When Is Reuse Superiority Licensed? Evidence Separation for Heterogeneous Offline Experience Reuse

Anonymous authors
Paper under double-blind review

Abstract

Comparisons of heterogeneous offline experience reuse often collapse distinct questions into a single statement that one method is "better": average outcome, target-wise benefit, source-selection quality, and deployed behavior are treated as interchangeable evidence. We study when such comparative claims are warranted. We introduce an executable, claim-specific evidence-separation protocol that first specifies what "better" means—including the estimand, population, reference, measurement contract, comparison budget, uncertainty unit, deployment contract, and scope—and then invokes the evidence required by that claim. In controlled finite-MDP problems with known answers, stronger audit configurations reduce wrong directional claims at matched claimed-decision coverage, while a critic-local forced-reuse score shows weak fine-grained agreement with exact downstream consequences but retains coarse consequence information. The result is a claim-licensing methodology for heterogeneous reuse, not a universal controller or a validated prospective harm predictor.

1 Introduction

Offline reinforcement learning (RL) learns from fixed experience and must therefore decide which data support a target policy (Levine et al., 2020; Fu et al., 2020). In multi-task and transfer settings, experience from one source may help one target while interfering with another, motivating conservative sharing, target-aware filtering, modular transfer, and routing (Yu et al., 2021; Yang et al., 2020; Sun et al., 2020; Bai et al., 2024; Wen et al., 2024). Yet heterogeneous-reuse comparisons often compress distinct scientific questions into one statement that a method is "better." Higher average return need not imply target-wise benefit; accepting less source experience can make a selector appear safer; many reuse events need not be independent replications; and a favorable process diagnostic need not imply a better deployed policy.

The issue is not simply that more metrics should be reported. A *mean outcome claim*, a *target-wise benefit claim*, a *selector quality claim*, a *mechanism attribution claim*, and a *deployment performance claim* ask different questions and therefore require different evidence contracts. Matching reuse quantity is important when isolating selector quality, for example, but need not be imposed when comparing complete deployment policies whose reuse rate is itself part of the method. Likewise, evidence about selected reuse events does not by itself establish a claim about future rollout performance.

We therefore ask: *when are comparative claims in heterogeneous offline experience reuse warranted by the available evidence?* Our answer is *claim-specific evidence separation*. As summarized in Figure 1, a comparison first specifies what "better" means and its intended scope, then invokes the evidence contract associated with that claim. The five branches are alternative claim types rather than sequential stages. Inference is performed at the scientific independence unit—such as a policy seed, source–target pair, or controlled problem seed—and the audit returns a bounded conclusion: Supported, Conditional, Unresolved, or Not Supported. Crucially, this makes evidence separation empirically testable rather than merely prescriptive: when the correct comparative direction is known, we can ask whether stronger audit configurations actually reduce wrong directional claims at matched claimed-decision coverage.

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.

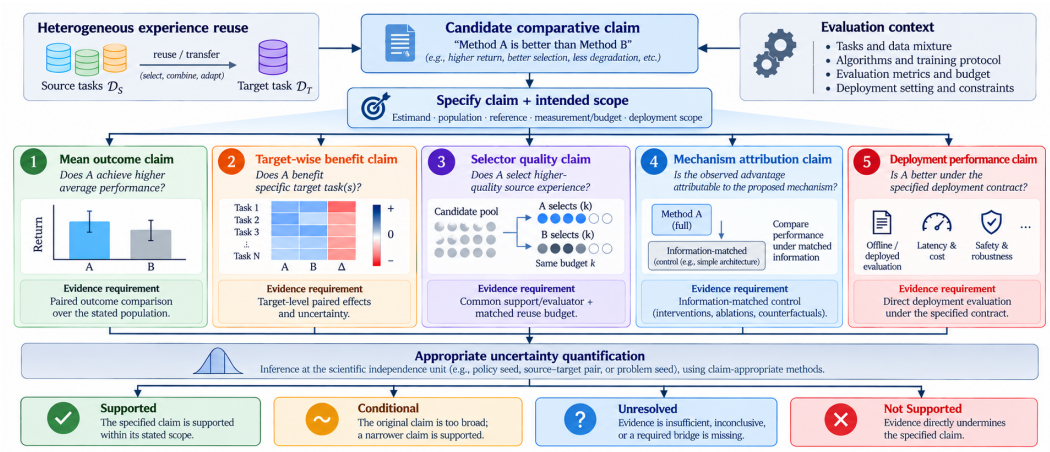


Figure 1: **Claim-specific evidence separation.** A comparison specifies what "better" means, invokes the evidence contract appropriate to that claim, performs inference at the relevant scientific independence unit, and licenses only the conclusion supported within its stated scope. The five branches are alternative claim types, not sequential stages.

We evaluate the framework in controlled finite-MDP problems with known answers. At matched claimed-decision coverage, progressively stronger audit configurations reduce wrong directional conclusions at several operating points, whereas the uncertainty-only configuration provides no additional discrimination at those same coverages. The same setting also exposes the limits of a critic-local forced-reuse diagnostic: fine-grained agreement with exact downstream consequences is weak, but high-score strata consistently exhibit worse exact outcomes, with a representative high–low exact-effect difference of -0.0236. Thus, the measurement retains coarse consequence information under the tested contracts rather than supporting precise event-level prediction.

Meta-World then serves as a claim-adjudication case study. Mean outcome does not establish target-wise benefit; pair-level inference reveals concentrated selector effects; an information-matched simple control weakens architectural attribution; and alternative measurement specifications alter fine-grained rankings. ManiSkill3 provides a second case in which process measurements, rollout outcomes, and target-wise degradation need not license the same conclusion, while D4RL supplies an additional diagnostic example outside Meta-World.

Our contributions are: **(i)** an executable, claim-specific audit protocol that maps comparative claims to the evidence contracts associated with their stated scope; **(ii)** known-answer validation showing measurable audit utility while exposing the limits of local process measurements; and **(iii)** claim-transition evidence across Meta-World, ManiSkill3, and D4RL showing that outcome, selection, mechanism, and deployment evidence can support different conclusions.

2 Related Work and Problem Formulation

Heterogeneous sharing and offline RL. Multi-task and transfer learning study when representations, parameters, or gradients should be shared across tasks (Caruana, 1997; Pan and Yang, 2010; Wang et al., 2019; Misra et al., 2016; Zamir et al., 2018; Standley et al., 2020; Yu et al., 2020a; Liu et al., 2021). Multi-task RL and modular routing similarly regulate cross-task transfer (Teh et al., 2017; Rosenbaum et al., 2018; Yang et al., 2020). Offline RL adds distribution shift and extrapolation error, motivating behavior constraints, conservative or pessimistic value estimation, and implicit regularization (Fujimoto et al., 2019; Kumar et al., 2019; 2020; Kidambi et al., 2020; Yu et al., 2020c; Fujimoto and Gu, 2021; Kostrikov et al., 2022). Selective sharing methods further show that source experience can help some targets while harming others (Yu et al., 2021; Bai et al., 2024; Wen et al., 2024; Qiao et al., 2025; Liu and Long, 2026). We do not claim to originate selective reuse. Our question is how comparative claims about such methods should be evaluated once "better" can refer to mean outcome, target-wise benefit, selector quality, mechanism attribution, or deployment performance.

Reliable comparison and selective decision-making. RL comparisons are sensitive to random seeds, aggregation choices, and the unit at which uncertainty is estimated (Henderson et al., 2018; Agarwal et al., 2021). Selective prediction studies the trade-off between issuing decisions and their conditional error (Geifman and El-Yaniv, 2017), while calibration and predictive uncertainty address related but distinct questions (Guo et al., 2017; Lakshminarayanan et al., 2017). Model-selection work further warns that repeatedly optimizing noisy evaluation criteria can introduce selection bias (Cawley and Talbot, 2010). Our contribution is not a new abstention rule or a general theory of risk–coverage trade-offs. We adapt these ideas to heterogeneous reuse, where the scientific claim determines the associated evidence contract. In particular, we distinguish `reuse_coverage`, the fraction of candidate source events accepted in a selector comparison, from `claimed_decision_coverage`, the fraction of comparison instances on which the audit issues a directional research claim. Matching the former controls selection opportunity; matching the latter tests whether a stricter audit improves claim quality rather than merely issuing fewer claims.

Problem formulation. For methods A and B, we represent a candidate comparative claim as

$$\mathcal{C} = (A, B, \theta, \mathcal{P}, R, m, b, u, d, s), \tag{1}$$

where θ is the estimand, $\mathcal{P}$ the evaluation population, R the reference comparator, m the measurement contract, b the comparison budget, u the scientific independence unit, d the deployment contract, and s the intended claim scope. These dimensions are an operational decomposition chosen to capture the reuse-specific identification failures studied here; they are not intended as an exhaustive representation of all possible machine-learning claims.

The reference R specifies what the claim is evaluated against—for example, a no-reuse baseline, a matched-budget selector, an information-matched architectural control, or another deployed policy. The measurement contract m specifies *how* the quantity of interest is measured, such as the loss or Bellman-family definition used by an operational diagnostic. The comparison budget b instead specifies *what opportunity is held comparable*, such as the accepted reuse count or selection quota. This distinction separates changes in the measurement itself from changes in the amount of reuse made available to a selector.

The claim type determines which components enter the corresponding evidence contract. For example, a selector-quality claim requires a comparable selection opportunity and measurement contract, whereas a deployment-performance claim requires direct evaluation under the specified deployment contract. Given evidence $\mathcal{D}$, the objective is not to produce a global ranking of A and B, but to determine the strongest scope of $\mathcal{C}$ warranted by $\mathcal{D}$. Section 3 operationalizes this problem as a claim-specific audit returning SUPPORTED, CONDITIONAL, UNRESOLVED, or NOT SUPPORTED. The resulting procedure is an operational evidence contract rather than a theorem of necessity or sufficiency; Section 4 evaluates its decision utility in settings where the relevant answers are known.

3 AN EXECUTABLE CLAIM-AUDIT PROTOCOL

Starting from $\mathcal{C}$ in equation 1, the audit identifies the claim type, records the evaluated evidence contract, applies a specified decision rule at the scientific independence unit, and returns a bounded conclusion. These operational contracts address the reuse-specific failures studied here; they are not a necessary or exhaustive theory of comparative evidence.

Claim-specific evidence contracts. Table 1 summarizes how the five claim types in Figure 1 are instantiated in this paper. The contracts are deliberately asymmetric: evidence supporting one type of statement is not automatically promoted to a stronger statement at another level.

Process and deployment evidence remain distinct unless an experiment directly evaluates their bridge. An unevaluated bridge is reported as unresolved, not as positive or negative evidence.

