorem that true dynamics are useless. The selector optimizes an expected-return quantity while the old endpoint normalizes regret by one realized future path; the observed policy main-effect shares are only about 0.2–0.4%.

C Expected-Return Sensitivity Analysis

The repair sensitivity analysis replaces the target with expected-return regret and does not inspect the selector outcome. Numerical precision reaches the pre-specified requirement at $K_{\mathrm{eval}} = 80{,}000$ in all six units.

Table 3: Repair sensitivity-analysis outer-Monte-Carlo precision diagnostic. Entries are the pre-specified median paired precision ratios.

Unit	10k	20k	40k	80k
static / 12-policy	0.154	0.109	0.077	**0.055**
static / 24-policy	0.228	0.161	0.114	**0.081**
fixed / 12-policy	0.145	0.106	0.076	**0.054**
replay / 12-policy	0.144	0.106	0.075	**0.054**
live / 12-policy	0.121	0.087	0.062	**0.044**
live / 24-policy	0.181	0.128	0.092	**0.065**

Table 4: Repair sensitivity-analysis information ladder. Regret decreases with increasing K in every unit.

Unit	K=1	10	50	200	1000	Random	Median Spearman
static / 12-policy	0.503	0.478	0.453	0.376	0.190	0.500	-0.90
static / 24-policy	0.502	0.491	0.473	0.424	0.258	0.500	-0.90
fixed / 12-policy	0.644	0.577	0.480	0.339	0.156	0.526	-1.00
replay / 12-policy	0.643	0.575	0.472	0.345	0.150	0.519	-1.00
live / 12-policy	0.573	0.446	0.343	0.234	0.136	0.430	-1.00
live / 24-policy	0.419	0.357	0.310	0.256	0.189	0.423	-1.00

All six K=1000-versus-random intervals lie below zero, and five of six units pass all three per-unit criteria. The static / 24-policy unit misses only the half-random criterion ($0.258 > 0.250$). Nevertheless, the pre-specified global outcome is MIXED: the label-permutation control in the fixed / 12-policy unit has an unadjusted interval $[0.003, 0.029]$, narrowly excluding zero under the original "all six intervals must contain zero" rule. The selector analysis was therefore not run on this block. The later family-wise rule was specified prospectively for the fresh replication and does not change the repair-analysis outcome.

D FRESH INFORMATION-SENSITIVITY REPLICATION

The fresh replication uses population seeds 4001–4100, with no outcome-dependent exclusion or replacement. The endpoint, policy pools, information ladder, $K_{\text{eval}} = 80{,}000$, 20 inner repetitions, sensitivity criteria, and anti-oracle control are unchanged from the repair sensitivity analysis. The only methodological change is the prospectively specified family-wise rule for the label-permutation control.

Table 5: Fresh sensitivity replication. "Prec." is the median paired outer-MC precision ratio at $K_{\text{eval}} = 80{,}000$. Perm. ctrl. p_H is the Holm-adjusted label-permutation p-value.

Unit	Prec.	K=1	K=1000	Random	$K1000 - \text{rnd}$ 95% CI	Spearman	Half-rnd	Perm. ctrl. p_H	Pass
static / 12-policy	0.0547	0.505	0.196	0.500	$[-0.314, -0.292]$	-0.90	yes	1.00	yes
static / 24-policy	0.0808	0.502	0.262	0.500	$[-0.249, -0.228]$	-0.90	no	1.00	no
fixed / 12-policy	0.0547	0.640	0.171	0.529	$[-0.368, -0.347]$	-1.00	yes	1.00	yes
replay / 12-policy	0.0537	0.643	0.160	0.517	$[-0.368, -0.347]$	-1.00	yes	1.00	yes
live / 12-policy	0.0447	0.570	0.139	0.433	$[-0.302, -0.287]$	-1.00	yes	1.00	yes
live / 24-policy	0.0659	0.429	0.192	0.426	$[-0.242, -0.226]$	-0.90	yes	1.00	yes

The replication passes the global gate: precision passes 6/6; sensitivity passes 5/6; there are zero reversed information trends; no label-permutation control rejects after Holm; and the anti-oracle control is worse than random in all six units. The repeated 24-policy static miss (0.258 in the repair analysis; 0.262 in the replication) is retained rather than tuned away.

