FROM MEMBERSHIP SCORES TO UNLEARNING CLAIMS IN OFFLINE REINFORCEMENT LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Machine unlearning aims to remove the influence of selected training data by matching a model retrained without them. Membership scores offer a convenient evaluation tool, but interpreting their changes requires knowing which properties of retraining they measure. We study this connection through controlled experiments in offline reinforcement learning, where simulator access and training records allow us to separate policy compatibility from training exposure. Naturally generated non-members receive member-like scores under a fixed audit, while interventions on reference selection and target adaptation reveal how additional information changes the conclusions of the same score family. We then examine whether score-based comparisons can recognize retain-only retraining itself. Two candidate audit designs frequently reject retraining controls despite consistently rejecting full-data comparators, showing why comparator separation alone does not validate an equivalence test. We explain these findings by distinguishing model-distribution distance from distances between individual models and by characterizing information lost in finite action-error scores. A three-action decision rule connects valid uncertainty bounds to support, rejection, or abstention, and a practical protocol specifies the claim, access assumptions, independent samples, and controls needed to apply it. The resulting case study shows how membership evidence can remain useful for diagnosis while supporting retraining-equivalence claims only through an explicitly validated statistical connection.

1 INTRODUCTION

Machine unlearning seeks to remove the influence of selected training data without repeating the full cost of training (Cao & Yang, 2015; Bourtoule et al., 2021; Nguyen et al., 2025). Its natural reference is a model trained from scratch on the retained data. Evaluation must therefore answer a counterfactual question: how would the learning procedure behave had the forget set been excluded? Membership inference offers an accessible proxy, since it tests whether forgotten examples still look different from non-members. The difficulty is deciding how much a change in this proxy establishes about the retraining reference.

Consider a known forget set and three models: the unchanged full-data model, an independent retain-only retrain, and a candidate obtained by refitting on the retained data. Membership of the forget set in the original training data is already known. The question is whether the candidate is sufficiently close to retraining, not whether deletion changes that historical fact. A useful evaluation must connect its observable scores to this comparison and distinguish evidence against equivalence from evidence supporting it.

This connection involves two separate checks. We must understand what the score measures, including how non-members, references, and target adaptation affect it. We must also validate the comparison rule against the retraining behavior it is intended to recognize. A score may distinguish selected cohorts without isolating training exposure, and a rule may reject the unchanged model while also rejecting independent retrains. In either case, separation alone leaves the unlearning question unanswered.

Figure 1 gives an overview of the measurement and retraining questions. We investigate these checks in offline Reinforcement Learning (RL), which provides complete simulator trajectories, recorded training membership, feasible retraining, and a separate measure of task utility (Fu et al., 2020;

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.

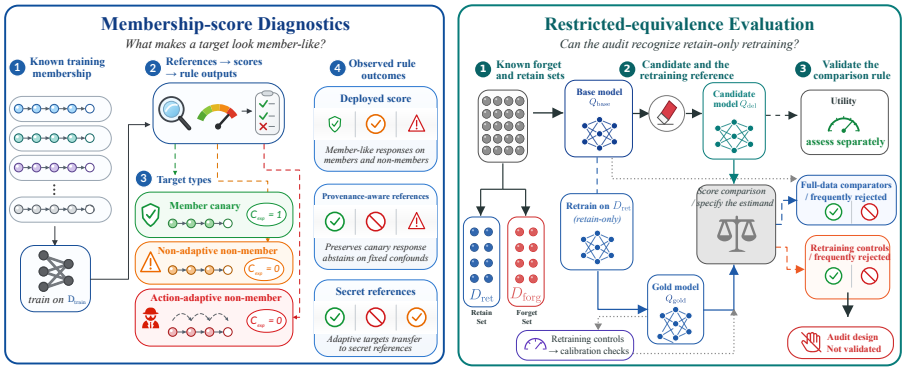


Figure 1: Two questions for interpreting membership evidence. Left: controlled targets and reference information reveal what membership scores measure. Right: a known forget set defines the comparison with retain-only retraining, with utility assessed separately. Exposure diagnosis is not a prerequisite for this comparison. Score differences and threshold outputs require validation against retraining controls before they support an equivalence decision.

Levine et al., 2020). These features let us vary audit information while keeping the scoring mechanism controlled. Our experiments concern action-error evidence on locomotion tasks, rather than an evaluation of all membership attacks. Policy generalization is especially relevant here: accurately reproducing a trajectory can reflect compatible behavior even when the trajectory never entered training. We use that mechanism to examine what reference calibration does and does not resolve.

Our evidence connects measurement behavior to retraining comparisons. Natural non-members expose low specificity at a fixed threshold. Reference provenance removes the evaluated non-adaptive confounds, but action-adaptive targets transfer to secret references. Finite-score geometry explains how distinct actions can preserve these observations. Turning to retraining, two candidate audit designs reject many retain-only controls as well as full-data comparators. This result reveals a gap between comparator separation and equivalence testing. A directly observed constructive control confirms that the decisive branches of our decision rule are attainable when the audited quantity is measured exactly, so the frequent rejection of retraining controls reflects the information content of heuristic scores rather than excessive caution of the rule itself. We explain why average distances between individual models need not measure the distance between randomized training distributions.

The general concern that membership inference does not prove training participation is established by Zhang et al. (2025). Retraining-grounded evaluation and selective auditing also have substantial precedents (Tu et al., 2025; Yang & Yeung, 2026). Our contribution is a controlled empirical account of how the evidence changes with natural target generation, reference provenance, and adaptation, followed by a direct test of retraining-control behavior. We use standard statistical principles to connect these observations to a concrete evaluation protocol, not to propose a universal certification method. This paper makes three contributions:

1. We provide a dynamics-valid stress test and a sequence of information-access interventions that distinguish policy compatibility from historical exposure in action-error audits.
2. We show that two candidate retraining audits reject many retraining controls, and explain why pairwise model distances and full-data rejection cannot by themselves establish algorithm-distribution equivalence or non-equivalence.
3. We organize these findings into a claim-specific evaluation protocol with explicit sample units, controls, and a three-action rule. Supporting results characterize the information and uncertainty conditions needed for decisive conclusions.

2 What Is Being Evaluated?

Retraining as the reference. An offline RL learner receives a trajectory multiset $\mathcal{D}$. For a known forget set $\mathcal{D}_f \subset \mathcal{D}$, let $\mathcal{D}_r = \mathcal{D} \setminus \mathcal{D}_f$. Conditional on these data and fixed learning recipes, training on $\mathcal{D}$, applying deletion, and independently retraining on $\mathcal{D}_r$ induce model distributions Q_{base}, Q_{del},

and Q_{gold}. These distributions include training randomness. A fixed checkpoint is instead a single model m, represented by the point mass δ_m when needed.

For a pre-specified class $\mathcal{F}$ of measurable $[0,1]$-valued model tests, define

$$d_{\mathcal{F}}(Q, Q') = \sup_{f \in \mathcal{F}} |\mathbb{E}_{m\sim Q} f(m) - \mathbb{E}_{m\sim Q'} f(m)|. \tag{1}$$

We study the restricted-equivalence claim $C^{\eta}_{\mathcal{F}\text{-eq}} = \mathbf{1}\{d_{\mathcal{F}}(Q_{\text{del}}, Q_{\text{gold}}) \leq \eta\}$ for a tolerance η fixed independently of the candidate evaluation. The class specifies which properties must agree. All measurable indicator tests recover total variation, $\mathrm{TV}(Q, Q') = \sup_B |Q(B) - Q'(B)|$, whereas tests built from action scores support a narrower comparison. Utility is a separate requirement: writing C^{u}_{util} for a pre-specified performance condition gives the useful-unlearning claim $C^{\eta,u}_{\text{useful-}\mathcal{F}\text{-unlearn}} = C^{\eta}_{\mathcal{F}\text{-eq}} C^{u}_{\text{util}}$.

Three levels of measurement. A model distribution, a fixed model, and the trajectory scores produced by that model are different statistical objects.

$$e(m, \tau) = \frac{1}{T d_a} \sum_{t=1}^{T} \|m(s_t) - a_t\|_2^2, \tag{2}$$

where $\tau = (s_t, a_t)_{t=1}^{T}$ consists of recorded states s_t and actions a_t, $m(s_t)$ is the action predicted by model m, T is trajectory length, and d_a is action dimension. Lower error means closer reproduction of recorded actions. Increasing the number of trajectories improves measurement of a score distribution conditional on the model, but does not supply additional independent training runs. In contrast, estimating Equation 1 for randomized algorithms requires information about variation across models. Section 4 examines the consequences of confusing these levels.

Exposure as a separate question. An external auditor without training records may instead ask whether a trajectory entered training. For a setting W with executed training multiset $\mathcal{D}_{\text{train}}(W)$, this claim is

$$C_{\text{exp}}(W, \tau) = \mathbf{1}\{\tau \in \mathcal{D}_{\text{train}}(W)\}. \tag{3}$$

Deletion leaves this historical fact unchanged, so exposure inference is not a prerequisite for evaluating a known forget set. We use controlled exposure tests to investigate the interpretation of membership scores, with experimenter-held training records supplying ground truth. When an audit can read reference training records, we state that additional access explicitly.

From observations to decisions. An audit protocol selects references, computes scores and other permitted observations, and applies a decision rule. We call everything visible to that rule its transcript Z. A setting W specifies the data, learning recipes, randomness laws, target generator, and information access, and induces a transcript law P^{Π}_{W}. A declared class $\mathcal{W}$ specifies the settings covered by a claim. We use three actions: *authorize* (A) supports the stated claim, *reject* (R) supports its negation, and *abstain* ($\perp$) leaves it unresolved, giving decision rules $\delta(Z) \in \{\mathsf{A}, \mathsf{R}, \perp\}$. An observed output of a candidate rule is not itself evidence that the rule has valid error control.

3 What Do Membership Scores Measure?

A score used in unlearning evaluation should first have a clear interpretation. We examine action-error evidence using natural trajectories, then change reference information and target access to isolate its behavior. The natural-trajectory experiment tests a fixed operating point. The subsequent fixed-observation constructions explain mechanisms under a broader target class. Appendix B specifies scoring and matching, and Appendices C–D give the complete protocols.

3.1 Natural Non-Members at a Fixed Operating Point

The first experiment asks whether member-like evidence can arise on complete simulator trajectories that never entered training, which we call *natural non-members*. We use the Walker2d-medium-replay-v2 task from the D4RL offline RL benchmark (Fu et al., 2020) and trajectories generated by

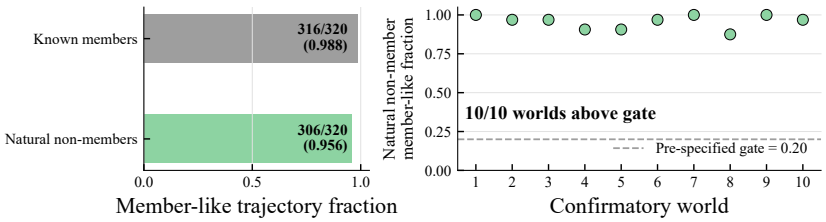


Figure 2: Natural trajectories expose low specificity at the fixed member-like threshold. The experiment uses ten confirmatory target models and a shared eight-model reference bank. Of 320 trajectories in each class, 316 members and 306 natural non-members exceed the threshold. The world-level gate tests whether the non-member response is consistently high under this recipe, not whether an unknown trajectory was in training.

Table 1: Post-hoc score analysis of natural trajectories. Pooled AUC includes cross-world pairs. Within-world AUC compares scores from the same target model. All values are descriptive and retain the frozen threshold.