Claim-specific decision rules. To make status assignment reproducible, each reported audit is paired with an explicitly specified decision rule

$$\delta_{\mathcal{C}}(\mathcal{D}) \in \{\text{SUPPORT}, \text{INCONCLUSIVE}, \text{CONTRADICT}\}, \tag{2}$$

where $\mathcal{D}$ denotes the evidence evaluated for claim $\mathcal{C}$. The instantiated rule specifies the comparison quantity, direction of the claim, scientific independence unit, and decision criterion used in that

Table 1: **Evidence contracts instantiated in this paper.** The middle column records the evidence used for each claim type; the final column records an interpretation that this evidence does not establish by itself.

Claim type	Evidence instantiated in this paper	Scope not established by that evidence alone
Mean outcome	Paired aggregate performance under the stated evaluation population and reference.	Benefit for every target.
Target-wise benefit	Target-level outcomes or deficits relative to the same reference, with uncertainty at the scientific independence unit used for that analysis.	Selector quality or mechanism attribution.
Selector quality under m	Process comparisons under the specified measurement contract m, with reuse quantity matched when the analysis is intended to separate selection quality from reuse amount.	Measurement-contract-invariant selector quality or superiority of the deployed system.
Mechanism attribution	Information-matched controls used to test whether the observed audit-time advantage requires the proposed architectural mechanism.	Causal sufficiency of that mechanism or deployment superiority.
Deployment performance	Direct rollout evaluation under the tested deployment contract d.	Retrospective attribution of the outcome to selector quality or mechanism.

Table 2: **Controlled audit configurations at matched claim coverage.** Let d_j and s_j be specification j's estimated difference and pair-based standard error, $r_j = (|d_j| - \epsilon)/\max(s_j, 10^{-12})$, and Q_j its frozen qualification flag. Rank eligible items by decreasing score and retain the required claim count. Direction is the sign of d_1; P3 also requires agreement.

Config.	Eligibility	Ranking score
P0	All candidate items	$\lvert d_1\rvert/\max(s_1, 10^{-12})$
P1	All candidate items	r_1
P2	Q_1	r_1
P3	$Q_1 \wedge Q_2 \wedge [\mathrm{sign}(d_1) = \mathrm{sign}(d_2)]$	$\min(r_1, r_2)$

analysis. The corresponding experiment section states the rule actually used; for the controlled benchmark, the rules are inherited from the frozen P0–P3 implementation. This explicit specification is a reproducibility device and should not be interpreted as prospective preregistration where the underlying provenance does not support that claim.

Audit outcomes and scope restriction. For a candidate claim C, the claim-specific decision rule $\delta_C(D)$ is evaluated at the designated scientific independence unit. Support at the stated scope yields SUPPORTED. If that scope is not licensed but an explicitly evaluated narrower claim is supported, the audit returns CONDITIONAL together with the narrower scope. Direct contradiction yields NOT SUPPORTED only when no evaluated narrower claim is licensed. All remaining cases, including inconclusive uncertainty, missing evidence, an unevaluated bridge, or an incomplete experiment, are reported as UNRESOLVED. Narrower claims are restricted to contracts, populations, operating points, or evidence domains that were actually evaluated; the audit does not search unrestricted post-hoc subsets.

Controlled audit configurations. Section 4 evaluates P0–P3, validation configurations rather than the five claim types in Figure 1. Their eligibility sets are nested, but P1 changes the uncertainty-based ranking rather than imposing an additional gate in the matched-coverage sweep. Table 2 specifies that sweep; Appendix B distinguishes it from the native fixed-threshold decisions.

Because qualification can reduce the number of issued directional claims, raw error rates at native coverage are not directly comparable. We therefore compare P0–P3 at matched *claimed-decision coverage*: the fraction of controlled decision items on which a directional claim is issued. This controls for abstention, but not by itself for differences in available information or eligibility rules.

The controlled study therefore evaluates a deliberately narrow proposition: whether the implemented P0–P3 configurations differ in directional wrong-claim rate at matched claimed-decision coverage. It does not establish that every audit component is individually necessary, that the four reporting outcomes are universally calibrated, or that the same configuration ordering must hold outside the controlled benchmark.

4 CONTROLLED VALIDATION

We evaluate the audit where incorrect conclusions are observable. In finite-MDP problems, both the process comparison used to score audit decisions and an independent downstream policy consequence can be computed exactly. The benchmark contains 1,200 held-out problem instances with development-disjoint seeds. Each problem is evaluated in six observation regimes, giving 7,200 decision items per configuration; the two skill levels contain 600 problems each. Exact Bellman and forced-reuse calculations agree to numerical precision (maximum residuals 4.32×10^{-15} and 5.33×10^{-15}). Oracle quantities are used only for evaluation. Generator details, additional stress regimes, arithmetic checks, and provenance are in Appendix B.

The benchmark addresses two distinct questions: *(i)* do stronger implemented P0–P3 audit configurations reduce incorrect directional claims at matched claimed-decision coverage; and *(ii)* does the critic-local forced-reuse measurement contain information about an independent exact downstream consequence? The first tests audit decision utility; the second bounds the meaning of one measurement used within the audit.

4.1 AUDIT UTILITY AT MATCHED CLAIMED-DECISION COVERAGE

For event e, define the exact forced-reuse degradation and accepted severity as

$$\Delta_e^\star = (q_e^+ - y_e)^2 - (q_e^- - y_e)^2, \tag{3}$$

$$S_{\mathrm{URR}}^\star = \frac{1}{|\mathcal{A}|} \sum_{e \in \mathcal{A}} \max(\Delta_e^\star, 0), \tag{4}$$

where each selector accepts its top-k events at the frozen reuse budget 50%. For methods A and B,

$$D_{A-B}^\star = S_{\mathrm{URR},A}^\star - S_{\mathrm{URR},B}^\star. \tag{5}$$

Because lower severity is better, the oracle label is A-BETTER if $D_{A-B}^\star < -\epsilon$, B-BETTER if $D_{A-B}^\star > \epsilon$, and PRACTICALLY-UNRESOLVED otherwise. The frozen tolerance is $\epsilon = 1.125378 \times 10^{-3}$, obtained on development problems as 0.05 times the median $|D^\star|$ with floor 10^{-6}. Among the 1,200 held-out problems, 615 are A-BETTER, 553 B-BETTER, and 32 PRACTICALLY-UNRESOLVED.

Let $a_k(n) \in \{-1, 0, +1\}$ denote the direction issued by configuration P_k, with 0 denoting abstention, and let y_n denote the corresponding oracle label, with 0 denoting PRACTICALLY-UNRESOLVED. We define

$$c_k = \frac{1}{N} \sum_n \mathbb{1}[a_k(n) \neq 0], \tag{6}$$

$$e_k = \frac{\sum_n \mathbb{1}[a_k(n) \neq 0]\mathbb{1}[y_n \neq 0]\mathbb{1}[a_k(n) \neq y_n]}{\sum_n \mathbb{1}[a_k(n) \neq 0]}, \tag{7}$$

where $N = 7,200$, c_k is `claimed_decision_coverage`, and e_k the opposite-direction claim rate. An explicit direction on a practically unresolved oracle item remains in the denominator but is counted as neither correct nor wrong. The strict sensitivity also counts these cases as wrong; its effect depends on the selected claim set (Appendix B).

Because stronger configurations may abstain more often, we compare P0–P3 at matched c_k. At 2%, 5%, 10%, and 20% coverage, P0 and P1 have identical reported rates: $2.78\%, 1.67\%, 2.50\%, 3.13\%$. P2 gives $0.00\%, 0.83\%, 1.11\%, 2.15\%$, and P3 gives $0.00\%, 0.00\%, 0.00\%, 0.56\%$ (Figure 2a). Identical P0/P1 rates do not imply identical selected claim sets.

P2 has lower point estimates than P1 at all four operating points. P3 equals P2 at 2% and is lower at 5%, 10%, and 20%; paired intervals exclude zero for the latter two. At 10% coverage, P3–P2 is

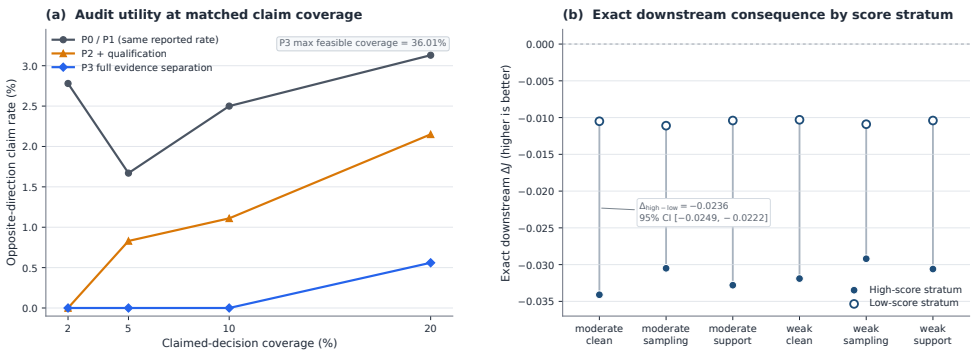


Figure 2: **Controlled known-answer validation.** **(a)** Opposite-direction claim rates for P0–P3 at matched `claimed_decision_coverage`. P0 and P1 coincide at the displayed operating points; lower rates emerge for P2 and P3, while P3 remains coverage-limited. **(b)** Exact downstream consequences for high- and low-score strata. Despite weak fine-grained rank association, high-score strata have worse exact outcomes in every primary regime. The supported interpretation is coarse consequence information, not precise event-level prediction.

-1.11 percentage points (95% CI $[-1.96, -0.42]$); at 20%, it is -1.60 ($[-2.48, -0.77]$). P3 is also lower than P0 at all four reported operating points.

Inference uses 10,000 paired resamples of `problem_seed`, retaining all six regimes of each problem together. Selected claim sets are fixed before resampling; thresholds are not reselected inside each replicate. The provenance record is reconstructed rather than prospectively preregistered. P3 qualifies 2,593 of 7,200 items, giving a maximum feasible *claimed-decision coverage* of 36.01%, not a reuse fraction.

A same-information control limits the attribution of this gain. At 2,593 claims, P3 makes 103 opposite-direction errors (3.972%), whereas selecting agreement items by minimum confidence makes five (0.193%). The latter admits 181 illegal-unit and 515 support-shift items, excluded by P3. Thus claim-count matching does not enforce the same qualification contract, and P3 is not the best same-information directional selector. The result supports bounded configuration utility, not the necessity or optimality of every P3 component (Appendix B).

4.2 Exact downstream relevance of the local measurement

We separately test whether the local process measurement contains downstream information. For target i, source j, and context x,

$$\Delta_{\mathrm{op}}(i, j, x) = \ell_i\big(h_{ij}^{+}(x)\big) - \ell_i\big(h_{ij}^{-}(x)\big)\,, \tag{8}$$

with target context and measurement contract fixed and only source inclusion changed. Positive Δ_{op} denotes larger target-side fitting loss under that contract, not an environment-safety violation.