E CONFIRMATORY SELECTOR AUDIT

The selector audit uses the same fresh block and is examined only after the sensitivity replication passes the pre-specified gate. The primary quantity is the population-seed mean $D^{\exp}_{\text{random}} = R^{\exp}_{\text{fit}} - R^{\exp}_{\text{unif}}$. Negative values favor held-out fit. Paired bootstrap uses $B = 10{,}000$ and analysis seed 90001; Holm correction is applied across the eight pre-declared primary means.

Table 6: Confirmatory selector analysis. Raw CIs are unadjusted paired-bootstrap 95% intervals; the inferential decision uses Holm-adjusted p_H.

Slice	Mean $D^{\exp}_{\text{random}}$	Raw 95% CI	Raw p	p_H	R_{fit}	R_{unif}
5M / 12-pol / static	-0.0227	$[-0.0452, -0.0016]$	0.0342	0.2736	0.4256	0.4483
5M / 12-pol / fixed	-0.0237	$[-0.0680, 0.0192]$	0.2724	0.8171	0.5648	0.5886
5M / 12-pol / replay	-0.0363	$[-0.0809, 0.0085]$	0.1156	0.4740	0.4914	0.5277
5M / 12-pol / live	-0.0450	$[-0.0966, 0.0085]$	0.0948	0.4740	0.4637	0.5086
8M / 12-pol / static	-0.0244	$[-0.0495, -0.0017]$	0.0346	0.2736	0.4390	0.4634
8M / 12-pol / live	-0.0096	$[-0.0680, 0.0511]$	0.7551	1.0000	0.4686	0.4782
8M / 24-pol / static	-0.0219	$[-0.0466, 0.0004]$	0.0554	0.3324	0.4470	0.4689
8M / 24-pol / live	-0.0059	$[-0.0512, 0.0420]$	0.7883	1.0000	0.4076	0.4135

Two raw intervals exclude zero, but neither survives the pre-specified multiplicity correction. The confirmatory summary is therefore 0/8 detectable fit advantages after Holm. All eight point estimates are negative, but the slices share seeds and overlapping model families; no aggregate sign test was pre-specified.

Table 7: Selector-audit secondary diagnostics. $D^{\exp}_{\text{class}}$ averages uniformly over unique induced-policy classes rather than model identities; its intervals are unadjusted and secondary.

Slice	Median $D^{\exp}_{\text{random}}$	Median raw CI	Mean $D^{\exp}_{\text{class}}$	D_{class} raw CI	Mean classes	Median ρ(fit, regret)
5M / 12-pol / static	0.0000	$[-0.0041, 0.0012]$	-0.0324	$[-0.0595, -0.0068]$	2.67	-0.107
5M / 12-pol / fixed	-0.0047	$[-0.0674, 0.0088]$	-0.0411	$[-0.0977, 0.0124]$	2.51	0.000
5M / 12-pol / replay	-0.0337	$[-0.1040, 0.0000]$	-0.0656	$[-0.1242, -0.0063]$	2.59	0.000
5M / 12-pol / live	-0.0259	$[-0.1107, 0.0122]$	-0.0425	$[-0.1040, 0.0211]$	2.59	0.000
8M / 12-pol / static	-0.0001	$[-0.0059, 0.0071]$	-0.0275	$[-0.0572, -0.0002]$	3.38	-0.052
8M / 12-pol / live	-0.0338	$[-0.0732, 0.0080]$	-0.0297	$[-0.0935, 0.0376]$	3.31	0.000
8M / 24-pol / static	0.0013	$[-0.0075, 0.0134]$	-0.0265	$[-0.0535, -0.0015]$	3.36	-0.062
8M / 24-pol / live	-0.0336	$[-0.0658, 0.0080]$	-0.0420	$[-0.0923, 0.0097]$	3.39	0.056

Within-class regret is identical in every analyzed seed, as implied by decision equivalence. Four of eight unadjusted D_{class} intervals exclude zero, but no multiplicity-controlled confirmatory rule was pre-specified for this secondary family. We use this analysis only to expose sensitivity to model-identity multiplicity.

