SAME WINNERS, DIFFERENT SUCCESS RATES: EVALUATING HOW LLM AGENTS RECOVER FROM FAILURES

Anonymous authors
Paper under double-blind review

ABSTRACT

Evaluating how LLM agents recover from mid-task failures is central to deploying reliable agentic systems. Existing checkpoint-based benchmarks measure recovery by comparing which action is selected as best across independent runs, a quantity known as set agreement. However, set agreement is a purely ordinal measure that records which action "wins" without reflecting the absolute level of performance. When all actions fail, they tie at zero reward, and independent runs produce the same tied set with high probability, creating an illusion of stability that masks near-zero recovery success. We formalize this limitation through a *set-path symmetry* result, proving that for equal-cost Bernoulli actions the success probabilities $(0.9, 0.8)$ and $(0.2, 0.1)$ yield identical best-action-set distributions at every sample size. No procedure based solely on which action wins can distinguish these two regimes. We further prove that certifying exact population ties is impossible in finite time, and that the assignment of outcomes to checkpoints carries information beyond marginal outcome distributions. The pooled success probability is the missing scalar that resolves the ordinal ambiguity. Experiments on 864 frozen RecoveryBench episodes and two planning cohorts totaling 3,456 responses confirm the theoretical predictions. Agreement and held-out quality can move in opposite directions, and permuting checkpoint-to-action bindings changes 8 to 13% of cell-level conclusions. Based on these findings, we propose reporting four diagnostic quantities (agreement, all-zero fraction, held-out success, and pooled success) that expose this failure mode with no additional data collection.

1 INTRODUCTION

As LLM-based agents are deployed in increasingly complex workflows, the ability to recover from mid-task failures has become a critical capability. When an agent fails at an intermediate step, a workflow may retry the operation, verify the current result, or replace the agent entirely. Evaluating these recovery strategies requires benchmarks that can reliably identify which action leads to successful task completion. Recent work has introduced checkpoint-based benchmarks such as RecoveryBench (Tan et al., 2025) and ToolMaze (Zhu et al., 2026) that replay failed agent states and measure recovery outcomes under controlled conditions.

A widely used stability metric in these benchmarks is *set agreement*, the frequency with which independent runs select the same best action. High agreement is typically interpreted as evidence that the evaluation is stable and the selected action is reliably superior. However, this interpretation assumes that agreement reflects action quality. We show that it does not, because set agreement conflates ordinal rankings with cardinal performance, and the same high agreement can arise from universal success or universal failure.

The core problem is that set agreement is a purely ordinal measure. It records which action "wins" but not the absolute level of performance. When all actions fail with near-certainty, they tie at zero reward, and independent runs produce the same tied set with high probability. The resulting agreement is an artifact of shared failure, not a signal of quality. We formalize this observation through a *set-path symmetry* result (Proposition 4.1). For two equal-cost Bernoulli actions, the parameters $(0.9, 0.8)$ and $(0.2, 0.1)$ produce identical distributions over the observed best-action set at every sample size. No procedure based solely on which action "wins" can distinguish these two regimes. The missing information is a single scalar, the pooled success probability, which requires

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.

observing actual outcomes. In addition, we prove that certifying exact population ties is impossible in finite time (Eq. (35)), and that even when outcomes are available, the assignment of outcomes to checkpoints carries information beyond marginal distributions. Permuting checkpoint-to-action bindings while preserving every action's outcome multiset changes 8 to 13% of cell-level conclusions.

We confirm this disconnect empirically across two regimes. On RecoveryBench, we evaluate 864 frozen episodes across three GPT-5.6 models (Luna, Terra, Sol) and three recovery actions (retry, verify, replace), with eight draws per action. Agreement reaches 85.4% at k=1, yet held-out success is only 1.4%, with the entire agreement driven by all-zero comparisons. Selecting models by agreement attains only 1.3% held-out success, compared to 6.5% when selecting by mean outcome. In two planning cohorts totaling 3,456 responses across 24 frozen checkpoints, agreement drops five-fold as the sample budget grows from 1 to 4 per action, while held-out success remains stable at approximately 76%. Decomposing agreement by outcome quality reveals that the decline is driven entirely by tie resolution. Multi-action matches account for 0.310 of 0.318 total agreement at k=1 and fall to 0.125 at k=4, while singleton agreement rises only slightly. A separate panel of Qwen3-8B across rank, path, and filter tasks confirms that the pattern generalizes beyond the GPT-5.6 family.

We additionally validate the Hoeffding certificate for statistical certification of best actions, showing that certification is practical only when margins are large (97.9% at 128 branches for a 0.40 gap) but remains below 0.1% even at 512 branches when the margin is 0.06. At the margins typically observed in agent recovery experiments (5 to 10 percentage points), certifying a unique best action would require hundreds or thousands of executions per action, far beyond current evaluation budgets.

Our contributions are as follows.

- **Identification theory.** We prove that best-action-set sequences identify the preference gap but not the absolute success probabilities (a property we call *set-path symmetry*). The pooled success probability resolves this ambiguity. We also prove that certifying exact population ties is impossible in finite time, and characterize the identified set as a two-element parameter family (Eq. (4)).
- **Empirical evidence.** We evaluate set agreement alongside held-out success on 864 frozen RecoveryBench episodes and two planning cohorts (3,456 responses across GPT-5.6 and Qwen3-8B models), showing that agreement and quality can move in opposite directions. Agreement decomposition, tolerance sensitivity, and cross-task analyses confirm that tie structure, not action quality, drives the observed agreement levels.
- **Record provenance.** We show that checkpoint-to-action bindings carry information beyond marginal outcome distributions. Permuting these bindings while preserving every action's outcome multiset changes 8 to 13% of cell-level conclusions, an error rate that exceeds typical measurement noise and is undetectable from outcome data alone.
- **Practical recommendations.** We propose reporting four diagnostic quantities (agreement, the fraction of all-zero comparisons, held-out success, and pooled success) that make the failure mode visible with no additional data collection. We provide statistical certification bounds and demonstrate their sample requirements under realistic margins.