Each intervention also has an exact downstream policy consequence $\Delta J_{\mathrm{exact}}$. Fine-grained association is weak: across the six primary skill/regime combinations, Spearman correlations between Δ_{op} and $-\Delta J_{\mathrm{exact}}$ are only 0.030–0.039. Thus Δ_{op} is not validated as a precise event-level downstream ranking.

The coarse comparison is stronger. In every primary regime, the high-score stratum has a more negative exact ΔJ than the low-score stratum (Figure 2b). For moderate-clean, the high–low difference is -0.0236 (95% CI $[-0.0249, -0.0222]$); for weak sampling noise it is -0.0183 ($[-0.0197, -0.0170]$). The remaining primary regimes preserve the same direction.

The two tests therefore support different conclusions: the first provides known-answer evidence for bounded directional decision utility of the implemented audit configurations; the second shows that the local measurement retains coarse downstream consequence information under the tested contracts without establishing fine-grained or prospective predictive validity.

Table 3: **Meta-World claim adjudication.** Each row gives a candidate claim, the evidence that changes its scope, and the strongest conclusion licensed under the evaluated contract.

Candidate claim	Decisive evidence	Licensed conclusion
Higher mean return implies target-wise benefit	15-seed target-level outcomes relative to paired Independent references	NOT SUPPORTED in the evaluated suite: higher mean performance can coexist with substantial target deficits.
RCER has uniformly better source selection under the canonical measurement contract	Matched within-panel reuse coverage, pair-balanced inference, and concentration sensitivity	CONDITIONAL: L5 has favorable pointwise pair-balanced evidence; corresponding L10 contrasts are unresolved and effect magnitude is concentrated. No multiplicity-adjusted joint superiority is claimed.
The complex RCER architecture is necessary for the observed audit ranking	Same-representation, same-label, same-event, matched-budget Exact-Simple-A	NOT SUPPORTED as a necessity claim: the observed audit ranking can be reproduced without the complex architecture under the matched information contract.
The canonical event ordering is invariant across measurement specifications	Closely related Bellman measurements and explicitly specified alternative operationalizations	CONDITIONAL: within-family measurements can agree strongly, but fine-grained ordering is not invariant across the evaluated measurement specifications.
Lower process degradation implies superior deployment	CQL process/rollout boundary test and source-context checks	UNRESOLVED: the available evidence does not establish a general process-to-deployment bridge.

5 META-WORLD AS A CLAIM-ADJUDICATION CASE STUDY

We next apply the audit to a heterogeneous offline-RL study in Meta-World (Yu et al., 2020b). IQL is the primary backbone (Kostrikov et al., 2022); CQL is used only as a secondary process/deployment boundary test (Kumar et al., 2020). The goal is not to recover a universal ranking of reuse methods, but to ask which initially plausible claims remain licensed once their evidence contracts are made explicit. Table 3 reports five consequential claim transitions; these are instantiations of the claim classes in Figure 1, not additional claim types. Complete contracts, provenance, and secondary analyses are in Appendices C–F.

Mean outcome and selector quality license narrower claims. The canonical IQL outcome rows report mean $\pm$ SE across seeds and use the paired Independent policy as the target-level reference. On L10, Shared Encoder attains mean return 390.7 ± 13.9, compared with 381.0 ± 14.7 for Independent and 384.2 ± 12.9 for RCER-Agg, yet both Shared Encoder and RCER-Agg have negative-transfer rate (NTR) 0.447. Thus the favorable mean does not establish target-wise benefit. A separate All-25 endpoint sweep likewise shows why the reference must remain explicit: zero reuse is the highest observed fixed endpoint, but we do not interpret that stress-test result causally because task composition and optimization conditions also change.

Selector-quality claims require a different contract. Within each set and measurement panel, target-wise top-k selection uses the same candidate source set and accepted source count. L5 has realized reuse coverage 0.75, whereas L10 has realized coverage $7/9 = 0.778$. Aggregate- and event-label panels instantiate different measurement contracts and are therefore never compared numerically across panels. Under the event contract, pooled S-URR is 2.264 for RCER-Event versus 3.392, 3.904, and 3.384 for Similarity, CDS, and Random. Because pooled events over-weight pairs with more retained events, we also use a pair-balanced finite-suite estimator. We first average complete, nonempty pair contrasts within each scorer seed, then resample the five shared seeds (42–46) as whole blocks, preserving cross-pair dependence. All $5^5 = 3,125$ ordered bootstrap resamples are enumerated; the intervals remain small-seed, pointwise summaries for the fixed suite, not guarantees for new tasks. On L5, RCER-Event S-URR differences are -1.040 versus Similarity (95% CI $[-1.933, -0.275]$), -1.656 versus CDS ($[-2.861, -0.518]$), and -1.350 versus Random ($[-2.156, -0.604]$); all corresponding L10 intervals include zero. CDS uses 15 complete pairs on L5 and 70 on L10; Similarity and Random use 20 and 90. Matching applies to the target-wise quota, not to the accepted count of every directed pair (Appendix D). A post-hoc concentration diagnostic further shows that the largest-magnitude 10% of pairs account for roughly 70–93% of absolute effect mass across the audited comparisons. The supported statement is therefore contract-specific and localized, not uniform pair-level superiority.

Matched controls narrow attribution and measurement claims. Exact-Simple-A receives the same target/source representation, counterfactual labels, training events, and realized selection budget as the corresponding RCER audit, but replaces the complex predictor with a shallow scalar MLP. On L10 event labels, its AUROC is 0.894 versus 0.884 for RCER and its matched S-URR is 1.992 versus

2.264. This argues against the stronger claim that RCER's architectural complexity is *necessary* for the observed audit-time ranking; it does not identify the separate causal contributions of architecture, supervision, representation, or optimization, nor establish an equivalent deployment controller.

The measurement-sensitivity results impose a second restriction. Canonical TD and held-out Bellman-error rankings agree strongly at the final checkpoint (Spearman 0.989 on L5 and 0.991 on L10), showing that closely related Bellman-family operationalizations can be stable. In contrast, normalized A′ measurements have near-zero correlation with the canonical A0 ordering on the L10 aggregate and event panels (0.001 and -0.068); additional alternatives can also change selector ordering. Thus the evidence supports local within-family agreement, not measurement-invariant fine-grained ranking. Importantly, the audit therefore preserves bounded positive findings—most clearly the L5 pair-balanced selector result and within-family Bellman agreement—rather than systematically rejecting comparative claims.

Process evidence does not close the deployment bridge. The CQL follow-up is a boundary test rather than a deployment ranking. At matched reuse coverage 0.75, the process and rollout point estimates do not induce the same ordering. Paired rollout-return inference resolves only part of this contrast: RCER exceeds Exact-Simple-A by $+0.616$ (95% CI $[0.024, 1.208]$), whereas the corresponding RCER contrasts against CDS, Similarity, and Random have intervals spanning zero. Thus, the batch does not establish a general correspondence between process and deployment rankings; the process-to-deployment bridge remains UNRESOLVED.

A source-context check addresses a narrower implementation ambiguity. The original evaluator uses target-copy representations in source slots; an actual-source prototype condition instead constructs them from the source task, while a self-only condition removes source context at the same checkpoint. On L10, prototype minus target-copy changes mean return by $+0.832$ (95% CI $[-8.503, 10.168]$). This interval does not establish equivalence; it simply does not resolve a large aggregate target-copy effect at the available precision. A separate 15-seed L5 re-evaluation gives prototype minus target-copy $+1.985$ (95% CI $[-1.034, 4.990]$) and prototype minus self-only $+32.028$ ($[14.580, 50.331]$), supporting only the latter return contrast under that deployment contract. The dedicated prospective bridge from local degradation to future rollout harm also remains incomplete and is therefore UNRESOLVED.

6 EXTERNAL CASES

We next ask whether the same claim discipline remains useful outside Meta-World. ManiSkill3 (Tao et al., 2025) provides a second simulation-based setting with process and rollout evidence, whereas D4RL (Fu et al., 2020) provides only a frozen action-prediction diagnostic. Neither case is used to claim universal controller transfer.

6.1 MANISKILL3: PROCESS AND DEPLOYMENT LICENSE DIFFERENT CLAIMS

ManiSkill3 uses five policy seeds and three locked rollout offsets per seed; uncertainty is clustered by policy seed, with offsets treated as within-seed repeats. The process audit shares candidate support, target observations, and a frozen target critic, but uses method-specific source-ON/OFF transformations (Appendix G). Thus the comparison concerns complete process pipelines, not an isolated selector under identical event-level losses.

At matched `reuse_coverage` 0.75, RCER–Similarity process contrasts are -0.0788 for URR (95% CI $[-0.1018, -0.0388]$) and -964.4 for S-URR ($[-1567.8, -444.7]$). Within its own panel, RCER also reduces S-URR relative to random matching by 120.954 (CI for RCER minus random: $[-153.484, -90.360]$); Similarity increases it by 150.608 ($[87.529, 213.687]$). These within-panel controls distinguish selection from accepted count without attributing the cross-pipeline gap solely to selector quality.

Deployment licenses a different statement. Relative to Similarity, RCER-Abstain has an unresolved mean-return difference of -0.113 (95% CI $[-0.331, 0.106]$), but lower NTR by -0.267 ($[-0.489, -0.044]$). Because abstention can dispatch to the Independent policy, this target-level difference belongs to the complete deployment contract rather than to the process score alone. Task

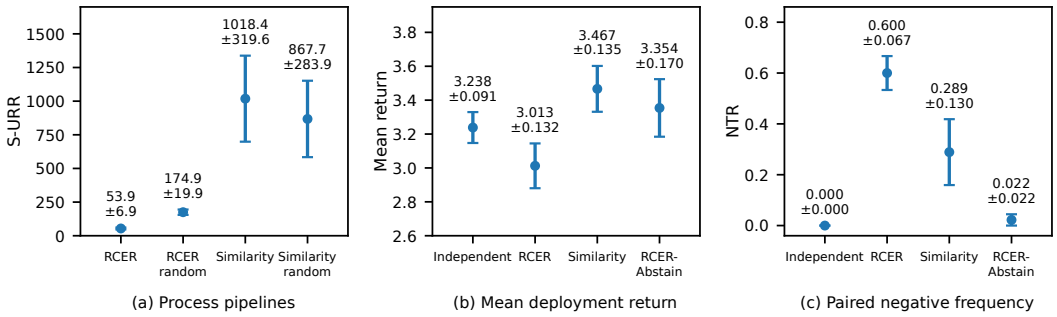


Figure 3: **ManiSkill3 evidence separation.** Points and error bars are means and standard errors over five policy seeds; deployment offsets are averaged within seed. **(a)** RCER has lower process-pipeline severity than Similarity at matched reuse count; source-ON/OFF transformations are method-specific. **(b)** Mean deployment return between RCER-Abstain and Similarity is unresolved. **(c)** RCER-Abstain has lower target-level degradation than Similarity.

success remains near zero for all evaluated methods, so these results do not establish task-solving superiority.