F CANDIDATE-FAMILY HEADROOM AND MUJOCO TRANSPORT

Post-hoc best-in-family and order-statistic diagnostic. After observing the selector analysis, we computed a hindsight best-in-family diagnostic from the existing per-seed outputs only; this analysis is exploratory and does not alter the confirmatory $0/8$ selector result. For each population seed, R_{best} is the minimum expected regret among the policies induced by the fixed candidate models. The raw selector opportunity is $R_{\text{fit}} - R_{\text{best}}$ and raw family headroom is $R_{\text{random policy}} - R_{\text{best}}$. Because taking a minimum over multiple candidates creates headroom by order statistics alone, we compare R_{best} with exact random-policy order-statistic references.

Table 8: Post-hoc family geometry. "Sel. opp." is $R_{\text{fit}} - R_{\text{best}}$. $n_{=0}$ counts seeds where held-out fit already selects a hindsight-best family candidate; $n_{>0.1}$ counts large misses. Lower regret is better.

Slice	$\bar{R}_{\text{fit}}$	$\bar{R}_{\text{best}}$	Random	Headroom	Sel. opp. mean / median	$n_{=0}$	$n_{>0.1}$
5M / 12-pol / static	0.426	0.350	0.500	0.150	0.076 / 0.000	51	17
5M / 12-pol / fixed	0.565	0.346	0.529	0.183	0.219 / 0.000	54	40
5M / 12-pol / replay	0.491	0.284	0.517	0.233	0.207 / 0.000	57	38
5M / 12-pol / live	0.464	0.221	0.433	0.212	0.243 / 0.010	50	41
8M / 12-pol / static	0.439	0.335	0.500	0.165	0.104 / 0.017	36	24
8M / 12-pol / live	0.469	0.215	0.433	0.218	0.254 / 0.051	42	42
8M / 24-pol / static	0.447	0.352	0.500	0.148	0.095 / 0.023	42	27
8M / 24-pol / live	0.408	0.221	0.426	0.206	0.187 / 0.052	36	38

Selector opportunity is right-skewed: the 25th percentile is zero in every slice, held-out fit selects a hindsight-best candidate in 36–57 of 100 seeds, and misses larger than 0.1 occur in 17–42 seeds. Thus the positive mean opportunity should not be read as a uniform per-seed deficit.

Table 9: Post-hoc order-statistic references. Entries are mean excess headroom, defined as reference regret minus R_{best}, with paired-bootstrap 95% CIs over population seeds. Positive values would mean the fixed family best outperforms the random-choice reference. M is the number of model identities; C is the number of distinct induced decision classes. The with-replacement C reference is family-favoring; the without-replacement reference conditions on C distinct random policies.

Slice	$\bar{C}$	Best-of-M	Best-of-C (repl.)	Best-of-C (no repl.)
5M / 12-pol / static	2.67	$-0.162\,[-0.201, -0.121]$	$-0.037\,[-0.072, +0.000]$	$-0.054\,[-0.090, -0.018]$
5M / 12-pol / fixed	2.51	$-0.195\,[-0.242, -0.149]$	$-0.045\,[-0.085, -0.004]$	$-0.065\,[-0.106, -0.025]$
5M / 12-pol / replay	2.59	$-0.138\,[-0.183, -0.092]$	$-0.004\,[-0.045, +0.040]$	$-0.025\,[-0.066, +0.019]$
5M / 12-pol / live	2.59	$-0.141\,[-0.171, -0.111]$	$-0.023\,[-0.050, +0.005]$	$-0.043\,[-0.070, -0.015]$
8M / 12-pol / static	3.38	$-0.228\,[-0.267, -0.186]$	$-0.069\,[-0.104, -0.033]$	$-0.093\,[-0.128, -0.058]$
8M / 12-pol / live	3.31	$-0.178\,[-0.207, -0.148]$	$-0.060\,[-0.085, -0.035]$	$-0.082\,[-0.107, -0.058]$
8M / 24-pol / static	3.36	$-0.194\,[-0.230, -0.156]$	$-0.041\,[-0.072, -0.008]$	$-0.051\,[-0.082, -0.018]$
8M / 24-pol / live	3.39	$-0.138\,[-0.162, -0.114]$	$-0.030\,[-0.051, -0.009]$	$-0.040\,[-0.060, -0.018]$