Statistic	**Value**
Frozen-threshold balanced accuracy	0.516
Pooled member–natural AUC	0.584
Mean within-world AUC	0.590
Within-world AUC range	0.479–0.780

a pre-trained Twin Delayed DDPG with Behavioral Cloning (TD3+BC) policy in the simulator (Fujimoto & Gu, 2021). Each trajectory follows consecutive simulator steps. Eight calibration target models set a cutpoint using member trajectories, and ten separately trained confirmatory target models each evaluate 32 members and 32 natural non-members. The reference bank contains eight fixed models. Natural non-members are absent from target training sets.

The evidence is $s(\tau) = \frac{1}{8}\sum_{r=1}^{8} e(m_r, \tau) - e(m_0, \tau)$, where m_0 is the target model and $m_1, \ldots, m_8$ are models from the reference bank. Larger values are more member-like. The pooled fifth percentile of calibration-member evidence fixes the threshold at approximately 0.00456. This construction aims for high member response, so its specificity must be tested on non-members rather than inferred from its calibration.

Natural non-members trigger nearly as often as members (Figure 2). All ten confirmatory worlds exceed the pre-specified non-member-rate gate of 0.20. The frozen threshold has balanced accuracy 0.516, demonstrating that high member response does not provide a reliable exposure decision under this recipe.

The full scores retain more information than this operating point suggests. A post-hoc descriptive analysis gives pooled member-versus-natural area under the receiver operating characteristic (ROC) curve (AUC) 0.584 and mean within-world AUC 0.590, with within-world values ranging from 0.479 to 0.780 (Table 1). AUC counts the probability that a member receives higher evidence than a natural non-member, assigning half weight to ties. These statistics neither retune the threshold nor treat pooled trajectories as independent training runs. Together, the threshold and ranking results support a specific interpretation: compatibility with the learned policy produces frequent member-like evidence on natural non-members, even though some worlds retain ranking signal. This is a measurement finding, not a verdict on a deletion operator.

3.2 Reference Provenance Changes the Interpretation

To isolate why scores become member-like, we hold observations, rewards, and lengths fixed and vary actions. The study uses Walker2d and HalfCheetah with five seeds per environment. Policies are two-layer 256-unit ReLU networks trained for 100,000 gradient steps and evaluated over 25 deterministic episodes (Appendix H, Table H.1). Injected canaries enter training as positive controls, with each canary repeated 16 times so that the positive signal is unambiguous. Retain-only controls

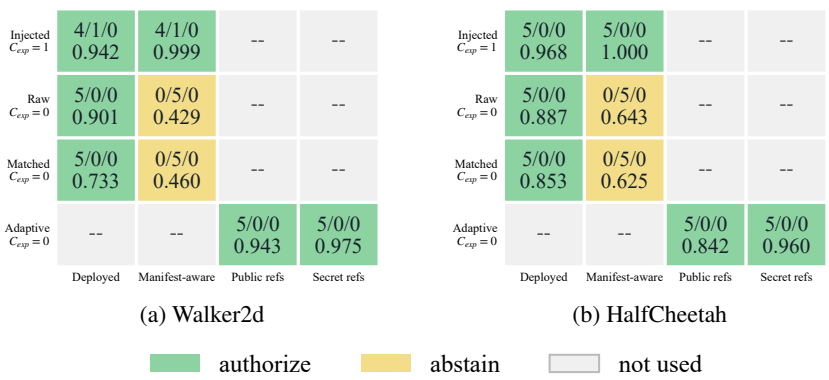


Figure 3: Information-access interventions on fixed-observation targets. Cells report authorize/abstain/reject counts over five seeds and mean signed AUC. Reference training records remove authorization of the non-adaptive cross-model confounds while preserving the canary response. Action-adaptive targets remain authorized with public or independently trained secret references. These are outputs of the evaluated exposure rules, not retraining-equivalence decisions.

exclude them, while two cross-model constructions use actions from an Implicit Q-Learning (IQL) teacher policy (Kostrikov et al., 2022), either directly or after matching action marginals. Training records establish the membership labels independently of the scoring rule.

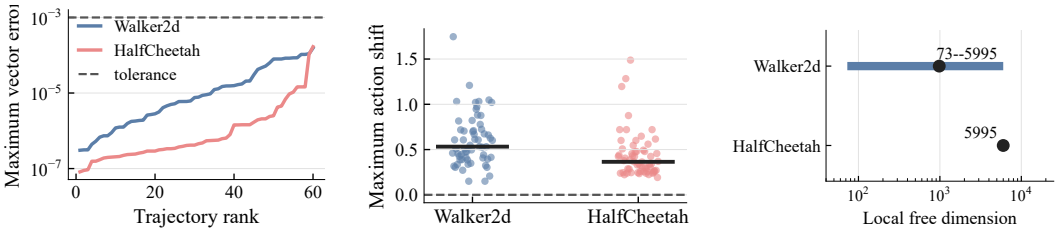
Raw action-error evidence responds to both training exposure and compatible teacher actions. In Figure 3, the candidate exposure rule authorizes both cross-model constructions in every environment-by-seed cell. Its signed AUC values are strongly member-like despite known absence from training (Appendix D, Table D.1). Signed AUC fixes the member-greater-than-non-member orientation in advance without folding, so values above 0.5 indicate member-like evidence. The rule uses directional AUC intervals, paired label swaps, and a utility check against a fixed floor (the larger of 20 and 40% of a reference D4RL normalized score, where 100 corresponds to expert performance; see Appendix B), with its exact form in Appendix D.

Reference calibration can remove these particular confounds when reference membership is checked. A role-only selector uses model and training-role metadata, which does not ensure that a target was excluded. A manifest-aware selector additionally reads reference training fingerprints and retains only target-absent checkpoints. Standardizing the deployed error against these references changes the fixed cross-model decisions from authorization to abstention while preserving canary behavior (Figure 3). The selector accesses reference records, not the deployed training membership label. Thus, this improvement uses additional information beyond scores. Appendix D (Table D.5) reports the 210 reference-absence checks.

3.3 Adaptation and Information Lost in Finite Scores

Verified reference absence addresses reference semantics, but target access introduces another source of ambiguity. We optimize actions against the deployed model and four public references, then evaluate the fixed targets against four independently trained secret references from the same recipe. Observations, rewards, and trajectory lengths remain unchanged. The resulting targets receive authorization in every evaluated cell under the secret-reference rule (Figure 3 and Appendix D, Table D.6). This demonstrates transfer within the evaluated action-only construction, without requiring access to the withheld checkpoints.

Finite score vectors explain how this ambiguity can persist even when several models are observed. On fixed observations, concatenate actions as $a \in \mathbb{R}^D$ and model predictions as p_j. The score $g_j(a) = D^{-1}\|a - p_j\|_2^2$ records a squared distance to one prediction vector. Matching differences between M scores fixes at most $M - 1$ linear directions. Matching the remaining norm can still leave a sphere of distinct action sequences. Section 5 formalizes this geometric observation.



(a) Maximum vector error (b) Maximum action shift (c) Local free dimension

Figure 4: Diagnostics for 120 non-identity, five-checkpoint score-vector collisions. (a) Maximum vector errors lie below the frozen 10^{-3} tolerance. (b) Every trajectory changes its actions, with black bars marking environment medians. (c) Rank-five Jacobians leave substantial local action dimensions outside the score row space, where intervals show observed ranges and dots mark medians.

Every collision reproduces the five-checkpoint score vectors below the frozen 10^{-3} tolerance (Figure 4a), every trajectory changes its actions (Figure 4b), and rank-five Jacobians leave substantial local action dimensions unconstrained (Figure 4c). The construction holds observations fixed and constrains actions to the allowed box (Appendix D).

These interventions identify information loss under fixed observations, rather than demonstrating dynamically reachable attacks. The common lesson is that a score must be interpreted relative to both its reference information and its allowed target class. Additional references can repair a particular comparison without making the finite score vector a record of training history.

4 Can the Audit Recognize Retraining?

For a known forget set, the central comparison is with retain-only retraining. We now ask whether score-based audit rules recognize that reference behavior. Returning to the three-model example, a rule should distinguish the candidate from the unchanged model only when the evidence supports that distinction. Independent retrains provide a direct check of how the rule behaves when the target and reference recipes agree.

4.1 Pairwise Distances Are Not Distributional Distances

A natural construction compares trajectory-score distributions from two checkpoints. Let $H_{m,c}$ be the empirical action-error distribution of model m on cohort c, where c is either the forget cohort or a matched held-out cohort. The tagged Kolmogorov–Smirnov (KS) distance is

$$d_{\text{pair}}(m, m_g) = \max_c \sup_{u \in \mathbb{R}} |H_{m,c}(u) - H_{m_g,c}(u)|, \tag{4}$$

where m_g denotes a reference retain-only retraining model, and 'tagged' indicates taking the maximum discrepancy over the cohort tags c. For a common fixed cohort and threshold class, this is $d_{\mathcal{F}}(\delta_m, \delta_{m_g})$ with tests $f_{c,u}(m) = H_{m,c}(u)$. Averaging this distance over candidate–reference model pairs estimates a different object from Equation 1. In particular,

$$d_{\mathcal{F}}(Q, Q') \le \mathbb{E}_{m\sim Q, m_g \sim Q'} d_{\mathcal{F}}(\delta_m, \delta_{m_g}). \tag{5}$$

The inequality follows by moving the absolute value and supremum inside the expectation, and need not hold with equality. If both model laws place equal probability on two models with pairwise distance one, their distributional distance is zero, but their expected pairwise distance is one half.

This distinction changes how an audit can be used. A valid upper bound on the expected pairwise distance can support a conservative upper bound on distributional distance. A large pairwise distance, however, supplies no corresponding lower bound on distributional distance. Alternatively, an audit may target a fixed candidate model, but that is a conditional claim rather than equivalence of randomized learning procedures.

Table 2: Development-only retraining-control results. Positive pass probability measures authorization of at least five of six slots. A/Ab/R are per-slot authorize/abstain/reject frequencies, reported for both the retraining controls and the full-data comparators. The rank row gives the maximum over eight candidates, none passing validation. Stop denotes the design-level outcome.

Design	Pos. pass prob.	Pos. A/Ab/R	Full-data A/Ab/R	Outcome
Tagged KS interval	0.1340	0.237/0.381/0.382	0.000/0.000/1.000	Stop
Split-conformal rank	0.2885	0.410/0.000/0.590	0.000/0.000/1.000	Stop

4.2 Control Tests of Two Candidate Designs

We evaluate two heuristic designs based on Equation 4: a tagged-KS interval rule and a split-conformal-style rank rule. Both use 75 forget trajectories and 75 held-out trajectories matched by return, length, and initial-state features. The development data contain three models per role. Each of 2,000 simulations resamples model sources and trajectory pairs to fill eight reference slots and six evaluation slots. These are resampling slots, not additional independent trained models. Retain-only models act as positive controls for agreement of the training recipes. Full-data models are comparators, whose training membership alone does not establish that their distributional distance exceeds the tolerance.

The development target is to authorize at least five of six positive-control slots with probability at least 0.8, while keeping their rejection frequency at most 0.2. The full-data comparison applies analogous gates. The tagged-KS rule constructs a t-style interval around a mean of pairwise distances and uses a reference-derived tolerance. The rank design examines eight acceptance/rejection quantile pairs. Appendix E specifies the sampling geometry and thresholds. We test these designs as proposed interpretations of the score, not as established confidence procedures for $d_{\mathcal{F}}(Q_{\mathrm{del}}, Q_{\mathrm{gold}})$.

Both designs reject retraining controls too frequently to support their intended interpretation (Table 2). The probability of authorizing at least five positive slots is only 0.134 and 0.289, below the 0.8 target. Neither family passes development validation.