6.2 D4RL: A BOUNDED DIAGNOSTIC CASE

D4RL supplies only a frozen action-MSE diagnostic, not rollout validation. Here NTR is defined relative to the paired Independent action-MSE reference: a target counts as degraded when its MSE is higher than the corresponding Independent value. Across Hopper, Walker2d, and Reacher with 15 seeds, Similarity Routing has the lowest observed mean MSE among the evaluated methods (0.2396 ± 0.0056) and NTR 0.0222 ± 0.0222. Thus the licensed positive claim is narrow: *Similarity performs best among the compared methods on this frozen action-MSE diagnostic*. By contrast, Shared Encoder improves mean MSE relative to Independent by -0.1228 (95% CI $[-0.1658, -0.0796]$) while retaining NTR 0.3111 ($[0.1778, 0.4667]$), so an average diagnostic improvement does not establish uniform target-wise benefit.

Together, the external cases show that the claim-specific evidence discipline can recur outside Meta-World; they do not establish a universal operational loss scale, selector ranking, deployment improvement, or reuse controller.

7 LIMITATIONS AND CONCLUSION

Limitations. Our audit is an operational claim-licensing procedure, not a general theory of comparative evidence. The controlled benchmark establishes only bounded decision utility: stronger audit configurations reduce opposite-direction claims at selected matched `claimed_decision_coverage` points under the evaluated finite-MDP generators and fixed rules. Audit statuses are not universally validated, bootstrap intervals condition on fixed selected claim sets, and a same-information agreement rule has lower directional error than P3 under a different eligibility contract.

Other conclusions are similarly scoped. The critic-local score retains coarse but not fine-grained downstream information. The Meta-World process-to-rollout bridge remains UNRESOLVED; ManiSkill3 is contract-dependent with near-zero task success; and D4RL provides only an action-MSE diagnostic. None establishes universal measurement, selector, or deployment transfer.

Conclusion. We study when comparative claims in heterogeneous offline experience reuse are warranted by their evidence. Known-answer experiments show bounded audit utility, while the case studies show that mean outcome, target-wise benefit, selector quality, mechanism attribution, and deployment performance are not interchangeable evidence. The reusable contribution is a discipline: *specify the claim, evaluate the corresponding evidence contract, and license only the conclusion supported within that scope.*

AI USE STATEMENT

Large language models were used for coding assistance, manuscript editing, statistical consistency checking, feedback on research methodology and statistical analysis, and interpretation of the supplied experimental records. No generated result is reported without an underlying experiment or artifact. Model assistance was not used as a substitute for scientific validation; the authors are responsible for the final analyses, citations, and claims.

REFERENCES

Joshua Achiam, David Held, Aviv Tamar, and Pieter Abbeel. Constrained Policy Optimization. In *Proceedings of the 34th International Conference on Machine Learning*, pp. 22–31, 2017.

Rishabh Agarwal, Dale Schuurmans, and Mohammad Norouzi. An Optimistic Perspective on Offline Reinforcement Learning. In *Proceedings of the 37th International Conference on Machine Learning*, pp. 104–114, 2020.

Rishabh Agarwal, Max Schwarzer, Pablo Samuel Castro, Aaron Courville, and Marc G. Bellemare. Deep Reinforcement Learning at the Edge of the Statistical Precipice. In *Advances in Neural Information Processing Systems*, 34, 2021.

Chenjia Bai, Lingxiao Wang, Jianye Hao, Zhuoran Yang, Bin Zhao, Zhen Wang, and Xuelong Li. Pessimistic Value Iteration for Multi-Task Data Sharing in Offline Reinforcement Learning. *Artificial Intelligence*, 326:104048, 2024.

Rich Caruana. Multitask Learning. *Machine Learning*, 28(1):41–75, 1997.

Gavin C. Cawley and Nicola L. C. Talbot. On Over-fitting in Model Selection and Subsequent Selection Bias in Performance Evaluation. *Journal of Machine Learning Research*, 11:2079–2107, 2010.

Zhao Chen, Vijay Badrinarayanan, Chen-Yu Lee, and Andrew Rabinovich. GradNorm: Gradient Normalization for Adaptive Loss Balancing in Deep Multitask Networks. In *Proceedings of the 35th International Conference on Machine Learning*, pp. 794–803, 2018.

Justin Fu, Aviral Kumar, Ofir Nachum, George Tucker, and Sergey Levine. D4RL: Datasets for Deep Data-Driven Reinforcement Learning. *arXiv preprint arXiv:2004.07219*, 2020.

Scott Fujimoto, David Meger, and Doina Precup. Off-Policy Deep Reinforcement Learning without Exploration. In *Proceedings of the 36th International Conference on Machine Learning*, pp. 2052–2062, 2019.

Scott Fujimoto, and Shixiang Shane Gu. A Minimalist Approach to Offline Reinforcement Learning. In *Advances in Neural Information Processing Systems*, 2021.

Javier Garcia, and Fernando Fernandez. A Comprehensive Survey on Safe Reinforcement Learning. *Journal of Machine Learning Research*, 16:1437–1480, 2015.

Yonatan Geifman, and Ran El-Yaniv. Selective Classification for Deep Neural Networks. In *Advances in Neural Information Processing Systems*, pp. 4878–4887, 2017.

Yonatan Geifman, and Ran El-Yaniv. SelectiveNet: A Deep Neural Network with an Integrated Reject Option. In *Proceedings of the 36th International Conference on Machine Learning*, pp. 2151–2159, 2019.

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.

Jinmin He, Kai Li, Yifan Zang, Haobo Fu, Qiang Fu, Junliang Xing, and Jian Cheng. Goal-Oriented Skill Abstraction for Offline Multi-Task Reinforcement Learning. In *Proceedings of the 42nd International Conference on Machine Learning*, pp. 22566–22579, 2025.

Peter Henderson, Riashat Islam, Philip Bachman, Joelle Pineau, Doina Precup, and David Meger. Deep Reinforcement Learning That Matters. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1), 2018.

Alex Kendall, Yarin Gal, and Roberto Cipolla. Multi-Task Learning Using Uncertainty to Weigh Losses for Scene Geometry and Semantics. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7482–7491, 2018.

Rahul Kidambi, Aravind Rajeswaran, Praneeth Netrapalli, and Thorsten Joachims. MOReL: Model-Based Offline Reinforcement Learning. In *Advances in Neural Information Processing Systems*, 2020.

Ilya Kostrikov, Ashvin Nair, and Sergey Levine. Offline Reinforcement Learning with Implicit Q-Learning. In *International Conference on Learning Representations*, 2022.

Aviral Kumar, Justin Fu, Matthew Soh, George Tucker, and Sergey Levine. Stabilizing Off-Policy Q-Learning via Bootstrapping Error Reduction. In *Advances in Neural Information Processing Systems*, 2019.

Aviral Kumar, Aurick Zhou, George Tucker, and Sergey Levine. Conservative Q-Learning for Offline Reinforcement Learning. In *Advances in Neural Information Processing Systems*, 2020.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles. In *Advances in Neural Information Processing Systems*, pp. 6402–6413, 2017.

Romain Laroche, Paul Trichelair, and Remi Tachet des Combes. Safe Policy Improvement with Baseline Bootstrapping. In *Proceedings of the 36th International Conference on Machine Learning*, pp. 3652–3661, 2019.

Sergey Levine, Aviral Kumar, George Tucker, and Justin Fu. Offline Reinforcement Learning: Tutorial, Review, and Perspectives on Open Problems. *arXiv preprint arXiv:2005.01643*, 2020.

Xi Lin, Hui-Ling Zhen, Zhenhua Li, Qing-Fu Zhang, and Sam Kwong. Pareto Multi-Task Learning. In *Advances in Neural Information Processing Systems*, 2019.

Bo Liu, Xingchao Liu, Xiaojie Jin, Peter Stone, and Qiang Liu. Conflict-Averse Gradient Descent for Multi-task Learning. In *Advances in Neural Information Processing Systems*, 2021.

Shikun Liu, Stephen James, Andrew J. Davison, and Edward Johns. Auto-Lambda: Disentangling Dynamic Task Relationships. *Transactions on Machine Learning Research*, 2022.

Wei Liu and Ting Long. Target-Aligned Bellman Backup for Cross-domain Offline Reinforcement Learning. *arXiv preprint arXiv:2605.22376*, 2026.

Mingsheng Long, Zhangjie Cao, Jianmin Wang, and Philip S. Yu. Learning Multiple Tasks with Multilinear Relationship Networks. In *Advances in Neural Information Processing Systems*, 2017.

Ishan Misra, Abhinav Shrivastava, Abhinav Gupta, and Martial Hebert. Cross-Stitch Networks for Multi-Task Learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3994–4003, 2016.

Aviv Navon, Aviv Shamsian, Idan Achituve, Haggai Maron, Kenji Kawaguchi, Gal Chechik, and Ethan Fetaya. Multi-Task Learning as a Bargaining Game. In *Proceedings of the 39th International Conference on Machine Learning*, pp. 16428–16446, 2022.

Sinno Jialin Pan, and Qiang Yang. A Survey on Transfer Learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359, 2010.

Zhongjian Qiao, Rui Yang, Jiafei Lyu, Chenjia Bai, Xiu Li, Zhuoran Yang, Siyang Gao, and Shuang Qiu. Cross-Domain Offline Policy Adaptation with Dynamics- and Value-Aligned Data Filtering. *arXiv preprint arXiv:2512.02435*, 2025.

Zhongjian Qiao, Jiafei Lyu, Chenjia Bai, Peisong Wang, Siyang Gao, and Shuang Qiu. Unifying Value Alignment and Assignment in Cross-Domain Offline Reinforcement Learning with Heterogeneous Datasets. *arXiv preprint arXiv:2605.24862*, 2026.

Clemens Rosenbaum, Tim Klinger, and Matthew Riemer. Routing Networks: Adaptive Selection of Non-Linear Functions for Multi-Task Learning. In *International Conference on Learning Representations*, 2018.

Ozan Sener, and Vladlen Koltun. Multi-Task Learning as Multi-Objective Optimization. In *Advances in Neural Information Processing Systems*, 2018.

Trevor Standley, Amir Zamir, Dawn Chen, Leonidas Guibas, Jitendra Malik, and Silvio Savarese. Which Tasks Should Be Learned Together in Multi-task Learning?. In *Proceedings of the 37th International Conference on Machine Learning*, pp. 9120–9132, 2020.

Ximeng Sun, Rameswar Panda, Rogerio Feris, and Kate Saenko. AdaShare: Learning What to Share for Efficient Deep Multi-Task Learning. In *Advances in Neural Information Processing Systems*, 2020.