No slice shows positive excess headroom with a 95% interval above zero. Excess headroom is below zero in 8/8 slices against best-of-M, 5/8 against best-of-C with replacement, and 7/8 against best-of-C without replacement; the remaining intervals include zero. Thus the raw hindsight headroom is explainable by order statistics alone and is consistent with limited decision enrichment of these fixed families. Because both the best-in-family diagnostic and the order-statistic references are post hoc, this does not establish a family-capacity mechanism. A prospective known-good-candidate or family-capacity experiment would be required to establish that mechanism.

Secondary continuous-control transport check. The earlier MuJoCo experiment uses the same nested model-selector $\rightarrow$ policy-selector regret object, but it does not pass through the full prospective sensitivity-gating chain. It is therefore transport evidence rather than replication.

Table 10: MuJoCo transport check ($n = 100$ reset seeds per task). Negative differences favor held-out fit.

Task	Mean D_{random} (95% CI)	Median D_{random} (95% CI)
HalfCheetah-v5	$-0.0515\,[-0.0717, -0.0309]$	$-0.0877\,[-0.1016, -0.0758]$
Hopper-v5	$-0.2761\,[-0.2988, -0.2529]$	$-0.2782\,[-0.3029, -0.2493]$
Walker2d-v5	$-0.1612\,[-0.2196, -0.1006]$	$-0.2209\,[-0.2920, -0.1736]$

The transport study contains only three upstream task-level selections. Its shared reward predictor is checked on true transitions, not separately on imagined out-of-distribution states. Accordingly, the result is used only to show that the broader audit machinery can yield positive selector evidence.

G Controlled Finance Simulator and Return

The synthetic environment is a controlled prediction-to-decision testbed rather than a market-realism claim. Horizon is $H = 50$. Background price noise follows an AR(1) process with coefficient $\phi = 0.4$ and stationary marginal variance one. Permanent impact has latent state

$$h_t = \delta h_{t-1} + \gamma f_t, \qquad dp_{\mathrm{mid},t} = \epsilon_t + (h_t - h_{t-1}), \tag{12}$$

with

$$\delta = \exp[-\log(2)/5] \approx 0.8706, \qquad \gamma = 0.06.$$

The four regimes differ only in total flow:

$$f_t = \begin{cases} 0, & \text{static/no impact,} \\ x_t, & \text{fixed impact,} \\ x_t + \mathrm{replay_flow}_t, & \text{replay/broken feedback,} \\ x_t + \left(\sum_{i=1}^{5} g_i\right) \mathrm{mean}(x_{t-9:t}), & \text{live aggregate feedback.} \end{cases} \tag{13}$$

Replay flow is generated once against a fixed TWAP-sell reference and then held fixed across evaluated candidate policies. In the live regime, the aggregate response is recomputed from the evaluated policy's recent signed trades.

The downstream objective is self-financing. Let $M_t = \sum_{u=t}^{H-1} dp_{\mathrm{mid},u}$ be the remaining midprice move from trade time t to horizon end. Then

$$J(x, dp_{\mathrm{mid}}) = \sum_{t=0}^{H-1} x_t M_t - \eta \sum_{t=0}^{H-1} x_t^2, \qquad \eta = 0.105. \tag{14}$$

Targeted round-trip property tests are non-positive in the implemented simulator, including 0 positive outcomes in a 600-trial randomized sweep. This is an implementation check, not a theorem that the impact model is manipulation-free in general.

H Candidate Models and Policy Pools

Synthetic candidate model families. The five-model family is the original sklearn-only family; the pooled family adds three sequence models.

Table 11: Candidate-model definitions used by the finance audit.