Full-data comparators are rejected throughout the same simulations. This contrast establishes the central finding: consistently rejecting a comparator does not demonstrate that a rule can recognize retraining. The positive-control results expose a problem with the candidate designs, rather than just conservative abstention on difficult cases. Their design-level outcome is to stop before applying them to an unknown deletion candidate. The pairwise-distance distinction explains why the desired equivalence interpretation requires additional justification.

4.3 What Additional Observations Can Establish

A directly observed comparison illustrates when a decisive result is possible. In a fixed-feature ridge problem, bounded coordinate probes compare two pre-specified numerical models with a retain-only reference. Subtracting the forget sufficient statistics yields probe distance approximately 10^{-11}, below the arithmetic tolerance, while the unchanged full-data solution is separated (Appendix E.2, Table E.1). Direct observation places the downdate below the tolerance and the unchanged model above it, supporting authorization and rejection, respectively. Section 5 gives the general confidence-set rule. The example illustrates sufficient information for a restricted claim, rather than validating the score-based designs.

Actual retain-refit outputs provide a complementary diagnostic. Their mean pairwise tagged distance is 0.276, compared with 0.369 for the unchanged models, but the overlapping bootstrap ranges do not support a calibrated equivalence decision (Appendix F, Tables F.1 and F.2, which also discuss the pairing and utility-calibration limitations). This preserves the distinction between observing a useful change in a metric and establishing the retraining claim that motivates it.

5 CONNECTING EVIDENCE TO A DECISION

The experiments distinguish two requirements: the observations must identify the intended property, and uncertainty must be calibrated for that property. We state the statistical implications here. These results organize the evidence rather than assuming that the empirical score comparisons already satisfy their premises.

5.1 INFORMATION LIMITS AND VALID DECISIONS

Fix a binary claim C, a class of settings $\mathcal{W}$, and an audit protocol Π. For a three-action rule $\delta(Z)$, let FA be the worst-case probability of authorization when $C = 0$, and FR the worst-case probability of rejection when $C = 1$. Define decisive coverage by $\mathrm{Cov}_W = \Pr_W(\delta \neq \perp)$. This coverage measures how often the rule decides, not confidence-interval coverage.

Theorem 1 (Hard-pair bound)**.** *Suppose $C(W_0) = 0$, $C(W_1) = 1$, and $d_{01} = \mathrm{TV}(P^{\Pi}_{W_0}, P^{\Pi}_{W_1})$. Every possibly randomized rule satisfies*

$$\Pr_{W_1}(\delta = \mathsf{A}) \leq \Pr_{W_0}(\delta = \mathsf{A}) + d_{01},$$
$$\Pr_{W_0}(\delta = \mathsf{R}) \leq \Pr_{W_1}(\delta = \mathsf{R}) + d_{01}. \tag{6}$$

If $\mathrm{FA} \leq \alpha$ and $\mathrm{FR} \leq \alpha_{\mathrm{R}}$, then

$$\max_{c \in \{0,1\}} \mathrm{Cov}_{W_c} \leq \min\{1, \alpha + \alpha_{\mathrm{R}} + d_{01}\}. \tag{7}$$

The result follows from the TV bound on event probabilities. When opposite claims produce similar observations, controlling both errors requires frequent abstention on that pair. It is not a bound on every setting. Our finite-score constructions explain information loss but do not estimate the full-transcript TV required to apply this bound numerically.

For a constructive counterpart, let $\psi(W)$ be a parameter and let $\Gamma(\vartheta) \subseteq \{0, 1\}$ contain the claim values compatible with parameter value ϑ. Set $D_c = \{\vartheta : \Gamma(\vartheta) = \{c\}\}$.

Theorem 2 (Confidence-set decision rule)**.** *Assume $C(W) \in \Gamma(\psi(W))$ for all $W \in \mathcal{W}$. If a confidence set $\mathcal{C}_\alpha(Z)$ satisfies*

$$\inf_{W \in \mathcal{W}} \Pr_W\{\psi(W) \in \mathcal{C}_\alpha(Z)\} \geq 1 - \alpha, \tag{8}$$

then authorizing on $\varnothing \neq \mathcal{C}_\alpha \subseteq D_1$, rejecting on $\varnothing \neq \mathcal{C}_\alpha \subseteq D_0$, and abstaining otherwise gives $\mathrm{FA} \leq \alpha$ and $\mathrm{FR} \leq \alpha$, assuming these events are measurable.

For restricted equivalence, take $\psi = d_{\mathcal{F}}(Q_{\mathrm{del}}, Q_{\mathrm{gold}})$. A valid interval $[L, U]$ authorizes when $U \leq \eta$, rejects when $L > \eta$, and abstains otherwise. The statistical work lies in establishing coverage for this distance. A percentile range for another statistic does not suffice. For a finite class of fixed bounded model tests, independent training runs support a standard Hoeffding interval (Appendix G.2). Three models per role generally give an uninformative bound, and repeated resampling cannot replace additional independent models.

5.2 WHY ACCESS AND SCORE CLASSES MATTER

Allowing more target generators reduces worst-case separation between opposite claims. Applying a common observation mechanism to two model distributions likewise contracts TV. These facts explain why access restrictions matter and why score agreement is weaker than full model agreement (Appendix G.1). For action-error scores the information loss has a direct geometric form.

Proposition 1 (Finite action-score ambiguity)**.** *Fix observations and let $g_j(a) = D^{-1}\|a - p_j\|_2^2$ for M frozen prediction vectors. Let $L = \mathrm{span}\{p_j - p_1 : 2 \leq j \leq M\}$ have dimension r. For any target action sequence $a_\star \in \mathbb{R}^D$, if the $L^\perp$ component of $a_\star - p_1$ is nonzero and $D - r \geq 2$, then $g^{-1}(g(a_\star))$ contains a sphere of dimension $D - r - 1$.*

Thus, even exact agreement of all finite scores can leave many actions undetermined. Dynamics and action constraints require separate analysis. The AUC mixture identity in Appendix G.1 makes a related point at the cohort level: interpreting AUC as exposure prevalence requires the non-member comparison law to represent the non-exposure component. These facts specify what additional assumptions are needed to interpret membership scores, without making them useless.

Table 3: Protocol card for interpreting an unlearning audit. Each field specifies part of the statistical comparison, not an additional certificate. A/Ab/R denote authorize, abstain, and reject frequencies.

Field	Required specification
Claim	Historical exposure, a fixed-model comparison, or algorithm-distribution equivalence.
Access	Training recipe, target generator, visible reference records, and public or secret models.
Measurement	Score, cohorts, test class, conditioning, and the exact parameter being estimated.
Uncertainty	Independent model and trajectory units, interval justification, and a fixed tolerance.
Controls	Retraining-control A/Ab/R rates, comparator rates, and grounds for any negative label.
Decision	Apply the interval rule only with justified coverage. Otherwise report diagnostics.
Utility & changes	Report task performance separately. Reassess calibration when recipes, generators, or access change.

5.3 A Practical Evaluation Protocol

The three-model example yields a concrete workflow (Table 3). Fix the claim, tests, tolerance, and information access before inspecting the candidate. Use independent retrains to validate control behavior, treat full-data models as comparators unless their non-equivalence is established, and report utility separately.

When interval premises are unavailable, scores still measure changes and investigate discrepancies, or refute claimed guarantees via statistical lower bounds. Supporting equivalence, however, requires an upper bound on distance rather than failing to find differences. This distinction gives membership evidence a defined role without treating it as a universal proxy for deletion. The stop outcomes in Section 4 illustrate this function: both candidate designs halt during development validation instead of authorizing, which is the intended behavior for rules whose coverage premises are unestablished.

6 Related Work and Discussion

From membership evidence to unlearning guarantees: Membership attacks measure distinguishability under specified access and non-member distributions (Shokri et al., 2017; Carlini et al., 2022; Watson et al., 2022). Zhang et al. (2025) explain why this differs from proving training participation. Tu et al. (2025) develop a game-based, retraining-grounded framework for empirical unlearning evaluation. Positive guarantees are possible with additional structure (Sepahvand et al., 2025; Sarmasarkar et al., 2026). These results separate rejecting an unsupported guarantee from establishing retraining equivalence, a distinction central to our protocol.

Recent work closely aligns with our focus (Yang & Yeung, 2026; Li et al., 2026): retraining controls, abstention, and retrain-calibrated evaluation. We contribute controlled interventions on natural trajectory generation, reference training information, and target adaptation, analyzing the statistical object measured by candidate score-based rules. This complements trajectory auditing and deletion in offline RL (Du et al., 2024; Gong et al., 2025; Ye et al., 2025) by examining the interpretation of evaluation evidence rather than proposing new deletion operators.

Our empirical findings concern action-error evidence in the evaluated locomotion settings. Appendix I states the resulting limits on generality and on statistical interpretation.

7 Conclusion

Membership evidence is most effective when aligned with the evaluated unlearning claim. Across controlled offline RL settings, natural non-members, reference provenance, and target adaptation substantially alter action-error evidence, while rejecting full-data comparators coexists with frequent rejection of retraining controls. These findings motivate a direct evaluation principle: specify the target property, validate decision rules against it, and separate diagnostic changes from calibrated support for equivalence, clarifying what is required to substantiate unlearning claims.

REFERENCES

Lucas Bourtoule, Varun Chandrasekaran, Christopher A. Choquette-Choo, Hengrui Jia, Adelin Travers, Baiwu Zhang, David Lie, and Nicolas Papernot. Machine unlearning. In *IEEE S&P*, pp. 141–159, 2021.

Yinzhi Cao and Junfeng Yang. Towards making systems forget with machine unlearning. In *IEEE S&P*, pp. 463–480, 2015.

Nicholas Carlini, Steve Chien, Milad Nasr, Shuang Song, Andreas Terzis, and Florian Tramèr. Membership inference attacks from first principles. In *IEEE S&P*, pp. 1897–1914, 2022.

Lili Chen, Kevin Lu, Aravind Rajeswaran, Kimin Lee, Aditya Grover, Michael Laskin, Pieter Abbeel, Aravind Srinivas, and Igor Mordatch. Decision transformer: Reinforcement learning via sequence modeling. In *NeurIPS*, pp. 15084–15097, 2021.

Georgina Cosma and Axel Finke. Ruler: Representation-level verification of machine unlearning. *arXiv preprint arXiv:2605.27569*, 2026.

Linkang Du, Min Chen, Mingyang Sun, Shouling Ji, Peng Cheng, Jiming Chen, and Zhikun Zhang. ORL-AUDITOR: Dataset auditing in offline deep reinforcement learning. In *NDSS*, 2024.

Justin Fu, Aviral Kumar, Ofir Nachum, George Tucker, and Sergey Levine. D4RL: datasets for deep data-driven reinforcement learning. *arXiv*, abs/2004.07219, 2020.

Scott Fujimoto and Shixiang Shane Gu. A minimalist approach to offline reinforcement learning. In *NeurIPS*, pp. 20132–20145, 2021.

Scott Fujimoto, David Meger, and Doina Precup. Off-policy deep reinforcement learning without exploration. In *ICML*, volume 97, pp. 2052–2062, 2019.

Antonio Ginart, Melody Y. Guan, Gregory Valiant, and James Zou. Making AI forget you: Data deletion in machine learning. In *NeurIPS*, pp. 3513–3526, 2019.

Aditya Golatkar, Alessandro Achille, and Stefano Soatto. Eternal sunshine of the spotless net: Selective forgetting in deep networks. In *CVPR*, pp. 9301–9309, 2020.

Chen Gong, Kecen Li, Jin Yao, and Tianhao Wang. Trajdeleter: Enabling trajectory forgetting in offline reinforcement learning agents. In *NDSS*, 2025.

Laura Graves, Vineel Nagisetty, and Vijay Ganesh. Amnesiac machine learning. In *AAAI*, volume 35, pp. 11516–11524, 2021.