Stone Tao, Fanbo Xiang, Arth Shukla, Yuzhe Qin, Xander Hinrichsen, Xiaodi Yuan, Chen Bao, Xinsong Lin, Yulin Liu, Tse-kai Chan, Yuan Gao, Xuanlin Li, Tongzhou Mu, Nan Xiao, Arnav Gurha, Viswesh Nagaswamy Rajesh, Yong Woo Choi, Yen-Ru Chen, Zhiao Huang, Roberto Calandra, Rui Chen, Shan Luo, and Hao Su. ManiSkill3: GPU Parallelized Robotics Simulation and Rendering for Generalizable Embodied AI. In *Robotics: Science and Systems*, 2025.

Yee Whye Teh, Victor Bapst, Wojciech M. Czarnecki, John Quan, James Kirkpatrick, Raia Hadsell, Nicolas Heess, and Razvan Pascanu. Distral: Robust Multitask Reinforcement Learning. In *Advances in Neural Information Processing Systems*, 2017.

Philip Thomas, Georgios Theocharous, and Mohammad Ghavamzadeh. High Confidence Policy Improvement. In *Proceedings of the 32nd International Conference on Machine Learning*, pp. 2380–2388, 2015.

Zirui Wang, Zihang Dai, Barnabas Poczos, and Jaime Carbonell. Characterizing and Avoiding Negative Transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11293–11302, 2019.

Xiaoyu Wen, Chenjia Bai, Kang Xu, Xudong Yu, Yang Zhang, Xuelong Li, and Zhen Wang. Contrastive Representation for Data Filtering in Cross-Domain Offline Reinforcement Learning. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 52720–52743, 2024.

Yifan Wu, George Tucker, and Ofir Nachum. Behavior Regularized Offline Reinforcement Learning. *arXiv preprint arXiv:1911.11361*, 2019.

Ruihan Yang, Huazhe Xu, Yi Wu, and Xiaolong Wang. Multi-Task Reinforcement Learning with Soft Modularization. In *Advances in Neural Information Processing Systems*, 2020.

Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Gradient Surgery for Multi-Task Learning. In *Advances in Neural Information Processing Systems*, 2020a.

Tianhe Yu, Deirdre Quillen, Zhanpeng He, Ryan Julian, Anisha Narayan, Hayden Shively, Adithya Bellathur, Karol Hausman, Chelsea Finn, and Sergey Levine. Meta-World: A Benchmark and Evaluation for Multi-Task and Meta Reinforcement Learning. In *Conference on Robot Learning*, 2020b.

Tianhe Yu, Garrett Thomas, Lantao Yu, Stefano Ermon, James Zou, Sergey Levine, Chelsea Finn, and Tengyu Ma. MOPO: Model-based Offline Policy Optimization. In *Advances in Neural Information Processing Systems*, 2020c.

Tianhe Yu, Aviral Kumar, Yevgen Chebotar, Karol Hausman, Sergey Levine, and Chelsea Finn. Conservative Data Sharing for Multi-Task Offline Reinforcement Learning. In *Advances in Neural Information Processing Systems*, 2021.

Amir R. Zamir, Alexander Sax, William Shen, Leonidas J. Guibas, Jitendra Malik, and Silvio Savarese. Taskonomy: Disentangling Task Transfer Learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3712–3722, 2018.

Appendix Overview

The appendix follows the paper's evidence chain rather than the chronological order in which experiments were conducted. Appendix A fixes definitions and inferential conventions; Appendix B expands the known-answer validation; Appendices C–F provide the complete Meta-World evidence underlying the claim transitions; Appendices G and H contain the external cases; and Appendix I maps each principal claim to the evidence that licenses it.

The appendix is intended to support scientific verification of the paper's claims. Reproduction assets, manifests, hashes, scripts, and unit-level derived data are provided separately in the supplementary artifact.

A Definitions and Protocol Details

This appendix collects the definitions and inferential conventions used throughout the paper. The goal is to make the claim-audit procedure operationally unambiguous without treating the decomposition as a universal theory of scientific evidence.

A.1 Claim specification and audit states

A candidate comparative claim is represented as

$$\mathcal{C} = (A, B, \theta, \mathcal{P}, R, m, b, u, d, s), \tag{9}$$

where A and B are the compared methods, θ is the estimand, $\mathcal{P}$ the evaluation population, R the reference comparator, m the measurement contract, b the comparison budget, u the scientific independence unit, d the deployment contract, and s the intended claim scope.

The five claim classes used in the paper are: *mean outcome*, *target-wise benefit*, *selector quality*, *mechanism attribution*, and *deployment performance*. They are alternative claim types rather than sequential stages.

For a claim-specific decision rule $\delta_{\mathcal{C}}(\mathcal{D}) \in \{\textsc{Support}, \textsc{Inconclusive}, \textsc{Contradict}\}$, the final audit status is assigned as follows:

- Supported: the submitted bounded claim is supported under its evaluated contract;
- Conditional: the original claim is too broad, but an explicitly evaluated narrower claim is supported;
- Unresolved: evidence is insufficient, a required bridge is missing, or uncertainty prevents directional adjudication;
- Not Supported: evidence contradicts the specified strong version of the claim.

A confidence interval crossing zero is not treated as equivalence, and an incomplete experiment is not treated as negative evidence.

A.2 Coverage semantics

We use three distinct notions of coverage and never exchange them:

`reuse_coverage` the fraction or quota of source experience accepted by a reuse selector;

`claimed_decision_coverage` the fraction of audit decision items on which an explicit directional claim is issued;

`endpoint_reuse_fraction` the fixed amount of positive reuse executed by a deployed endpoint.

Matched reuse coverage is used when the scientific question is *selector quality at fixed reuse quantity*. It is not imposed on every deployment comparison, because reuse amount may itself be part of a complete deployed method.

A.3 Outcome metrics

For method m, target i, and policy seed s, define

$$D_{is}^{m} = \left[R_{is}^{\mathrm{ind}} - R_{is}^{m}\right]_{+}. \tag{10}$$

The canonical target-level metrics are

$$\mathrm{NTR}_m = \frac{1}{|\mathcal{I}||\mathcal{S}|} \sum_{i,s} \mathbb{1}[D_{is}^{m} > 0], \tag{11}$$

$$\mathrm{NTM}_m = \frac{1}{|\mathcal{I}||\mathcal{S}|} \sum_{i,s} D_{is}^{m}, \tag{12}$$

$$\mathrm{MaxDeg}_m = \frac{1}{|\mathcal{S}|} \sum_{s} \max_{i} D_{is}^{m}. \tag{13}$$

Thus NTR measures the frequency of target-level degradation relative to the paired Independent reference, NTM measures its mean magnitude, and MaxDeg is the seed-average of the worst target deficit. NTR is an empirical target$\times$seed negative-difference frequency, not the fraction of tasks with statistically established or practically important harm.

A.4 Process metrics

For accepted event set $\mathcal{A}$ and operational degradation $\Delta_{\mathrm{op}}(e)$, define

$$\mathrm{URR} = \frac{1}{|\mathcal{A}|} \sum_{e \in \mathcal{A}} \mathbb{1}[\Delta_{\mathrm{op}}(e) > 0], \tag{14}$$

$$\text{S-URR} = \frac{1}{|\mathcal{A}|} \sum_{e \in \mathcal{A}} [\Delta_{\mathrm{op}}(e)]_{+}. \tag{15}$$

We additionally report accepted p_{95} of Δ_{op} when the event-level distribution is available. URR, S-URR, and accepted p_{95} answer different questions: frequency, mean severity, and tail severity, respectively.

A.5 Uncertainty units

The uncertainty unit follows the scientific object being generalized over. Controlled protocol utility uses `problem_seed`; Meta-World outcome tables use policy seeds; pair-balanced selector analyses macro-average complete pairs within each scorer seed and resample whole shared scorer-seed blocks; ManiSkill3 clusters deployment uncertainty by policy seed, with rollout offsets treated as within-seed repeats. No event count is interpreted as an independent replication merely because it is numerically large.

B Controlled Benchmark

This appendix expands the two controlled tests in Section 4: protocol decision utility and exact downstream relevance of the local measurement.

B.1 Finite-MDP construction and exactness checks

The benchmark contains 1,200 held-out finite-MDP problem instances with development and test problem seeds disjoint. Exact Bellman solutions and forced-reuse arithmetic are used only to construct evaluation oracles; they are not inputs to the audit configurations. The maximum numerical Bellman residual is 4.32×10^{-15}, and the maximum forced-intervention arithmetic error is 5.33×10^{-15}.

The broader benchmark contains clean, sampling-noise, support-shift, shared-bias, legal-affine, and illegal-unit stress regimes. The six primary skill/regime combinations used for the main exact-downstream result are reported separately from secondary stress tests.

For protocol utility, development seeds are 120000–120259 and test seeds 220000–221199, alternating moderate and weak skill. Each problem samples 20–50 states, 3–5 actions, and 4–10 source labels. Transition rows follow $\mathrm{Dirichlet}(0.6\mathbf{1})$, rewards are standard normal, and the reference policy is the softmax of standard-normal logits. With $\gamma = 0.95$, solve $(I - \gamma P_\pi)V = r_\pi$ and set $y(s,a) = R(s,a) + \gamma P(\cdot \mid s,a)^\top V$. Each problem has five target labels with 120 events each; each selector accepts 300 of the pooled 600 events. The six observation regimes produce $1,200 \times 6 = 7,200$ decision items, not 7,200 independent problems.

For independent standard-normal features x_r, x_u, x_a, the generator uses $q^- = y + 0.35\eta + 0.25x_a + b_t$ and $q^+ = q^- + 0.32x_r - 0.28x_u + b_s + 0.25\eta'$, with $b_t \sim N(0, 0.25^2)$ and $b_s \sim N(0, 0.35^2)$. The A/B ranking scores are $-0.55x_r + 0.35x_u - 0.15x_a + \xi_A$ and $-0.35x_r + 0.52x_u + 0.20x_a \operatorname{sign}(\zeta) + \xi_B$. Their noise standard deviations are $(0.9, 1.05)$ for moderate skill and $(1.6, 1.75)$ for weak skill. These scores, not oracle degradation, select source events. Full observation-generation code is supplied in `manifests/source_scripts/protocol_utility_validation_20260909.py`. The separate exact-downstream experiment is not an additional copy in the 7,200-item protocol-decision denominator.

B.2 Exact process oracle

For event e, the exact forced-reuse degradation is

$$\Delta_e^\star = (q_e^+ - y_e)^2 - (q_e^- - y_e)^2. \tag{16}$$

At matched reuse budget 0.50, each selector accepts its own top-k events. Its exact accepted severity is

$$S_{\mathrm{URR}}^\star = \frac{1}{|\mathcal{A}|} \sum_{e \in \mathcal{A}} \max(\Delta_e^\star, 0). \tag{17}$$

For methods A and B,

$$D_{A-B}^\star = S_{\mathrm{URR},A}^\star - S_{\mathrm{URR},B}^\star. \tag{18}$$

Because lower severity is better, the oracle direction is A-Better if $D_{A-B}^\star < -\epsilon$, B-Better if $D_{A-B}^\star > \epsilon$, and Practically-Unresolved otherwise. The frozen tolerance is

$$\epsilon = 0.0011253780328593782, \tag{19}$$

derived on development problems as 0.05 times the median $|D^\star|$, subject to a floor of 10^{-6}.