Reporting label	Implementation role	Specification
AR1	linear baseline	synthetic feature index [0] (lag-1 own trade)
LR-flow	linear flow model	three-lag flow linear regression
LR-flow-dp	linear flow+target model	six-feature linear regression
MLP-small	sklearn MLP	one hidden layer (8), max_iter 800
MLP-big	sklearn MLP	hidden layers (16,8), max_iter 800
GRU-small	sequence model	1 layer, hidden size 8, linear head
GRU-big	sequence model	2 layers, hidden size 32, linear head
Transformer-tiny	sequence model	2 layers, 2 heads, $d_{\mathrm{model}} = 16$, FF 32, dropout 0

The sklearn MLPs use Adam with learning rate 0.001, $\alpha = 10^{-4}$, batch size "auto", and random state 0. Neural models use MSE loss, Adam with learning rate 10^{-2}, full-batch updates, and 60 epochs. The same trained model instances are used throughout the selector analysis.

Policy pools. The original 12-policy pool combines signed TWAP, front/back-loaded execution, momentum, and mean-reversion policies. The expanded pool contains all original policies plus slower/larger variants.

Table 12: Policy-family parameterization. The fixed policy specification defines the exact signs and ordering used in the implementation.

Family	Original-pool settings	Additional expanded-pool settings
TWAP	signed totals ± 12, ± 20 over H	signed totals ± 6, ± 30
Front/back loaded	signed total ± 12, active over 20% of H	signed total ± 12, active over 40% of H
Momentum	base magnitude 0.3, lookback 5	lookback 15 variants
Mean reversion	sign-reversed momentum rule, lookback 5	lookback 15 variants

For a stateful momentum policy,

$$x_t = b \operatorname{sign}(\bar{d}_t) \min(1, 3|\bar{d}_t| + 0.1),$$

where $\bar{d}_t$ is the recent mean target increment; mean reversion reverses the sign. During model rollouts, $\bar{d}_t$ is computed from model-generated target history. Rollouts are recursive: after initialization, state history and policy actions depend on previous model predictions rather than the true simulator history.

I Statistical Design, RNG, and Multiplicity

Table 13: Pre-specified confirmatory statistical design.

Component	Setting
Population block	seeds 4001–4100; full block used; no replacement/exclusion
Seed disjointness	zero overlap with earlier diagnostic ranges 0–5, 10–14, 101–210, 1001–1100, 3001–3100
Scientific unit	population seed
Outer estimator	$K_{\text{eval}} = 80{,}000$ common-random-number trajectories shared across policies
Information ladder	$K \in \{1, 10, 50, 200, 1000\}$ from an independent RNG stream
Inner repetitions	20 per seed and ladder level
Permutation control	two-sided sign-flip randomization, $B = 10{,}000$, Holm across 6 units
Anti-oracle control	argmin anti-oracle; must be worse than random in all units
Primary selector statistic	mean $D^{\text{exp}}_{\text{random}}$, paired percentile bootstrap $B = 10{,}000$, seed 90001
Selector multiplicity	Holm across 8 pre-declared primary slices
Secondary	medians, D_{class}, absolute regret, class count/entropy, top-2 margin, fit-regret association
Stopping rule	A mixed/failed sensitivity result stops analysis before selector outcomes are examined; no K_{eval} escalation or outcome-dependent seed changes

The outer estimator, selector evaluation, and label-permutation control use deterministic RNG streams keyed by arm, seed, and range. Repeated execution therefore reproduces the same generated values for the fresh block 4001–4100.

J Pre-Specification and Execution Integrity

Before the confirmatory selector outcomes were examined, we fixed the population seed block, endpoint, candidate model families, policy pools, sensitivity criteria, multiplicity procedures, and analysis code. A mixed or failed sensitivity result would have stopped the analysis before the selector outcome was inspected. No threshold, seed, endpoint, candidate family, or decision rule was changed after outcomes became available.

The sensitivity and selector analyses use the same population block, so pre-specification protects against adaptive analysis changes but does not create sample independence. A fully independent selector block would provide a stronger design. Machine-readable per-seed outputs and the analysis code used to produce the reported summaries are included with the supplementary materials.

K SUPPLEMENTARY REPRODUCIBILITY MATERIALS

The supplementary package contains the machine-readable per-seed diagnostic and selector outputs, expected-value estimates, sensitivity-ladder results, exact model and policy specifications, and the analysis code needed to reproduce the tables and figures in the paper. The manuscript reports rounded summaries; the supplementary machine-readable files retain the full numerical precision used in the analyses.