Chuan Guo, Tom Goldstein, Awni Y. Hannun, and Laurens van der Maaten. Certified data removal from machine learning models. In *ICML*, pp. 3832–3842. PMLR, 2020.

Zachary Izzo, Mary Anne Smart, Kamalika Chaudhuri, and James Zou. Approximate data deletion from machine learning models. In *AISTATS*, volume 130, pp. 2008–2016, 2021.

Michael Janner, Qiyang Li, and Sergey Levine. Offline reinforcement learning as one big sequence modeling problem. In *NeurIPS*, volume 34, pp. 1273–1286, 2021.

Ilya Kostrikov, Ashvin Nair, and Sergey Levine. Offline reinforcement learning with implicit q-learning. In *ICLR*, 2022.

Aviral Kumar, Aurick Zhou, George Tucker, and Sergey Levine. Conservative q-learning for offline reinforcement learning. In *NeurIPS*, volume 33, pp. 1179–1191, 2020.

Meghdad Kurmanji, Peter Triantafillou, Jamie Hayes, and Eleni Triantafillou. Towards unbounded machine unlearning. In *NeurIPS*, volume 36, pp. 1957–1987, 2023.

Sergey Levine, Aviral Kumar, George Tucker, and Justin Fu. Offline reinforcement learning: Tutorial, review, and perspectives on open problems. *arXiv preprint arXiv:2005.01643*, abs/2005.01643, 2020.

Jiazhuo Li, Yu Zhang, Yiming Fei, Kangkang Dong, Xiaojun Zhu, Houde Liu, and Jinze Tao. Rethinking demonstration unlearning in imitation learning for robotics. *arXiv preprint arXiv:2608.20784*, 2026.

Sijia Liu, Yuanshun Yao, Jinghan Jia, Stephen Casper, Nathalie Baracaldo, Peter Hase, Yuguang Yao, Chris Yuhao Liu, Xiaojun Xu, Hang Li, et al. Rethinking machine unlearning for large language models. *Nature Machine Intelligence*, 7(2):181–194, 2025.

Milad Nasr, Reza Shokri, and Amir Houmansadr. Comprehensive privacy analysis of deep learning: Passive and active white-box inference attacks against centralized and federated learning. In *IEEE S&P*, pp. 739–753, 2019.

Seth Neel, Aaron Roth, and Saeed Sharifi-Malvajerdi. Descent-to-delete: Gradient-based methods for machine unlearning. In *Algorithmic Learning Theory*, pp. 931–962. PMLR, 2021.

Thanh Tam Nguyen, Thanh Trung Huynh, Zhao Ren, Phi Le Nguyen, Alan Wee-Chung Liew, Hongzhi Yin, and Quoc Viet Hung Nguyen. A survey of machine unlearning. *ACM Transactions on Intelligent Systems and Technology*, 16(5):1–46, 2025.

Maria Rigaki and Sebastian Garcia. A survey of privacy attacks in machine learning. *ACM Computing Surveys*, 56(4):1–34, 2023.

Ahmed Salem, Yang Zhang, Mathias Humbert, Pascal Berrang, Mario Fritz, and Michael Backes. ML-Leaks: Model and data independent membership inference attacks and defenses on machine learning models. In *NDSS*, 2019.

Sahasrajit Sarmasarkar, Anastasia Koloskova, and Sanmi Koyejo. Auditing of unlearning algorithms. *arXiv preprint arXiv:2607.05898*, 2026.

Ayush Sekhari, Jayadev Acharya, Gautam Kamath, and Ananda Theertha Suresh. Remember what you want to forget: Algorithms for machine unlearning. In *NeurIPS*, pp. 18075–18086, 2021.

Nazanin Mohammadi Sepahvand, Anvith Thudi, Berivan Isik, Ashmita Bhattacharyya, Nicolas Papernot, Eleni Triantafillou, Daniel M. Roy, and Gintare Karolina Dziugaite. Leveraging per-instance privacy for machine unlearning. In *ICML*, pp. 53906–53922, 2025. URL `https://proceedings.mlr.press/v267/sepahvand25a.html`.

Yingdan Shi, Sijia Liu, Kaize Ding, and Ren Wang. Tackling fake forgetting through uncertainty quantification. In *ICML*, 2026.

Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *IEEE S&P*, pp. 3–18, 2017.

Liwei Song and Prateek Mittal. Systematic evaluation of privacy risks of machine learning models. In *USENIX*, pp. 2615–2632, 2021.

Thomas Steinke, Milad Nasr, and Matthew Jagielski. Privacy auditing with one (1) training run. In *NeurIPS*, 2023.

Jialong Sun, Zeming Wei, Jiaxuan Zou, Jiacheng Gong, Jie Fu, Chengyang Dong, Heng Xu, Jialong Li, and Bo Liu. SMI: Statistical membership inference for reliable unlearned model auditing. In *AAMAS*, 2026.

Anvith Thudi, Gabriel Deza, Varun Chandrasekaran, and Nicolas Papernot. Unrolling SGD: understanding factors influencing machine unlearning. In *IEEE EuroS&P*, pp. 303–319, 2022.

Yiwen Tu, Pingbang Hu, and Jiaqi Ma. A reliable cryptographic framework for empirical machine unlearning evaluation. In *NeurIPS*, 2025.

Zhenyi Wang, Enneng Yang, Li Shen, and Heng Huang. A comprehensive survey of forgetting in deep learning beyond continual learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(3):1464–1483, 2025.

Lauren Watson, Chuan Guo, Graham Cormode, and Alexandre Sablayrolles. On the importance of difficulty calibration in membership inference attacks. In *ICLR*, 2022.

Sen Yang and Yuen-Hei Yeung. Unlearning as distribution restoration: A controlled counterfactual study, a validated selective screen, and the limits of oracle-free certification. *arXiv preprint arXiv:2607.19442*, 2026.

Dayong Ye, Tianqing Zhu, Congcong Zhu, Derui Wang, Kun Gao, Zewei Shi, Sheng Shen, Wanlei Zhou, and Minhui Xue. Reinforcement unlearning. In *NDSS*, 2025.

Samuel Yeom, Irene Giacomelli, Matt Fredrikson, and Somesh Jha. Privacy risk in machine learning: Analyzing the connection to overfitting. In *IEEE CSF*, pp. 268–282, 2018.

Chenhao Zhang, Muxing Li, Feng Liu, Weitong Chen, and Miao Xu. Unlearning evaluation through subset statistical independence. In *ICLR*, 2026.

Jie Zhang, Debeshee Das, Gautam Kamath, and Florian Tramèr. Position: Membership inference attacks cannot prove that a model was trained on your data. In *IEEE SaTML*, 2025. doi: 10.1109/SaTML64287.2025.00025.

Kairan Zhao, Meghdad Kurmanji, George-Octavian Bărbulescu, Eleni Triantafillou, and Peter Triantafillou. What makes unlearning hard and what to do about it. In *NeurIPS*, 2024.

Part I
Appendix

A Supplement Roadmap and Related Work

The main paper presents the central argument. This supplement separates the direct confirmatory result from mechanistic diagnostics, restricted-equivalence protocol development, secondary non-authorizing evaluations, proofs, and reproducibility details. Table A.1 maps each central claim to its evidence role, while Table A.2 records the scale of the complementary evaluation components.

Table A.1: Claim–evidence crosswalk mapping central claims to evidentiary roles and supplement sections. Bold in Status highlights confirmatory status or verified coverage counts across evaluated cells.

Main claim	Main evidence	Supplement	Status
Natural non-members trigger false authorization	Fig. 2	App. C	**Confirmatory**; fixed recipe
Reference records remove fixed confounds	Fig. 3	App. D	Mechanistic diagnostic
Adaptive actions transfer to secret references	Fig. 3	App. D	**10/10** cells authorize
The finite score map is non-injective	Sec. 3	App. D	**120/120** collisions
Candidate rules reject retraining controls	Table 2	App. E	**2/2** families stop

A.1 Detailed Related Work

Unlearning objectives and update mechanisms. Machine unlearning methods include exact or partitioned retraining, certified removal, and approximate updates. Sharded, Isolated, Sliced, and Aggregated (SISA) training provides the canonical partitioned design (Bourtoule et al., 2021), while certified methods bound the discrepancy from retraining under assumptions that are strongest for convex or otherwise controlled learners (Guo et al., 2020; Ginart et al., 2019). Approximate methods use gradient updates, influence approximations, or representation changes to reduce deletion cost (Neel et al., 2021; Sekhari et al., 2021; Golatkar et al., 2020; Graves et al., 2021; Izzo et al., 2021). Repeated deletion and application-dependent objectives introduce additional evaluation demands (Kurmanji et al., 2023). These approaches differ in the object they guarantee, the randomness covered by the guarantee, and whether utility is part of the claim. Analyses of approximate unlearning further show that update closeness and retraining equivalence require careful definitions (Thudi et al., 2022). Broad surveys organize these algorithmic families and their evaluation assumptions (Nguyen et al., 2025; Wang et al., 2025). We study when an observed audit transcript identifies the historical or distributional claim used to interpret these operators.

Membership inference, calibration, and proof of exposure. Membership inference distinguishes training members from non-members (Shokri et al., 2017; Yeom et al., 2018). Subsequent black-box and white-box attacks broadened the access models, while systematic evaluations showed that success varies with access, overfitting, data difficulty, and the construction of the comparison cohort (Salem et al., 2019; Nasr et al., 2019; Song & Mittal, 2021; Rigaki & Garcia, 2023). Modern evaluations therefore emphasize calibrated likelihood ratios, matched examples, and operating points appropriate to the claimed threat model (Carlini et al., 2022; Watson et al., 2022). These requirements become stricter when a score is used as evidence of historical exposure instead of as an attack benchmark. The membership-proof framework shows that a score cannot establish training participation without controlling the alternative ways in which a non-member can appear member-like (Zhang et al., 2025). Privacy auditing based on controlled canaries provides another route when the experimental intervention is available (Steinke et al., 2023). Our exposure diagnostics specify

Table A.2: Scope of the evaluation program across the complementary analyses. Scale distinguishes trained models, trajectory constructions, and resampling draws. Bold in Outcome highlights decisive actions, complete collision rates, zero reference absence failures, and stopped calibration families.

#	Scientific question	Scale	Inference / audit	Outcome
1	Natural non-members cross a member-like cutpoint	8 calibration and 10 confirmatory worlds; 640 confirmatory trajectories	Trajectory indicators and world-level study rule	**Conditional false authorization**
2	When AUC identifies exposure	2 environments; 5 seeds/environment; 5 prevalence levels; 500 Monte Carlo draws	Monte Carlo and seed-cluster bootstrap	Closed-world bridge; open-set sensitivity
3	Whether a finite score vector is injective	2 environments; 10 cells; 120 trajectory optimizations	Constrained action optimization and lineage gates	**120/120** approximate in-box collisions
4	Whether public-model spoofing transfers	2 environments; 10 cells; 8 secret checkpoints	Paired bootstrap, permutation, and Holm correction	Transfer authorizes **10/10** secret-reference cells
5	Whether reference selection is implementable	24 environment-selector-family summaries; 210 manifest-aware reference checks	Hash deduplication and training-fingerprint audit	Manifest-aware selection: **0** absence failures
6	Whether restricted rules authorize known positives	3 model clusters/role; 2 rule families; 2,000 development draws/family	Development-only control simulation	**2/2** candidate families stop
7	Whether the three-action calculus is constructively attainable	2 fixed candidates; 6,150 bounded coordinate probes	Directly observed singleton identified set	**Downdate authorizes; no-op rejects**

the non-exposure alternatives and target-generator access, then test the effects of natural generation, reference provenance, and adaptation.