Among the 1,200 held-out problem instances, the oracle labels are 615 A-Better, 553 B-Better, and 32 Practically-Unresolved.

B.3 Directional-error semantics

For audit configuration P_k, let $a_k(n) \in \{-1, 0, +1\}$ be the issued direction on decision item n, where 0 denotes abstention, and let $y_n \in \{-1, 0, +1\}$ denote the oracle label, where 0 denotes Practically-Unresolved. Claimed-decision coverage is

$$c_k = \frac{1}{N} \sum_{n=1}^{N} \mathbb{1}[a_k(n) \neq 0]. \tag{20}$$

The opposite-direction claim rate is

$$e_k = \frac{\sum_n \mathbb{1}[a_k(n) \neq 0]\mathbb{1}[y_n \neq 0]\mathbb{1}[a_k(n) \neq y_n]}{\sum_n \mathbb{1}[a_k(n) \neq 0]}. \tag{21}$$

An explicit direction issued on a practically unresolved oracle item remains in the issued-claim denominator but is counted as neither correct nor wrong. This follows the frozen implementation.

The native fixed-$z = 1.96$ run reports 192 unresolved-oracle claims, all from P0; counting them as errors changes its native rate from 0.1497 to 0.1764. This is not the matched-coverage sweep. In

Table 4: Opposite-direction claim rate (%) at matched `claimed_decision_coverage`.

Coverage	P0	P1	P2	P3
2%	2.78	2.78	0.00	0.00
5%	1.67	1.67	0.83	0.00
10%	2.50	2.50	1.11	0.00
20%	3.13	3.13	2.15	0.56

that sweep, unresolved counts for P0/P1/P2/P3 are $(0, 0, 0, 0)$ at 2%, $(1, 1, 0, 0)$ at 5%, $(1, 1, 0, 1)$ at 10%, and $(4, 4, 5, 2)$ at 20%; the strict rate adds each count to the numerator while retaining the same issued-claim denominator.

B.4 Matched claimed-decision coverage

The issued-claim counts are 144, 360, 720, and 1,440 at the four coverages. Corresponding wrong counts (P0/P1/P2/P3) are $(4, 4, 0, 0)$, $(6, 6, 3, 0)$, $(18, 18, 8, 0)$, and $(45, 45, 31, 8)$. P0/P1 have identical rates, not necessarily identical selected sets. P3 equals P2 at 2% and has lower point estimates at the other three points; the paired interval excludes zero at 10% and 20%, but reaches zero at 5%. Updated P3–P2 differences are -1.11 percentage points ($[-1.96, -0.42]$) and -1.60 points ($[-2.48, -0.77]$) at 10% and 20%.

The current artifact materializes the pre-selection mask in `controlled/data/protocol_candidate_decisions.csv`. P3 qualifies 2,593/7,200 items: its maximum feasible *claimed-decision coverage* is 36.01%. This is neither source-event reuse coverage nor the native fixed-z issued-claim fraction. The updated analysis uses 10,000 paired problem-seed bootstrap draws (RNG seed 20260926), keeping all six regimes of a problem together and fixing each selected claim set before resampling. No threshold is reselected. These are pointwise, fixed-selection intervals, not simultaneous or threshold-selection guarantees.

B.5 Audit-configuration implementation

For specification j, d_j is the accepted-severity difference between A and B. The ranking standard error s_j is the standard deviation of target–source pair mean differences divided by the square root of the number of pairs accepted by both methods; this is not the final inferential unit. The supplied frozen qualification rule uses a deterministic 50/50 split:

$$Q_j = \mathbb{1}\left[\frac{\mathrm{MSE}_{\mathrm{heldout}}(q_j^-, y_j)}{\max\{\mathrm{MSE}_{\mathrm{heldout}}(\bar{y}_{j,\mathrm{train}}, y_j), 10^{-12}\}} \le 0.43629700891852\right] \mathbb{1}[\mathrm{support_rate}_j \ge 0.92]. \tag{22}$$

The loss-ratio threshold is the development 70th percentile; the support threshold is the clipped development 20th percentile. The two specifications use separate noisy predictions and targets. The sweep's materialized qualification flags are the authoritative inputs for its reconstruction.

Table 2 gives the sweep rules. At fraction c, retain $\lceil 7200c \rceil$ eligible rows by decreasing confidence, with ties resolved by problem seed and regime. The selected counts reproduce the supplied sweep at all four displayed points. In the separate native fixed-z implementation, P1 requires $|d_1| - \epsilon > 1.96 s_1$, P2 additionally requires Q_1, and P3 requires both qualified specifications to pass that test and agree. Native fixed-threshold records and sweep observations are distinct; a fixed-z issued count must not replace the sweep's eligibility ceiling.

Same-information control. At 2,593 claims, agreement-then-minimum-confidence makes 5 errors (0.193%) versus P3's 103 (3.972%); unresolved-oracle counts are 3 and 33. Both use the two recorded differences and uncertainties, but only P3 requires both qualification flags. The simple rule accepts 181 illegal-unit and 515 support-shift items; P3 accepts none. Thus P3 is not the best same-information *directional* selector, while matched count does not equate admissibility contracts. All 14 rules are retained in `scientific_closure/same_info_audit/same_information_matched_count_summary.csv`. Oracle labels are not used to select claims; these are retrospective controls, not a prospectively calibrated optimal-selector study.

B.6 EXACT DOWNSTREAM RELEVANCE OF THE LOCAL MEASUREMENT

Construction and policy consequence. This is a separate experiment from the protocol-utility run in Appendix B. For each skill setting (moderate and weak), it uses 1,200 problem seeds 220000–221199 and samples 120 of 600 generated events per problem without replacement (sampling seed 20260909). The same sampled events are evaluated in five observation regimes; Figure 2b shows clean, sampling-noise, and support-shift conditions for both skills. The downstream generator uses 20–50 states, 3–5 actions, transition rows from $\mathrm{Dirichlet}(0.6\mathbf{1})$, standard-normal rewards, and a reference policy π_b given by the softmax of standard-normal logits. Exact evaluation gives Q^{π_b}. The no-source policy estimate is $\widehat{Q}^0 = Q^{\pi_b} + \sigma_Q Z$, with independent standard-normal Z and $\sigma_Q = 0.30$ (moderate) or 0.45 (weak).

For event $e = (s_e, a_e)$, the source effect z_e is $0.32x_{r,e} - 0.28x_{u,e} + b_{s(e)} + 0.25\eta'_e$, plus an independent $0.20\eta''_e$ for weak skill. The features and η variables are standard normal, and source biases have standard deviation 0.35. Force-reject keeps $\widehat{Q}^-_e = \widehat{Q}^0$; force-accept changes only entry (s_e, a_e):

$$\widehat{Q}^+_e = \widehat{Q}^0 + z_e E_{s_e a_e}, \qquad \pi^{\pm}_e(a \mid s) = \mathrm{softmax}_a\big(\widehat{Q}^{\pm}_e(s,a)/0.35\big). \tag{23}$$

Here $E_{s_e a_e}$ is one at that entry and zero elsewhere. Each event starts from the same no-source estimate; interventions are not accumulated. For $\gamma = 0.95$ and a uniform initial-state distribution, compute

$$V^\pi = (I - \gamma P^\pi)^{-1} r^\pi, \qquad J(\pi) = |\mathcal{S}|^{-1}\mathbf{1}^\top V^\pi, \qquad \Delta J_{\mathrm{exact},e} = J(\pi^+_e) - J(\pi^-_e). \tag{24}$$

Thus the downstream target is an exact discounted-return difference, not an auditor prediction or a Monte Carlo rollout estimate.

Operational measurement and grouping. The event-local no-source value is $q^-_e = Q^{\pi_b}(s_e, a_e) + 0.35\eta_e + 0.25x_{a,e} + b_{t(e)}$, where $b_t \sim \mathcal{N}(0, 0.25^2)$, and $q^+_e = q^-_e + z_e$. The local squared-loss difference uses target $Q^{\pi_b}(s_e, a_e)$. The clean observed score adds zero-mean Gaussian noise with standard deviation 0.25 (moderate) or 0.45 (weak) times the within-problem standard deviation of that local difference. Sampling noise adds 0.65 times the clean-score standard deviation; support shift adds 0.35 times that deviation and clips at the 88th percentile of the absolute perturbed score. Within each problem and regime, the bottom and top groups are defined by the lower and upper score tertiles, respectively. Correlations in Section 4.2 are pooled event-level Spearman correlations. The reported high–low effect instead averages within-problem contrasts; its 95% percentile interval uses 5,000 problem-level bootstrap resamples within each skill/regime cell (seed 20260909), retaining each problem's paired high/low groups. Events are not independent bootstrap units. The source is `manifests/source_scripts/finite_mdp_exact_downstream_consequence_20260909.py`, functions `make_problem`, `exact_downstream_for_events`, `classify_groups`, and `summarize`. Inputs and recorded summaries are in `controlled/summaries/exact_downstream/`.

Retained results. For target i, source j, and context x, the local measurement is

$$\Delta_{\mathrm{op}}(i, j, x) = \ell_i(h^+_{ij}(x)) - \ell_i(h^-_{ij}(x)). \tag{25}$$

Only source inclusion changes; the target context and measurement contract are held fixed.

Every controlled intervention also has an exact downstream policy consequence $\Delta J_{\mathrm{exact}}$. Across the six primary skill/regime combinations, fine-grained rank association is weak, with Spearman correlations approximately 0.030–0.039. The score is therefore not validated as a precise event-level downstream ranking.

The coarse stratum comparison is more favorable. For moderate-clean, the high-score stratum has mean exact consequence -0.0341, compared with -0.0105 for the low-score stratum, giving

$$\Delta J_{\mathrm{high-low}} = -0.0236, \qquad 95\%\ \mathrm{CI} = [-0.0249, -0.0222]. \tag{26}$$

For weak sampling noise, the corresponding contrast is -0.0183 with CI $[-0.0197, -0.0170]$. All primary regimes preserve the same direction.

The supported interpretation is therefore coarse consequence information under the tested finite-MDP generators and intervention contracts, not precise future-harm prediction.

Table 5: Canonical Meta-World IQL outcomes. Values are mean ± SE across policy seeds.

Set	Method	Avg ↑	NTR ↓	NTM ↓	MaxDeg ↓
L5	Independent	264.7 ± 7.4	0.000 ± 0.000	0.0 ± 0.0	0.0 ± 0.0
L5	Full Sharing	242.8 ± 9.1	0.613 ± 0.069	40.2 ± 6.8	144.1 ± 25.2
L5	Shared Encoder	287.9 ± 7.9	0.400 ± 0.052	22.3 ± 5.3	103.3 ± 26.5
L5	Similarity	242.0 ± 7.0	0.507 ± 0.067	41.8 ± 7.1	147.5 ± 25.8
L5	CDS	245.7 ± 4.1	0.587 ± 0.053	38.3 ± 7.0	158.8 ± 33.8
L5	RCER-Agg	344.3 ± 15.0	0.280 ± 0.055	8.4 ± 2.3	39.3 ± 10.8
L10	Independent	381.0 ± 14.7	0.000 ± 0.000	0.0 ± 0.0	0.0 ± 0.0
L10	Full Sharing	229.2 ± 6.4	0.640 ± 0.043	160.6 ± 15.3	1253.8 ± 137.2
L10	Shared Encoder	390.7 ± 13.9	0.447 ± 0.036	50.1 ± 14.5	419.8 ± 146.5
L10	Similarity	362.4 ± 18.4	0.480 ± 0.039	72.1 ± 15.3	545.2 ± 110.6
L10	CDS	267.4 ± 10.0	0.520 ± 0.035	131.1 ± 16.6	1090.1 ± 152.2
L10	RCER-Agg	384.2 ± 12.9	0.447 ± 0.031	50.1 ± 11.2	409.8 ± 99.7

B.7 Analytic shared-bias sensitivity

Under additive shared bias b, the local degradation obeys

$$\Delta_b = \Delta + 2b(q_+ - q_-). \tag{27}$$

The identity is numerically verified to maximum error 5.33×10^{-15}. This analysis illustrates why the numerical operational score depends on the measurement contract even when the underlying intervention pair is fixed.

C Meta-World Outcome Evidence

This appendix contains the full Meta-World outcome evidence underlying the claim transitions in Section 5.

C.1 Canonical IQL outcome protocol

Canonical IQL outcome tables report mean ± SE over policy seeds. Independent is the paired target-level reference. Primary baselines and RCER-Agg use 15 seeds; canonical RCER-Event rollout results use five where reported. These marginal SEs are method-level uncertainty summaries and are not paired difference intervals.

C.2 All-25 endpoint stress test

The All-25 experiment compares fixed `endpoint_reuse_fraction` values. Among the evaluated endpoints, zero reuse has the highest observed mean return: 680.452 ± 29.667. Positive fixed fractions produce substantially lower observed means: 329.744 at 0.25, 330.838 at 0.50, 342.152 at 0.75, and 357.675 at 1.00.

These endpoints are not used to infer that heterogeneity causally makes reuse harmful because task composition, capacity, and optimization conditions differ from L5/L10. They establish only that the no-reuse reference cannot be omitted from the evaluated reuse frontier.

D Matched Process Inference

D.1 Within-panel matched coverage

Matched process comparisons use the same candidate support and accepted source count within each measurement panel. Both L5 and L10 are evaluated under separate aggregate- and event-label measurement contracts. The realized reuse coverage is 0.75 on L5 and $7/9 = 0.778$ on L10. Comparisons are made only within a measurement contract; values from aggregate- and event-label panels are not compared numerically.

Table 6: Canonical matched-coverage process summaries stratified by measurement contract. "Aggregate" denotes the RCER-Agg-label contract and "Event" the RCER-Event-label contract. Within each Set×Contract panel, methods share the candidate support and accepted source count. Values are not compared numerically across measurement contracts. Lower is better.

Set	Contract	Method	Coverage	URR ↓	S-URR ↓	Accepted p95 ↓
L5	Aggregate	RCER-Agg	.750	.288	.452	.888
L5	Aggregate	Similarity	.750	.318	.499	1.100
L5	Aggregate	CDS	.750	.309	.487	1.145
L5	Aggregate	Random	.750	.311	.720	1.264
L5	Event	RCER-Event	.750	.313	.436	1.051
L5	Event	Similarity	.750	.359	1.448	2.995
L5	Event	CDS	.750	.410	2.095	7.588
L5	Event	Random	.750	.403	1.835	5.658
L10	Aggregate	RCER-Agg	.778	.320	1.210	1.581
L10	Aggregate	Similarity	.778	.360	1.734	2.809
L10	Aggregate	CDS	.778	.372	1.920	3.002
L10	Aggregate	Random	.778	.372	1.828	3.151
L10	Event	RCER-Event	.778	.423	2.264	5.792
L10	Event	Similarity	.778	.505	3.392	9.869
L10	Event	CDS	.778	.519	3.904	12.585
L10	Event	Random	.778	.515	3.384	10.843

Table 7: Pair-balanced RCER-Event S-URR contrasts under shared scorer-seed resampling. Negative values favor RCER-Event. Intervals are pointwise; $|P_{AB}|$ is the fixed complete, nonempty pair set.

Set	Baseline	$\lvert P_{AB}\rvert$	Effect	95% CI
L5	Similarity	20	-1.040	$[-1.933, -0.275]$
L5	CDS	15	-1.656	$[-2.861, -0.518]$
L5	Random	20	-1.350	$[-2.156, -0.604]$
L10	Similarity	90	-0.070	$[-0.254, 0.180]$
L10	CDS	70	-0.245	$[-0.981, 0.331]$
L10	Random	90	-0.239	$[-0.873, 0.249]$

D.2 Pair-balanced estimand

For directed pair p and scorer seed s, M_{ps}^{m} is accepted S-URR. Budgets are matched target-wise, not pair-wise; a method may accept no events for a pair. For each contrast, fix P_{AB} to pairs with finite accepted means for both methods in all five scorer seeds (42–46), and compute

$$d_{ps} = M_{ps}^{A} - M_{ps}^{B}, \qquad D_s = \frac{1}{|P_{AB}|} \sum_{p \in P_{AB}} d_{ps}, \qquad \hat{\theta}_{\text{pair}} = \frac{1}{5} \sum_{s} D_s. \tag{28}$$

Resampling whole shared scorer-seed blocks preserves cross-pair dependence. We enumerate all $5^5 = 3,125$ ordered bootstrap resamples and use percentile intervals. Five seeds limit precision; exhaustive enumeration eliminates simulation noise, not finite-sample uncertainty or bias.

Point estimates reproduce the original pair-balanced summaries; intervals replace pair–then–seed resampling, which can break cross-pair dependence. These estimands differ from pooled-event averages and, for CDS, exclude empty-acceptance pairs; missing accepted means are not filled with zero. The result supports a fixed-suite, contract-specific mean contrast, not uniform benefit for every pair or new task. No joint superiority claim is made across the six comparisons. Native inputs are `metaworld/pair_inference/pair_seed_metrics.csv` and `metaworld/pair_inference/pair_effects.csv`.

Table 8: Post-hoc pair-concentration sensitivity. The trimmed estimator is diagnostic and never replaces the complete-data estimate.

Set	Baseline	All pairs	Drop 10%	Abs. mass
L5	Similarity	-1.040	-0.121	85.4%
L5	CDS	-1.656	-0.117	92.6%
L5	Random	-1.350	-0.187	85.7%
L10	Similarity	-0.070	-0.017	82.0%
L10	CDS	-0.245	-0.121	74.8%
L10	Random	-0.239	-0.132	69.9%

Table 9: RCER versus Exact-Simple-A under matched information. AUROC is higher-better; S-URR is lower-better.

Set	Panel	RCER / Simple-A AUROC	RCER / Simple-A S-URR
L5	Aggregate	.542 / .770	.452 / .458
L5	Event	.794 / .819	.436 / .428
L10	Aggregate	.538 / .836	1.210 / 1.029
L10	Event	.884 / .894	2.264 / 1.992

D.3 Concentration and influence

The top-10% concentration analysis ranks pairs by $|d_p|$. The largest-magnitude 10% consist of two pairs on L5, nine on L10 for Similarity/Random, and seven on L10 for CDS. They account for 85.4–92.6% of $\sum_p |d_p|$ on L5 and 69.9–82.0% on L10.

The absolute-mass statistic is not the fraction of beneficial effects or the fraction of the signed mean. The trimming rule is outcome-dependent and post hoc.

E Measurement Sensitivity and Attribution Controls

E.1 Exact-Simple-A

Exact-Simple-A receives the same target/source representation, Operator-A counterfactual labels, training events, and realized selection budget as the corresponding RCER audit, but uses a shallow scalar MLP. It is an audit-time ranking/selection control, not a formal rollout controller.

The control shows that the complex architecture is not necessary for the observed audit-time ranking under the matched information set. It does not identify the separate causal contribution of supervision, representation, optimization, or model capacity.

E.2 Within-family and cross-specification ranking

At the final checkpoint, canonical TD ordering agrees strongly with held-out Bellman-error ordering: Spearman correlation is 0.989 on L5 and 0.991 on L10 (Kendall 0.916 and 0.922). This establishes a local stability domain among closely related Bellman-family measurements.

The stronger claim of measurement-invariant fine ranking is not supported. Normalized A$'$ auditors can be internally coherent while correlating near zero with canonical A0 ordering. On the L10 aggregate and event panels, A0–A$'$ Spearman correlations are 0.001 and -0.068, respectively. B0 can additionally change selector ordering under an independently specified target.

E.3 Held-out policy-checkpoint qualification

Fresh policy checkpoints 47–56 test the frozen A$'$ auditor away from the training checkpoint population. Prediction–target Spearman remains approximately 0.87–0.90 across L5/L10 and Huber/MSE

variants. However, only approximately 9–23% of evaluated cells improve on a training-mean predictor in absolute prediction error.

These results are intentionally interpreted separately: relative prediction-target ordering can remain strong while absolute prediction quality remains weak. They do not establish that the induced forced-ON/OFF degradation differences are themselves accurately ranked.

E.4 Binary incidence and continuous severity

In L25/L49 and controlled diagnostic pools, positive incidence remains 1.0 through margin $\delta = 0.1$ and approximately 0.957–0.980 at $\delta = 1$, whereas continuous accepted p_{95} spans approximately 33.57–65.24. This demonstrates saturation of the binary incidence statistic under those diagnostic contracts, not a universal superiority theorem for continuous severity.

F Deployment Contracts and Source Context

F.1 CQL boundary test

CQL is used as a secondary process/deployment boundary test rather than as evidence for a universal controller ranking. The fixed-budget batch evaluates reported reuse coverage 0.75 for the compared methods. These are the original fixed-budget batch, not the autonomous current-15-seed CQL summaries included separately in the supplement; the latter have different realized reuse rates and are not pooled with this comparison.

Table 10 reports the process and rollout summaries.

Table 10: Fixed-budget CQL process and rollout summaries. The original batch uses reuse coverage 0.75; lower S-URR and higher return are favorable.