Offline reinforcement learning and trajectory-level auditing. Offline RL learns from a fixed dataset and includes behavior-constrained value methods, sequence models, conservative objectives, and implicit value learning (Fujimoto et al., 2019; Levine et al., 2020; Chen et al., 2021; Janner et al., 2021; Kumar et al., 2020; Kostrikov et al., 2022). D4RL supplies common datasets and normalized utility measures for comparing these learners (Fu et al., 2020). At the data level, ORL-AUDITOR studies whether trajectories can be detected from trained policies, whereas TrajDeleter couples trajectory auditing with a deletion procedure (Du et al., 2024; Gong et al., 2025). Reinforcement unlearning broadens the problem beyond one offline trajectory-deletion construction (Ye et al., 2025). These works motivate trajectory-level evaluation. We ask which exposure or equivalence claims the resulting scores can support. Our controlled benchmark uses feasible retraining, simulator provenance, and utility measurement to test this distinction.

Evaluation beyond offline RL. The same score-to-claim gap appears in settings where retain-only retraining is expensive or unavailable. Work on large-model unlearning distinguishes behavioral suppression from stronger removal objectives and documents how evaluation can overstate deletion when a model conceals rather than loses information (Liu et al., 2025; Zhao et al., 2024). Recent proposals accordingly explore statistical membership audits, representation-level verification, subset-independence tests, and uncertainty-aware detection of fake forgetting (Sun et al., 2026; Cosma & Finke, 2026; Zhang et al., 2026; Shi et al., 2026). These approaches expand the observable transcript, but their conclusions remain tied to their declared nulls, interventions, and audit classes. Our distinction between exposure and retraining equivalence helps state those boundaries without assuming that one score family transfers unchanged across modalities.

B SHARED EVALUATION INSTRUMENT

B.1 SCORE AND COMPARISON COHORTS

Let $\mathcal{D}$ be the full offline dataset, let $\mathcal{D}_{\mathrm{f}} \subset \mathcal{D}$ be the forget set with $|\mathcal{D}_{\mathrm{f}}|/|\mathcal{D}| = 0.10$, and let $\mathcal{D}_{\mathrm{r}} = \mathcal{D} \setminus \mathcal{D}_{\mathrm{f}}$. Individual policies are models m drawn from the corresponding distributions defined in Section 2.

We use the action error $e(m, \tau)$ defined in Equation 2. Lower values indicate closer reproduction of the recorded actions. For two score cohorts, signed AUC uses the membership orientation fixed in advance. We also report descriptive folded separability

$$\mathrm{sep} = \max(\mathrm{AUC}, 1 - \mathrm{AUC}), \tag{9}$$

but do not use folding for directional confidence intervals or authorization.

Forget trajectories are compared with held-out trajectories matched on z-normalized episodic return, episode length, and initial-state features. The basic matching features are $z(\tau) = (z_R, z_T, z_{s_0,1}, \ldots, z_{s_0,d_s})$, using the preprocessing means and standard deviations. Pair costs are $\|z(\tau) - z(\tau')\|_2$, with equal weight per standardized coordinate. The matched-cohort procedure solves a minimum-cost one-to-one assignment within a distance threshold, with unmatched dummy assignments penalized by $h + \max(1, 0.1h)$ for threshold h. A second assignment can match remaining trajectories without the threshold. The HalfCheetah secondary cohort uses $h = 0.88275$ and matches all 14 forget trajectories to distinct members of a 31-trajectory held-out pool without this fallback. This cohort size reflects the forget split, not rejection of additional forget trajectories by matching. Known-membership controls use the same score and matching structure as the target audit. A positive control establishes power for its declared member/non-member alternative, while a same-stratum negative control detects separation introduced by the evaluation construction.

For a deployed model m_0, target trajectory τ, and selected reference set $\mathcal{R}$, let $e_j(\tau) = e(m_j, \tau)$ for the deployed index $j = 0$ and every reference index $j \in \mathcal{R}$. Reference-calibrated analyses replace the deployed error with

$$z_{m_0,\mathcal{R}}(\tau) = \frac{\frac{1}{|\mathcal{R}|}\sum_{r\in\mathcal{R}} e_r(\tau) - e_0(\tau)}{\max\{\mathrm{sd}_{r\in\mathcal{R}}\, e_r(\tau), 10^{-8}\}}, \tag{10}$$

where the standard deviation uses the $|\mathcal{R}| - 1$ denominator. Larger values are more member-like because they indicate lower error for the deployed policy than for its references. Each target seed uses the selector-specific target-absent reference set, and signed AUC is computed from the paired z-scores with the same bootstrap, label-swap, Holm, and three-action rule as the deployed analysis.

B.2 UNCERTAINTY AND THREE-ACTION DECISIONS

Signed-AUC intervals use paired percentile bootstrap with 10,000 resamples. Paired permutation tests swap target and negative labels within each pair. The stress-test analyses apply Holm correction over the declared family of cells. Multi-seed analyses resample seed as the highest independent unit and then resample trajectories within seed when required. Evaluations that add base models without a paired retraining control are reported separately from fully paired base, retraining-reference, and control evaluations.

The ordinary evaluations use the utility floor $\mathrm{floor} = \max(20, 0.4 \times \mathrm{D4RL}_{\mathrm{ref}})$, where $\mathrm{D4RL}_{\mathrm{ref}}$ denotes the reference policy normalized score on D4RL. A wide interval remains an underpowered boundary. A failed positive control makes the score diagnostically uninterpretable (audit-unknown) for the evaluated model and cohort. Passing a control establishes score behavior on the tested alternative, not completeness of the non-exposure null.

Table B.1 reports secondary settings with different numbers of model runs. These rows describe the evaluated cases rather than a balanced cross-environment comparison. In particular, the single Hopper run does not establish a population-level failure of its audit.

Table B.1: Secondary score and utility diagnostics for evaluated policy settings. Signed-AUC intervals use seed as the highest sampling unit. PC denotes known-membership positive control; a dash indicates uninterpreted scores due to sub-floor utility. Bold highlights passing validation evidence (PC > 0.85 or exact residual); underline indicates sub-floor utility or failed control gates that halt authorization.

Case	Sample unit	Utility	Target signed AUC	Validation evidence	Supported conclusion
Walker2d IQL base	3 seeds, 75 pairs each	$66.3 > 31.85$ floor	0.519 [0.461, 0.578]	PC mean **0.924**, minimum lower bound **0.865**	Validated score, but the complete chain does not establish liveness.
HalfCheetah IQL base	4 base seeds, 14 pairs each	$39.0 > 20$ floor	0.588 [0.469, 0.704]	PC lower bound **0.909** on one seed	Base-only runs lack paired retraining controls, so liveness is unresolved.
Hopper IQL base	1 seed, matched cohort	$44.7 > 20$ floor	0.528 [0.460, 0.597]	PC <u>0.515</u>, lower bound <u>0.442</u>	Control fails in this run, so target score remains uninterpreted.
Walker2d medium-replay fixed-feature ridge	1 restricted learner	<u>6.80</u> < 31.85 floor	–	Downdate residual 1.875×10^{-11}	Exact restricted update (utility-vacuous).

C Confirmatory Exposure Evidence

C.1 Dynamics-Valid Natural Confirmatory Test

The dynamics-valid test fixes Walker2d-medium-replay-v2, one public pre-trained TD3+BC policy as the trajectory generator, eight member-excluded reference worlds, eight calibration target worlds, and ten confirmatory target worlds. Each target world contains 32 simulator-generated known-member trajectories and 32 simulator-generated natural trajectories. Every trajectory stores the complete sequence produced by environment reset and step calls. The natural cohorts remain absent from the corresponding executed training multisets, while only the member cohorts enter their target-world training data.

Evidence is the mean reference action error minus the target-model action error, without standard-deviation normalization. Larger values are more member-like. The eight calibration member cohorts determine one pooled fifth-percentile evidence cutpoint. Calibration natural scores are not computed. The ten confirmatory worlds are scored once against the fixed reference bank. A trajectory receives a member-like indicator when its evidence exceeds the cutpoint. The study-level action additionally requires the member-control gate, at least nine of ten natural world rates above 0.20, the fixed sign-test gate, and the interval-width gate.

Table C.1: Dynamics-valid evidence under the fixed Walker2d TD3+BC recipe. Member-like indicates trajectory evidence exceeding the calibration-member cutpoint. Bold highlights severe false authorization rates on unexposed natural non-members; underline denotes the known-member calibration baseline.

Cohort	Worlds	Trajectories	Member-like	World support
Known members	10	320	<u>316/320 (0.988)</u>	<u>10/10 above 0.80</u>
Dynamics-valid natural non-members			**306/320 (0.956)**	**10/10 above 0.20**

Table C.1 separates three units that should not be conflated. The 320 trajectory indicators define the finite member-like fraction, the ten world rates define the support count, and one study-level action states whether the pre-specified conditional false-authorization endpoint is met. The observed interval is conditional on the fixed reference bank and is not a population guarantee over alternative reference recipes.

Table 1 in the main text summarizes the complete-score analysis.

The descriptive AUC analysis compares all member and natural scores with ties split equally. Pooled AUC uses all cross-class pairs, including pairs from different target worlds. The mean within-world

Table D.1: Exposure-authorization stress test across five seeds per environment. A/Ab/R denotes authorize/abstain/reject counts followed by mean signed AUC. Bold highlights desirable audit outcomes (canary detection or suppression of false authorization to 0/5/0) and successful adaptive spoof transfer; underline indicates severe false authorizations on unexposed confounds under deployed scores.

Target	Exposed	Audit variant	Walker2d	HalfCheetah
Injected signal	yes	Deployed score	4/1/0 (0.942)	5/0/0 (0.968)
		Manifest-aware	**4/1/0 (0.999)**	**5/0/0 (1.000)**
Raw cross-model confound	no	Deployed score	5/0/0 (0.901)	5/0/0 (0.887)
		Manifest-aware	**0/5/0 (0.429)**	**0/5/0 (0.643)**
Matched cross-model confound	no	Deployed score	5/0/0 (0.733)	5/0/0 (0.853)
		Manifest-aware	**0/5/0 (0.460)**	**0/5/0 (0.625)**
Model-adaptive spoof	no	Public references	5/0/0 (0.943)	5/0/0 (0.842)
		Secret references	**5/0/0 (0.975)**	**5/0/0 (0.960)**

Table D.2: Complete Walker2d pre-specified three-action decisions across prospective seeds. Bold marks correct actions (authorization for exposed canaries and abstention for unexposed retain-only controls); underline highlights false authorizations on unexposed cross-model confounds.

Seed	Injected (Exposed)	Retain-only (Non-exp.)	Raw cross (Non-exp.)	Matched cross (Non-exp.)
10	0.986 / abstain	0.549 / **abstain**	0.840 / authorize	0.736 / authorize
11	0.979 / **authorize**	0.611 / **abstain**	0.917 / authorize	0.757 / authorize
12	0.896 / **authorize**	0.444 / **abstain**	0.924 / authorize	0.729 / authorize
13	0.951 / **authorize**	0.556 / **abstain**	0.993 / authorize	0.757 / authorize
14	0.896 / **authorize**	0.562 / **abstain**	0.833 / authorize	0.688 / authorize

AUC assigns equal weight to the ten worlds. Neither analysis changes the threshold or introduces an independence claim for pooled trajectories.

D Mechanistic and Adversarial Diagnostics

D.1 Construction, Blinding, and the Fixed Rule

Each environment contains five seeds from each of four construction families: injected signal, retain-only, raw cross-model, and action-marginal-matched cross-model. Random seeds are fixed in advance across two disjoint groups (seeds 10–14 for this suite and seeds 30–33 for the secret references described below). Each cell contains twelve paired targets and controls. Models are trained for 100,000 gradient steps and evaluated for 25 deterministic episodes. Injected canaries are repeated 16 times. The two cross-model variants within a seed share a student model and held-out state-reward cohort, so seed is the highest resampling unit.

The uncalibrated scorer receives utility, signed AUC, its interval, a paired permutation statistic, and non-label metadata. It authorizes when utility is at least 20, the AUC lower bound exceeds 0.5, and the adjusted p-value is below 0.05. It rejects when utility is adequate and the AUC upper bound is at most 0.5, and abstains otherwise. The calibrated analyses use the statistic in Equation 10 and retain the same decision form.

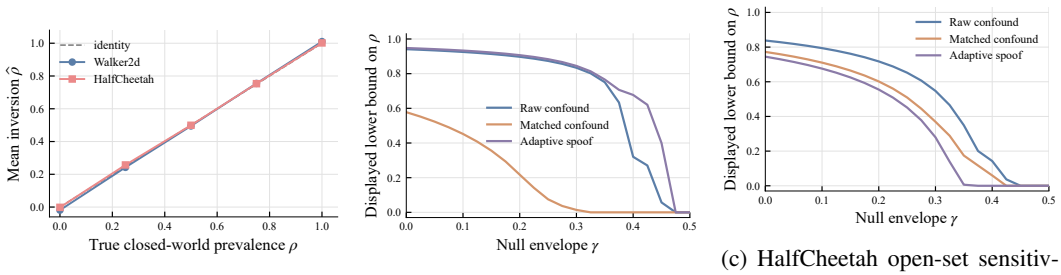
Table D.2 gives the per-seed Walker2d decisions, and Table D.3 compares the two environments. The latter separates directional score significance from training-membership verification.

D.2 Closed-World AUC Bridge and Open-Set Misspecification

The synthetic bridge resamples empirical known-member and matched-nonmember score cohorts from seeds 10–14. For each environment and prevalence $\rho \in \{0, 0.25, 0.5, 0.75, 1\}$, it performs 500 Monte Carlo draws and 500 seed-cluster bootstrap draws. The raw estimator has small bias, but the

Table D.3: Cross-environment reproduction under lineage-verified construction families. Bold in Verified confirms target absence from training multisets; underline indicates complete (5/5) false significance on unexposed cross-model confounds.

	Raw cross-model confound		Matched cross-model confound		Verified
Environment	**AUC [95% CI]**	**Significant**	**AUC [95% CI]**	**Significant**	
Walker2d	0.901 [0.853, 0.950]	5/5	0.733 [0.710, 0.753]	5/5	**5/5**
HalfCheetah	0.887 [0.864, 0.914]	5/5	0.853 [0.815, 0.890]	5/5	**5/5**



(a) Closed-world identification (b) Walker2d open-set sensitivity (c) HalfCheetah open-set sensitivity

Figure D.1: Closed-world identification and open-set sensitivity of the AUC bridge. The synthetic mixtures recover the injected prevalence on average, whereas applying the same inversion to known non-exposure constructions yields large apparent prevalence. The sensitivity panels show how an allowed null deviation γ weakens the lower bound without estimating γ from the target cohort.

naive five-seed interval undercovers. The open-set rows apply the closed-world inversion to known non-exposure constructions. Their large apparent prevalence values measure null misspecification rather than exposure (Table D.4). Figure D.1 summarizes both directions: Figure D.1a shows recovery in the synthetic mixtures, while Figures D.1b and D.1c show how the assumed null deviation changes the open-set bound.

For the sensitivity analysis, γ ranges from zero to 0.5 in increments of 0.025. A row is marked incompatible when the algebraic lower bound exceeds one and unresolved when the positive-control denominator is too small after the null envelope is applied. No value of γ is estimated from the target cohort.

D.3 Reference-Selector Audit

Across the evaluated roles in each environment, the retain-only retraining reference and the deployed student policy share the same trained checkpoint within each seed, yielding ten distinct reference models. Deduplicating identical checkpoints prevents repeated weighting. The same-family selector reads the target construction family. The role-only selector reads environment, architecture, training role, and seed from benchmark summary metadata. The manifest-aware selector additionally reads training membership records and retains only target-absent checkpoints.

Table D.5 reports that the role-only and same-family selectors choose identical checkpoints in every cell. Their 30 absence failures all occur for injected targets. The manifest-aware selector passes all 210 reference-target absence checks and preserves every three-action decision relative to the same-family calibrated selector. Relative to the deployed score, it changes the fixed cross-model targets from authorization to abstention. It therefore repairs the reference semantics rather than acting as a score-only selector.

D.4 Public-to-Secret Adaptive Transfer

The adaptive generator observes the deployed checkpoint and four public same-recipe references. It changes actions while fixing observations, rewards, and lengths. Four secret references per environment, indexed by seeds 30–33 and giving eight checkpoints in total across the two environments, are trained only after the forged actions are fixed. Final inference uses the deployed model and all four

Table D.4: Finite-sample behavior of the closed-world AUC mixture identity and its open-set failure mode. Bold in closed-world recovery highlights small synthetic bias; underline in open-set columns indicates severe false identification of non-exposed cohorts as high-prevalence exposure ($\hat{\rho}$).

	Closed-world synthetic recovery		Open-set apparent exposure ($\hat{\rho}$)		
Environment	**Bias range**	**CI coverage range**	**Raw cross**	**Matched cross**	**Adaptive spoof**
Walker2d	**[-0.019, +0.010]**	0.812–0.848	<u>0.945</u>	<u>0.578</u>	<u>0.969</u>
HalfCheetah	**[-0.002, +0.007]**	0.828–0.856	<u>0.838</u>	<u>0.772</u>	<u>0.745</u>

Table D.5: Reference-selector audit across both environments. Checks counts verified reference models; Absence failures indicates reference sets containing target data. Bold denotes zero absence failures (full isolation); underline indicates reference contamination.

Selector	**Visible information**	**Checks**	**Absence failures**	**Parity vs. same family**
Same family	Construction family	160	<u>30</u>	–
Role only	Training role metadata	160	<u>30</u>	**yes**
Manifest aware	Executed-training fingerprints	210	**0**	**yes**

secret references, with Holm correction over ten environment-by-seed cells. Training membership records verify absence of the original and forged cohorts from every secret training multiset.

Figure D.2 summarizes the reference interventions: Figure D.2a shows reference-absence checks, Figure D.2b compares public and secret AUC, and Figure D.2c reports leave-one-reference sensitivity. The leave-one-reference analysis repeats the complete ten-cell inference four times, once for each possible omission. It is a robustness diagnostic and does not define the primary transfer result.

D.5 Complete Five-Checkpoint Score-Vector Collision

For each cell, the target vector contains the five action-error values of an exposed canary. The optimization begins from a distinct matched nonmember observation, reward, and length skeleton and changes only its actions. Adam runs for 800 steps with learning rate 0.03, action regularization 10^{-4}, and box $[-1, 1]$. Success requires maximum absolute vector error at most 10^{-3}, maximum action change above 10^{-6}, altered complete trajectory sequences, preserved state-reward sequences, and absence from all five public training multisets. Figure 4 in Section 3 displays the maximum vector error, maximum action shift, and local free dimension distributions across the 120 constructed collisions.

Table D.7 reports the approximation error and action changes. All score Jacobians have numerical rank five at relative threshold 10^{-5}. The reported local free dimension is the flattened action dimension minus this rank. The resulting decision statistics are descriptive because the exposed vectors are not strong positive controls under the raw-family models.

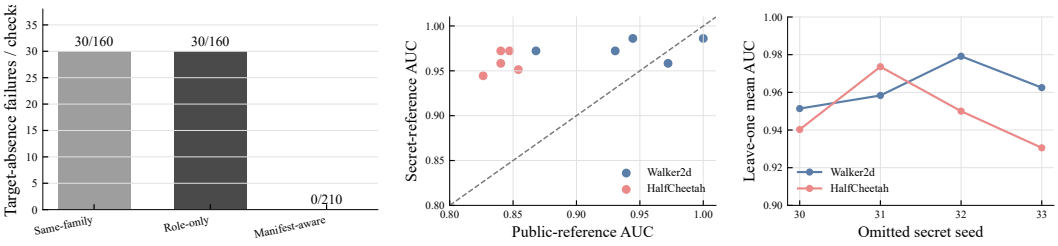
E Restricted-Equivalence Protocol Development

E.1 Development-Only Control Designs

Both calibration designs use a Walker2d cohort of 75 forget and 75 matched held-out trajectories. Three model sources per role are resampled with replacement to fill eight reference slots and six evaluation slots. Trajectory positions are also resampled and shared across the model comparisons within a draw. These operations retain dependence from repeated model sources. The tagged-KS design uses the 95th reference quantile as a random candidate tolerance and a t-style interval for a mean of target–reference empirical distances. The split-conformal design uses the same role geometry and evaluates the eight-pair grid $q_{\mathrm{acc}} \in \{0.5, 0.625, 0.75, 0.875\}$ and $q_{\mathrm{rej}} \in \{0.875, 1.0\}$. Each design uses 2,000 resamples. The development criteria require authorization of at least five of six retraining-control slots with probability at least 0.8 and positive-control rejection frequency at most 0.2. Analogous gates require frequent rejection and infrequent authorization of full-data

Table D.6: Public-to-secret reference transfer of the model-adaptive action spoof. Bold denotes complete transfer authorization (5/0/0) and invariant decision stability across all leave-one-reference omissions.

	Action decision (A/Ab/R)		Signed AUC		5/5 stable
Environment	Public	Secret	Secret	Leave-one range	
Walker2d	5/0/0	**5/0/0**	**0.975**	0.951–0.979	**yes**
HalfCheetah	5/0/0	**5/0/0**	**0.960**	0.931–0.974	**yes**



(a) Selector audit (b) Public vs. secret AUC (c) Leave-one-reference

Figure D.2: Reference evidence atlas. The selector audit counts failures of the requirement that a reference training multiset exclude the target under both complete and state-reward views. The paired scatter compares public- and secret-reference AUC in all ten adaptive cells, and the leave-one-reference panel reports environment-level means after omitting each secret seed. Manifest-aware selection satisfies this declared target-absence requirement, while adaptive transfer persists under secret references and every omission diagnostic.

comparators. Neither design evaluates a deleted model. Full-data status supplies a comparator role, not independent proof of non-equivalence. Table 2 reports the control results.

The tagged-KS candidate fails both the positive correct-count power and positive false-rejection gates. Every split-conformal candidate also fails those gates. Across the eight grid points, positive correct-count power is 0.006, 0.006, 0.017, 0.017, 0.095, 0.095, 0.289, and 0.289, in the lexicographic order of the displayed grid. Table 2 reports the grid maximum. Full-data comparators are rejected throughout. The rank grid has no selected candidate that passes all criteria. The reported maximum is descriptive, not a validated selection. Because the data contain only three model sources per role, these rates characterize the candidate designs and do not establish population coverage. In addition, the mean pairwise statistic differs from the model-distribution distance, as shown in Equation 5. The simulations therefore do not instantiate the coverage premise of Theorem 2.