Method	S-URR	Mean rollout return
Exact-Simple-A	10.218	60.915
Similarity	25.398	60.926
Random	66.260	60.614
RCER forced .75	71.725	61.531
CDS	82.770	60.686

The process and rollout point estimates do not induce the same ordering. Paired rollout-return contrasts over the 15 matched training seeds further quantify this boundary. Using RCER as the reference,

$$\begin{aligned}
\Delta_{\mathrm{Avg}}(\mathrm{RCER} - \text{Exact-Simple-A}) &= +0.6156, \quad 95\%\ \mathrm{CI} = [0.0236, 1.2076],\\
\Delta_{\mathrm{Avg}}(\mathrm{RCER} - \mathrm{CDS}) &= +0.8449, \quad 95\%\ \mathrm{CI} = [-0.0308, 1.7205],\\
\Delta_{\mathrm{Avg}}(\mathrm{RCER} - \mathrm{Similarity}) &= +0.6052, \quad 95\%\ \mathrm{CI} = [-0.0911, 1.3015],\\
\Delta_{\mathrm{Avg}}(\mathrm{RCER} - \mathrm{Random}) &= +0.9167, \quad 95\%\ \mathrm{CI} = [-0.0526, 1.8861].
\end{aligned}$$

Only the RCER–Exact-Simple-A return contrast excludes zero; the other three intervals remain unresolved. Together with the differing process-side ordering, these results do not establish a general mapping from lower process degradation to better deployment performance. The CQL experiment therefore remains a boundary test rather than validation of a process-to-deployment bridge.

F.2 Source-context evaluation

The canonical RCER-Agg evaluator uses target-copy representations in source slots. The prototype condition instead constructs source representations from the actual source task. The self-only condition removes source context while retaining the same policy checkpoint.

On L10,

$$\text{prototype} - \text{target-copy} = +0.832, \qquad 95\%\ \text{CI} = [-8.503, 10.168], \tag{29}$$

$$\text{prototype} - \text{self-only} = +61.975, \qquad 95\%\ \text{CI} = [24.923, 99.027]. \tag{30}$$

The first interval does not establish equivalence; it only fails to resolve a large aggregate target-copy effect at the available precision. The second shows a favorable source-context contrast relative to self-only inference under the evaluated checkpoint.

The separate L5 re-evaluation uses 15 policy seeds. Prototype, target-copy, and self-only mean returns are 345.729 ± 14.655, 343.744 ± 14.696, and 313.700 ± 15.024 (mean $\pm$ SE). Paired prototype minus target-copy is $+1.985$ (95% CI $[-1.034, 4.990]$); prototype minus self-only is $+32.028$ ($[14.580, 50.331]$). These are source-context *rollout* contrasts, not replacements for scorer-seed process contrasts or the canonical outcome table. We do not use this batch's NTR columns to strengthen the conclusion: some rows report zero NTR with positive NTM, inconsistent with the zero-threshold definition in Appendix A.

F.3 Incomplete prospective bridge

The dedicated prospective bridge from local operational degradation to future Meta-World rollout harm was not completed under the final protocol. Its audit status is therefore UNRESOLVED. No missing result is converted into negative evidence.

G ManiSkill3 Contracts and Full Results

G.1 Comparability contract

The supplied contract records common candidate support, target observations, action/reward/next-state/done values, and a frozen target critic. It also records *method-specific* source transformations and checkpoint semantics. RCER uses `no_reuse_feature` for source-OFF; Similarity fuses the target feature with a zero source feature. Source-ON uses each method's source encoder and fusion. A common target-loss convention and accepted count therefore do not imply identical per-event intervention losses. Table 11 summarizes these relationships.

Table 11: ManiSkill3 process comparability contract. Shared target-side quantities do not imply identical source-ON/OFF intervention losses.

Contract field	Recorded relationship
Candidate support; target observations; target critic	Shared
Loss definition; units; accepted count	Shared convention; reuse fraction 0.75
Source-ON/OFF transform; source/fusion checkpoint	Method-specific
Process-to-deployed-policy link	Method-specific; not a causal bridge

This permits comparison of complete process pipelines, not isolation of selector quality under identical event losses. Accepted and rejected means imply full-pool severity $c\bar{\ell}_{\text{acc}} + (1-c)\bar{\ell}_{\text{rej}}$ of 173.987 for RCER versus 868.984 for Similarity at $c = 0.75$. Each agrees with its own random partition, not the other method's pool, consistent with the method-specific transformations. Within-panel random comparisons are reported separately. Raw per-event accepted-mask lineage remains incomplete.

G.2 Process results

The main process analysis uses matched `reuse_coverage` 0.75. Under the shared target-loss convention but method-specific transformations, the reported RCER–Similarity process-pipeline contrasts are:

$$\Delta\text{URR} = -0.0788, \qquad 95\%\ \text{CI} = [-0.1018, -0.0388], \tag{31}$$

$$\Delta\text{S-URR} = -964.4, \qquad 95\%\ \text{CI} = [-1567.8, -444.7]. \tag{32}$$

Table 12: D4RL frozen action-MSE diagnostic over 15 seeds. Values are mean ± reported SE.

Method	Mean MSE ↓	Worst MSE ↓	NTR ↓	NTM-MSE ↓
Independent	0.4408 ± 0.0176	0.8114 ± 0.0555	0.0000 ± 0.0000	0.0000 ± 0.0000
Full Sharing	0.4253 ± 0.0462	0.6536 ± 0.0679	0.5333 ± 0.0783	0.0856 ± 0.0237
Shared Encoder	0.3180 ± 0.0100	0.6069 ± 0.0273	0.3111 ± 0.0761	0.0181 ± 0.0063
Similarity Routing	0.2396 ± 0.0056	0.5169 ± 0.0126	0.0222 ± 0.0222	0.0005 ± 0.0005

Complete accepted-p_{95} and severe-accepted contrasts are provided in the artifact and full result table.

Random-matched controls are reported separately because they answer a different question: whether the observed process summary can be explained solely by the accepted count within each panel. RCER minus its random-matched control has S-URR difference -120.954 (95% policy-seed CI $[-153.484, -90.360]$); Similarity minus its own random control is $+150.608$ ($[87.529, 213.687]$). The cross-pipeline contrast is not attributed to the selector alone.

G.3 Deployment protocol

The deployment study uses five policy seeds and three locked rollout offsets per seed. Offsets are treated as repeated evaluations nested within policy seed; policy seed is the clustering unit.

Relative to Similarity, RCER-Abstain has

$$\Delta\mathrm{Avg} = -0.113, \qquad 95\%\ \mathrm{CI} = [-0.331, 0.106], \tag{33}$$

$$\Delta\mathrm{NTR} = -0.267, \qquad 95\%\ \mathrm{CI} = [-0.489, -0.044]. \tag{34}$$

The mean-return contrast is unresolved, while the target-level degradation contrast favors the RCER-Abstain deployment under this contract.

Relative to unwrapped RCER, RCER-Abstain increases mean return by $+0.342$ (95% CI $[0.085, 0.567]$) and reduces NTR by -0.578 ($[-0.711, -0.422]$).

Because abstention may dispatch to Independent, these deployment contrasts are properties of the complete wrapper, not causal evidence for the local process score alone. The artifact therefore reports realized reuse rate, fallback rate, and the frozen calibration rule for every policy seed.

Task success is near zero for all evaluated methods. The experiment does not establish task-solving superiority.

H D4RL Diagnostic Results

D4RL is used only as an external frozen action-prediction diagnostic. It is not a rollout-validation experiment.

The evaluated domains are Hopper, Walker2d, and Reacher. Action MSE is lower-better. For each target and seed, NTR is defined relative to the paired Independent action-MSE reference: a target counts as degraded if its MSE exceeds the corresponding Independent MSE.

Relative to Independent, Shared Encoder improves mean MSE by -0.1228 with 95% CI $[-0.1658, -0.0796]$, while its NTR remains 0.3111 with CI $[0.1778, 0.4667]$. Full Sharing has mean-MSE difference -0.0155 ($[-0.1015, 0.0704]$) and NTR 0.5333 ($[0.3778, 0.6667]$).

The bounded positive claim retained by the audit is that Similarity Routing has the lowest observed mean MSE among the evaluated methods on this frozen diagnostic. No action-MSE result is promoted to rollout superiority.

I Claim–Evidence Matrix

Table 13 maps the principal comparative claims in the paper to their decisive evidence, audit status, and remaining boundary. The purpose is to make clear which experiment licenses which conclusion and to prevent evidence from one scope being silently promoted to another.

Table 13: **Final claim–evidence matrix.** Statuses refer to the explicitly stated claim scope, not to universal method quality.

Claim	Most direct evidence	Status	Remaining boundary
Stronger implemented audit configurations can reduce directional claim error at matched claim coverage	Finite-MDP exact process oracle; equal `claimed_decision_coverage`; problem-seed bootstrap	SUPPORTED	Selected operating points; P3 coverage limitation; fixed selected-set inference.
P3 is the best same-information directional selector	Agreement/minimum-confidence control at 2,593 claims	NOT SUPPORTED	5 versus 103 errors; the control admits items outside P3's qualification set.
The critic-local score provides precise event-level downstream ranking	Exact $\Delta J_{\mathrm{exact}}$ comparison	NOT SUPPORTED	Fine Spearman is weak; only coarse consequence separation is retained.
High-score controlled strata contain worse exact downstream outcomes	High/low exact-consequence contrasts across primary finite-MDP regimes	SUPPORTED	Evaluated generators and intervention contracts only.
Higher average Meta-World return implies target-wise benefit	15-seed IQL outcome and paired Independent deficits	NOT SUPPORTED	Mean and target-wise estimands are distinct.
RCER-Event has uniformly better source selection	Matched process panel, pair-balanced inference, concentration sensitivity	CONDITIONAL	L5 shared-seed pointwise evidence; L10 unresolved; complete nonempty pairs only.
RCER architectural complexity is necessary for the observed audit ranking	Exact-Simple-A matched-information control	NOT SUPPORTED	Does not identify separate causal contribution of supervision, representation, or optimization.
Canonical event ranking is invariant across measurement specifications	TD/Bellman, A′, B0	CONDITIONAL	Strong within-family agreement, but fine cross-specification ordering changes.
Lower process degradation implies better Meta-World deployment	CQL boundary test and source-context evaluation	UNRESOLVED	No completed prospective process-to-rollout bridge.
ManiSkill3 process comparison favors RCER	Matched count and target critic; method-specific ON/OFF transforms	CONDITIONAL	Joint process-pipeline contrast, not isolated selector quality.
RCER-Abstain has higher mean ManiSkill3 return than Similarity	Policy-seed-clustered rollout comparison	UNRESOLVED	Mean-return CI crosses zero; task success near zero.
RCER-Abstain has lower ManiSkill3 target degradation than Similarity	Policy-seed-clustered NTR comparison	SUPPORTED	Property of complete deployment/fallback contract, not process score alone.
Similarity has the strongest D4RL diagnostic result	Frozen Hopper/Walker2d/Reacher action-MSE panel	SUPPORTED	Among evaluated methods on action-MSE only; no rollout claim.
Local degradation prospectively predicts future Meta-World rollout harm	A1 prospective bridge	UNRESOLVED	Final protocol incomplete; neither support nor refutation.