E.2 Exact-Observation Control

The constructive control uses Walker2d medium-replay data with 756 training trajectories partitioned into 75 forget and 681 retain trajectories. A public-state random Fourier feature map has 1,024 random features and a bias coordinate, and ridge regression uses $\lambda = 10^{-2}$. The two candidate worlds share the data, feature map, canonical parameterization, and independently rescanned retain-only reference. Let $\theta(M) \in \mathbb{R}^{6150}$ be the fixed vectorization of the parameter matrix of model M, let θ_g denote the reference parameters, and set $S = \max\{1, \|\theta_g\|_\infty\}$. Define $\Delta_j(M) = \theta_j(M) - \theta_{g,j}$. For each coordinate j, the audit class contains

$$f_j(M) = \mathrm{clip}\left(\frac{1}{2} + \frac{\Delta_j(M)}{2S}, 0, 1\right). \tag{11}$$

We set $\mathcal{F} = \{f_j : 1 \le j \le 6150\}$. The reference-dependent centering and scale are fixed for both candidates. For their deterministic model distributions, $d_{\mathcal{F}}$ is the maximum coordinate-probe difference from the reference. Let $A_R = \Phi_R^\top \Phi_R + \lambda I$ be the regularized retain system, $\kappa = \kappa_2(A_R)$, and u be 64-bit machine precision. We fix $\tau = \max\{10^{-8}, 100\kappa u\}$ as the tolerance for the scaled parameter difference $\|\theta(M) - \theta_g\|_\infty / S$. The probe-distance tolerance is $\eta = \tau/2 = 6.35 \times 10^{-8}$ because each unclipped probe scales a coordinate displacement by $1/(2S)$.

Table D.7: Non-identity collision of complete five-checkpoint action-error vectors on fixed state-reward skeletons. Bold denotes 100% collision success across cells and trajectories, and tight vector error bounds below 2×10^{-4}.

Environment	Cells	Trajectories	Max vector error	Max action shift range	Local free dimension	Forged AUC
Walker2d	**5/5**	**60/60**	$\mathbf{1.56 \times 10^{-4}}$	0.150–1.747	73–5995	0.613
HalfCheetah	**5/5**	**60/60**	$\mathbf{1.68 \times 10^{-4}}$	0.192–1.488	5995–5995	0.433

Table F.1: Pairwise score diagnostics for Walker2d TD3+BC across three seeds. Gold repeatability tolerance is 0.160. The retain-refit mean score is 76.54. Bold highlights favorable point estimates for deleted policies; underlined interval bounds highlight overlapping bootstrap distributions that preclude definitive equivalence verdicts.

Comparison	Observed mean	Bootstrap percentiles (95%)	Mean D4RL	Interpretation
Base vs. gold	0.369	[0.347, 0.449]	71.61	Diagnostic only
Deleted vs. gold	**0.276**	[0.262, 0.356]	**76.54**	

Table E.1: Exact-observation constructive control for the identified-set rule. Bold highlights decisive actions and the probe distance satisfying $d_{\mathcal{F}} \leq \eta$ under the 64-bit parameter audit class.

Candidate	Construction	$d_{\mathcal{F}}$	η	Action
Algebraic downdate	Full minus forget statistics	$\mathbf{1.02 \times 10^{-11}}$	6.35×10^{-8}	**authorize**
No-op base	Unchanged full-data solution	7.08×10^{-2}		**reject**

The algebraic path subtracts the forget sufficient statistics from the full-data statistics before solving, whereas the reference path rescans retain data. Their parameter L_∞ difference is 3.23×10^{-10}, and the corresponding probe distance is 1.02×10^{-11}, below η. The unchanged full-data solution has probe distance 7.08×10^{-2} and is rejected. The trajectory-content multiset partition is exact and disjoint, sufficient-statistic cancellation errors are below 2.8×10^{-12}, and all relative Karush–Kuhn–Tucker (KKT) residuals are below 5.5×10^{-15}. These checks make the example an end-to-end arithmetic control rather than an object-identity comparison. Because the observed 64-bit arrays define the numerical models and determine the distance exactly, the confidence set is a singleton with coverage one over this two-world class. The construction does not provide a numerical error bound for an ideal real-arithmetic solution or a coverage result for stochastic score audits. This restricted-equivalence-only control does not include a utility condition and is distinct from the utility-assessed fixed-feature configuration discussed below.

F Secondary and Non-Authorizing Evaluations

F.1 Secondary Restricted-Distance Diagnostics

The secondary diagnostics compare Walker2d TD3+BC base, retain-only retraining, and retain-refit deletion scores for three seeds. Each evaluation uses 75 forget and 75 matched-negative scores. The retain-refit policy is trained after removal of the forget split and refitting on the retained data. Pairwise gold variation supplies the repeatability tolerance η: for every bootstrap draw, the maximum tagged distance across the three gold-seed pairs is computed, and η is its 95th percentile over 5,000 paired trajectory bootstrap draws. The liveness margin is defined as $\lambda = \eta$, so base liveness requires a lower bootstrap percentile above λ and deleted authorization requires an upper bootstrap percentile at or below η. Hierarchical seed-and-trajectory bootstrap ranges compare base and deleted distances. No tolerance in the examined grid simultaneously authorizes the deleted upper bound and establishes the corresponding negative-control separation.

Table F.2 compares 14 forget and 14 matched held-out trajectories in HalfCheetah under KS, energy, and radial-basis-function maximum mean discrepancy (RBF-MMD) statistics. For score samples

Table F.2: Secondary pairwise-distance diagnostics for HalfCheetah TD3+BC under three score distances. Bold highlights the smaller observed distance between base and deleted policies in each audit class; all rows remain descriptive without population equivalence guarantees.

		Base policy		Deleted policy (retain-refit)		Interpretation
Audit class	η	Observed	Seed range	Observed	Seed range	
Tagged KS	0.357	**0.238**	[0.214, 0.286]	0.262	[0.214, 0.286]	Diagnostic only
Tagged energy	1.59×10^{-3}	2.08×10^{-3}	$[1.71 \times 10^{-3}, 2.38 \times 10^{-3}]$	$\mathbf{1.69 \times 10^{-3}}$	$[1.39 \times 10^{-3}, 1.91 \times 10^{-3}]$	
Tagged RBF-MMD	2.26×10^{-3}	6.30×10^{-3}	$[5.90 \times 10^{-3}, 6.79 \times 10^{-3}]$	$\mathbf{5.48 \times 10^{-3}}$	$[4.74 \times 10^{-3}, 6.16 \times 10^{-3}]$	

$x_1, \ldots, x_n$ and $y_1, \ldots, y_m$, the energy statistic is

$$E = \frac{2}{nm} \sum_{i,j} |x_i - y_j| - \frac{1}{n^2} \sum_{i,i'} |x_i - x_{i'}| - \frac{1}{m^2} \sum_{j,j'} |y_j - y_{j'}|. \tag{12}$$

The squared MMD uses the V-statistic

$$\widehat{\mathrm{MMD}}^2 = \frac{1}{n^2} \sum_{i,i'} k(x_i, x_{i'}) + \frac{1}{m^2} \sum_{j,j'} k(y_j, y_{j'}) - \frac{2}{nm} \sum_{i,j} k(x_i, y_j), \tag{13}$$

including diagonal terms, with $k(x, y) = \exp(-(x - y)^2/(2\sigma^2))$. The bandwidth $\sigma = 0.04441168$ is the median nonzero absolute pairwise difference among gold scores pooled across both cohorts and all three seeds, fixed throughout resampling. Both statistics are clipped at zero for numerical stability and are reported without taking square roots. For each statistic, tagged means the maximum of its separate forget- and held-out-cohort values.

The observed estimate averages the three seed-specific tagged distances, and the reported range spans the three per-seed values. With three seeds and fourteen pairs per cohort, resampling-based intervals would not be meaningful for these nonlinear statistics, so the table reports the per-seed spread as a descriptive range without a coverage claim. The cohort contains all 14 forget trajectories matched under the caliper in Appendix B. Its size reflects the 10% forget split rather than case selection. The small cohort makes these rows sensitivity diagnostics rather than a replacement for the retraining-control tests.

As Table F.1 shows, the mean pairwise tagged distance is 0.369 for base and 0.276 for retain-refit, with descriptive bootstrap ranges $[0.347, 0.449]$ and $[0.262, 0.356]$, respectively. The retain-refit upper endpoint exceeds the base lower endpoint, so no common threshold separates these ranges in the intended directions. For the Walker2d comparison, model and trajectory pairs are drawn by resampling rather than by predetermined trajectory identities, and the utility threshold is descriptive rather than independently fixed. These limitations and the distinction between pairwise and model-distribution distances preclude a formal equivalence or non-equivalence verdict.

Underpowered base signals and inconclusive diagnostic cells. Walker2d and HalfCheetah IQL base-policy estimates remain too uncertain to establish a live base signal with paired retrain-gold evidence. A high-observed-utility Walker2d medium-expert additive downdate and a Hopper IQL cell have favorable membership readings, but their positive controls do not become confirmatory. For the Hopper cell, the first seed already fails the positive-control gate, so the pre-specified protocol stops evaluation there instead of interpreting the target score or spending additional runs on a failed control. These cells remain diagnostically uninterpretable (audit-unknown) rather than supplying evidence of deletion.

Exact but utility-vacuous deletion. A separate Walker2d fixed-feature ridge configuration matches the algebraic deletion update to numerical precision but obtains a D4RL score of 6.8, below its 31.85 utility floor. Exactness for this restricted learner therefore does not authorize useful unlearning. The example separates computational deletion correctness from the behavioral utility condition.

AntMaze target alignment and dependence. The operator chain evaluates a 120-trajectory forget set, whereas the available positive-control construction yields 138 trajectories. The five shadows use random 80% subsets with subset-specific normalization. Their signed AUC values range from

0.585 to 0.688, while same-stratum negative-control values range from 0.466 to 0.550. The training subsets overlap by 0.797 on average, and scores on shared trajectories have mean Pearson and Spearman correlations of 0.920 and 0.928. These intervals describe variation within one finite source pool rather than independent environment replication.

Third-algorithm reproduction boundary. An isolated Conservative Q-Learning (CQL) implementation (Kumar et al., 2020) exchanges data and policies through fixed arrays and TorchScript serialization. Three 500,000-step runs obtain 100-episode D4RL scores of 82.043, 80.441, and 66.352 with exact action parity. Because the learner has no known-membership control, retain-only reference, or deletion operator, it provides base-policy reproduction but no unlearning authorization.

G Proofs of the Claim-Specific Results

For probability measures P, Q on a common measurable space, we use

$$\mathrm{TV}(P,Q) = \sup_B |P(B) - Q(B)|. \tag{14}$$

Every measurable h with values in $[0,1]$ then satisfies

$$\left| \int h \, d(P-Q) \right| \le \mathrm{TV}(P,Q). \tag{15}$$

G.1 Additional Statements

For nonempty claim classes, define

$$\varepsilon_\star(C; \mathcal{W}, \Pi) = \inf_{C(W_0)=0,\, C(W_1)=1} \mathrm{TV}(P^\Pi_{W_0}, P^\Pi_{W_1}). \tag{16}$$

Proposition 2 (Generator-class monotonicity)**.** *For $\mathcal{W}' \subseteq \mathcal{W}$ with both claim values present in each class, $\varepsilon_\star(C; \mathcal{W}', \Pi) \ge \varepsilon_\star(C; \mathcal{W}, \Pi)$.*

Proposition 3 (Audit contraction)**.** *For a common Markov kernel K generating observations from models, $\mathrm{TV}(Q_{\mathrm{del}}\mathsf{K}, Q_{\mathrm{gold}}\mathsf{K}) \le \mathrm{TV}(Q_{\mathrm{del}}, Q_{\mathrm{gold}})$.*

The common-kernel requirement includes the conditional mechanism generating references and metadata. Changing that mechanism changes the experiment.

Proposition 4 (Closed-world AUC identity)**.** *Let $\theta(P,Q)$ denote tie-aware AUC with a fixed member-like orientation. If $P_T = \rho P_M + (1-\rho)P_N$, then*

$$\theta(P_T, P_N) = \tfrac{1}{2} + \rho\big(\theta(P_M, P_N) - \tfrac{1}{2}\big). \tag{17}$$

When $\theta(P_M, P_N) \ne 1/2$, this identifies the mixture prevalence ρ.

If instead $P_T = \rho P_M + (1-\rho)Q$ and $\mathrm{TV}(Q, P_N) \le \gamma$, then for $\theta(P_M, P_N) > 1/2 + \gamma$,

$$\rho \ge \left[\frac{\theta(P_T, P_N) - (1/2+\gamma)}{\theta(P_M, P_N) - (1/2+\gamma)} \right]_+ . \tag{18}$$

Here γ is a sensitivity parameter, not an estimated TV. A right-hand side greater than one indicates incompatibility with the stipulated mixture. The result concerns cohort prevalence rather than individual exposure.

G.2 A Confidence Interval for Fixed Finite Tests

For K fixed $[0,1]$-valued model tests and independent samples of n_{del} and n_{gold} models, let $\widehat{d}$ be the largest absolute difference between test sample means. Hoeffding bounds and a union bound give radius

$$r_\alpha = \sqrt{\frac{\log(4K/\alpha)}{2n_{\mathrm{del}}}} + \sqrt{\frac{\log(4K/\alpha)}{2n_{\mathrm{gold}}}}. \tag{19}$$

Then $[\max(0, \widehat{d} - r_\alpha), \min(1, \widehat{d} + r_\alpha)]$ covers $d_{\mathcal{F}}$ with probability at least $1 - \alpha$. Each of the $2K$ sample means deviates from its expectation by its corresponding radius with probability at most $\alpha/(2K)$. On the simultaneous event, the triangle inequality bounds every difference, and taking maxima preserves the bound. Tests must be fixed independently of these samples, and their values must be observed or supplied with separate sampling-error bounds. Repeated resampling of a finite model bank does not increase the number of independent models.

G.3 Proof of Theorem 1

For a randomized rule, define $a(z) = \delta(\mathsf{A} \mid z)$ and $r(z) = \delta(\mathsf{R} \mid z)$. Both functions are measurable and take values in $[0, 1]$. Equation 15 gives

$$P^{\Pi}_{W_1}(\delta = \mathsf{A}) = \int a \, dP^{\Pi}_{W_1} \le \int a \, dP^{\Pi}_{W_0} + d_{01}, \tag{20}$$

$$P^{\Pi}_{W_0}(\delta = \mathsf{R}) = \int r \, dP^{\Pi}_{W_0} \le \int r \, dP^{\Pi}_{W_1} + d_{01}. \tag{21}$$

These are the two pairwise action bounds. If false authorization is at most α and false rejection is at most α_{R}, then

$$\mathrm{Cov}_{W_1}(\delta) \le \alpha + d_{01} + \alpha_{\mathrm{R}}, \tag{22}$$

$$\mathrm{Cov}_{W_0}(\delta) \le \alpha + \alpha_{\mathrm{R}} + d_{01}. \tag{23}$$

Taking the minimum with one proves the coverage statement. As an additional consequence, if false non-authorization is at most β, then

$$1 - \beta \le P^{\Pi}_{W_1}(\delta = \mathsf{A}) \le \alpha + d_{01}, \tag{24}$$

which gives $\alpha + \beta \ge 1 - d_{01}$. Finally, the definition of the separation modulus supplies, for every $\xi > 0$, a cross-claim pair with $d_{01} \le \varepsilon_\star + \xi$. □

G.4 Proof of Theorem 2

Let $E = \{\psi(W) \in \mathcal{C}_\alpha(Z)\}$. Fix a world with $C(W) = 0$. On E, compatibility gives $0 \in \Gamma(\psi(W))$. The event $\delta = \mathsf{A}$ requires a nonempty confidence set contained in D_1, which is incompatible with E. Therefore

$$\{\delta = \mathsf{A}\} \subseteq E^c, \tag{25}$$

and uniform coverage bounds false authorization by α. The same argument for $C(W) = 1$ shows that $\{\delta = \mathsf{R}\} \subseteq E^c$, bounding false rejection by α.

For a fixed world with claim value c, the event $\{\varnothing \ne \mathcal{C}_\alpha(Z) \subseteq D_c\}$ makes the rule choose the correct non-abstaining action. If this event has probability at least $1 - \kappa$, every abstention lies in its complement and has probability at most κ. □

G.5 Proof of Proposition 2

For nested world classes $\mathcal{W}' \subseteq \mathcal{W}$, the cross-claim pair set $\mathcal{W}'_0 \times \mathcal{W}'_1$ is a subset of $\mathcal{W}_0 \times \mathcal{W}_1$. Taking the infimum of the same transcript-TV function over the smaller set cannot decrease its value. □

G.6 Proof of Proposition 3 and Its Converse Boundary

For a transcript event B, define $h_B(m) = \mathsf{K}(m, B)$. Since K is a Markov kernel, $0 \le h_B \le 1$. Therefore

$$(Q_{\mathrm{del}}\mathsf{K})(B) - (Q_{\mathrm{gold}}\mathsf{K})(B) = \int h_B \, d(Q_{\mathrm{del}} - Q_{\mathrm{gold}}). \tag{26}$$

Applying Equation 15 and taking the supremum over B proves the contraction. The common-kernel assumption is essential: references and metadata must belong to the compared model state or be generated conditionally by the same kernel.

For the converse, let f be deterministic and non-injective. Choose $m_0 \ne m_1$ with $f(m_0) = f(m_1)$ and set $Q_0 = \delta_{m_0}$ and $Q_1 = \delta_{m_1}$. Their pushforwards are equal, while the event $\{m_0\}$ shows $\mathrm{TV}(Q_0, Q_1) = 1$. □

G.7 Proof of Proposition 4

For a score law P against P_N, define

$$\theta(P, P_N) = \Pr[S(X) < S(Y)] + \frac{1}{2}\Pr[S(X) = S(Y)], \tag{27}$$

where $X \sim P$ and $Y \sim P_N$ are independent. The comparison kernel is bounded and integration is linear in the first law. Hence

$$\theta(P_T, P_N) = \rho\theta(P_M, P_N) + (1-\rho)\theta(P_N, P_N) \tag{28}$$
$$= \rho\theta(P_M, P_N) + (1-\rho)/2. \tag{29}$$

Exchangeability of two draws from P_N, with ties split equally, gives the second equality. Rearrangement proves the closed-world identity.

For the open-set bound, define

$$h(x) = \Pr_{Y\sim P_N}[S(x) < S(Y)] + \frac{1}{2}\Pr_{Y\sim P_N}[S(x) = S(Y)]. \tag{30}$$

Then $0 \le h \le 1$ and

$$|\theta(Q, P_N) - 1/2| = \left|\int h\, d(Q - P_N)\right| \le \mathrm{TV}(Q, P_N) \le \gamma. \tag{31}$$

Writing $a = \theta(P_M, P_N)$ and $b = \theta(Q, P_N)$ gives

$$\theta(P_T, P_N) = \rho a + (1-\rho)b \le \rho a + (1-\rho)(1/2 + \gamma). \tag{32}$$

When $a > 1/2 + \gamma$, rearrangement yields the stated lower bound, while a nonpositive numerator gives the trivial bound $\rho \ge 0$. □

G.8 Proof of Proposition 1

For $j \ge 2$, expanding the squared distances yields

$$D\big(g_j(a) - g_1(a)\big) = \|p_j\|_2^2 - \|p_1\|_2^2 - 2\langle a, p_j - p_1\rangle. \tag{33}$$

Matching all score differences to those of $a_\star$ is therefore equivalent to

$$\langle a - a_\star, v\rangle = 0 \quad \text{for every } v \in L. \tag{34}$$

Write $a - p_1 = u + v$ with $u \in L$ and $v \in L^\perp$. The difference constraints fix $u = u_\star$. Matching g_1 further requires

$$\|v\|_2^2 = \|v_\star\|_2^2. \tag{35}$$

Thus v ranges over a sphere in $L^\perp$ with positive radius. Since $\dim(L^\perp) = D - r \ge 2$, the sphere has dimension $D - r - 1$ and contains infinitely many points. Every point matches the complete vector g. □

The proof concerns Euclidean action space at fixed observations. An action box, simulator dynamics, return, and transcript fields outside the score vector require separate conditions. Approximate equality of deterministic score vectors does not imply closeness in TV.

H Reproducibility Details

The fixed-observation TD3+BC and IQL models are trained on an NVIDIA RTX 5070 Ti. Reported scoring and analysis use a Linux workstation with an AMD Ryzen 9 7950X 16-core CPU (32 threads), 93 GB RAM, and one NVIDIA GeForce RTX 4070 Ti SUPER GPU with 16 GB memory (driver 580.65.06). The principal library versions used for the reported results are PyTorch 2.13.0 with CUDA 13.0, NumPy 2.5.1, SciPy 1.18.0, scikit-learn 1.9.0, Gymnasium 1.3.0, Hydra-core 1.3.4, OmegaConf 2.3.1, h5py 3.16.0, and Matplotlib 3.11.0.

Table H.1: Model configuration for the fixed-observation stress tests. TD3+BC is the deployed student and IQL supplies cross-model teacher actions. Merged columns indicate shared hyperparameter values across algorithms.

Setting	TD3+BC	IQL
Network	2×256 ReLU MLP	
Batch size	256	
Training steps	100,000	
Optimizer	Adam	
Discount factor	0.99	
Gradient clip	10	
Learning rates	actor/critic 3×10^{-4}	actor/critic/value 3×10^{-4}
Target update	$\tau = 0.005$, policy delay 2	not applicable
Algorithm-specific values	policy noise 0.2, clip 0.5, BC weight 2.5	expectile 0.7, temperature 3, maximum weight 100

H.1 Model and Data Configuration

Table H.1 gives the configuration shared by the TD3+BC and IQL models in the fixed-observation stress tests. Both use two hidden ReLU layers of width 256 without dropout, Adam with zero weight decay, batches of 256, and 100,000 gradient steps. States are normalized with statistics of the relevant training split. The 10% forget split is removed only for retain-only and retain-refit roles. The matched-cohort distance and assignment procedure are specified in Appendix B. The fixed-observation exposure suite uses seeds 10–14 in each environment, 12 paired targets and controls per seed, 16 canary repetitions, and 25 deterministic evaluation episodes.

H.2 Code and Reproducibility

The reproducibility materials comprise model definitions, data-processing utilities, training and evaluation scripts, and figure generators. The experiment descriptions distinguish independent trained models from resampling draws and identify which comparisons are descriptive.

I Scope and Limitations

The empirical findings concern action-error evidence in the evaluated locomotion settings. The natural-trajectory result uses one generator and reference recipe, while the adaptive and geometric studies hold observations fixed. They do not establish dynamics-valid adaptive attacks or empirical failure of stronger audits in other domains. The retraining-control study uses only three model sources per role and evaluates development rules, not population coverage. Its main implication is that these rules require further statistical justification, not that score-based equivalence testing is impossible. The exact-observation example illustrates a sufficient-information case. For stochastic deep policies with few independent training runs, finite-sample intervals for model-distribution distances remain an open problem. Until such intervals are available, abstention on heuristic score rules is the intended output of the protocol, and it is what prevents unsupported equivalence claims.

AI Use Statement

In this work, we used generative AI tools to implement methods, assist with translation, draft parts of the paper, and edit the paper to improve readability. All AI-assisted outputs were reviewed and verified by the authors: generated text was checked against the experimental results and revised where it did not match them, and generated code was tested for correctness. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

Reproducibility Statement

Appendices A–F describe score definitions, data construction, reference access, control protocols, and diagnostic analyses. Appendix G gives proofs, and Appendix H records model configurations

and computational resources. The reported evidence distinguishes confirmatory trajectory tests from development simulations and post-hoc descriptive analysis.