2 Related Work

Agent recovery. ReAct (Yao et al., 2023), Reflexion (Shinn et al., 2023), and AutoGen (Wu et al., 2023) provide agent architectures for reasoning, self-improvement, and multi-agent coordination. AgentBench (Liu et al., 2024) and WorkflowLLM (Fan et al., 2025) evaluate agents across environments. Replay and recovery studies (Gonuguntla, 2026; Shah, 2026; Qi et al., 2026) examine interventions from partial executions. Execution-edit safety work studies checkpoint, fork, restore, and merge semantics (Zheng et al., 2026). RecoveryBench (Tan et al., 2025) derives model comparisons from recovery success on replayed failed states. We show that its set-agreement readout does not distinguish successful recovery from shared failure.

Identification and best-arm selection. Partial identification (Manski, 2003; Christensen et al., 2022) represents targets compatible with observations, and set-valued decisions (Fuentes-Vicente et al., 2026) retain unresolved alternatives. Fixed-confidence best-arm identification (Garivier & Kaufmann, 2016; Kaufmann & Kalyanakrishnan, 2013) returns optimal arms, while multiple-answer formulations (Degenne & Koolen, 2019) permit any correct answer. Our setting differs because the

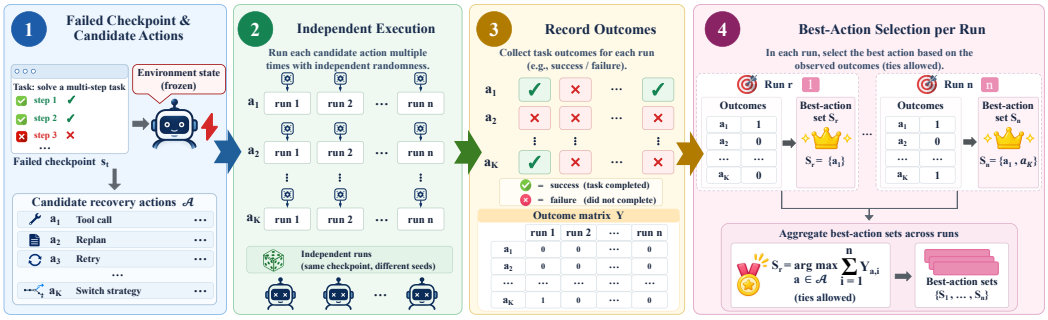


Figure 1: Overview of the checkpoint evaluation pipeline. **Step 1.** A failed checkpoint state is frozen and paired with multiple candidate recovery actions (e.g., tool call, replan, retry, switch strategy). **Step 2.** Each action is executed multiple times independently from the same checkpoint with different random seeds. **Step 3.** Binary outcomes (success or failure) are recorded for every action–run pair, forming an outcome matrix. **Step 4.** For each run, the best-action set is computed by selecting all actions with the highest total score (ties retained), yielding a collection of best-action sets that serves as the basis for agreement and quality analysis.

maximizer set's cardinality can change under arbitrarily small perturbations, yielding finite-time impossibility of certifying exact ties (Eq. (35)). Selective prediction (Geifman & El-Yaniv, 2017; Liu et al., 2026) allows abstention, and our certification rule shares this spirit. EC^2 (Golovin et al., 2010) serves as our retrieval baseline.

Repeated evaluation and provenance. Stable Ranking Probability (Riley et al., 2024) and repeated LLM evaluations (Alvarado Gonzalez et al., 2025) motivate reporting stability and quality separately. Provenance methods (Wang, 2026a; Liao, 2026; Zhu & Chang, 2026) connect actions to supporting records. We add the formal association between outcome laws and checkpoint assignments. Appendix C compares targets, observations, and assumptions across these settings.

3 Checkpoint Setting and Decision Targets

Setup and notation. We formalize the evaluation pipeline shown in Figure 1. Fix a checkpoint state s (e.g., a failed task with its current partial result), a finite set of recovery actions A_0, and a scoring contract. Each action a has cost $C(a)$ weighted by $\lambda \geq 0$. After executing action a, we observe a binary outcome. We write $\widehat{q}_s(a)$ for its sample mean minus the cost penalty $\lambda C(a)$. A reference model provides the expected score $Q_{\mathrm{sc}}(s,a)$ under the same cost adjustment. The key sets are

$$R_s = \arg\max_{a\in A_0} Q_{\mathrm{sc}}(s,a), \qquad \widehat{L}_s^{\mathrm{obs}} = \arg\max_{a\in A_0} \widehat{q}_s(a), \qquad I_s^{\mathrm{obs}} = R_s \cap \widehat{L}_s^{\mathrm{obs}}. \quad (1)$$

Here R_s is the reference-best set, $\widehat{L}_s^{\mathrm{obs}}$ is the *observed best-action set* (retaining all ties), and I_s^{obs} is their intersection. *Exact-set agreement* compares $\widehat{L}_s^{\mathrm{obs}}$ across independent execution blocks at the same checkpoint. Two blocks agree when they yield the same set.

Expected vs. observed preference. Let P be the true outcome law for binary success $\overline{Y}_{s,a} \in \{0,1\}$. The *expected* best-action set is

$$A^\star(P) = \arg\max_{a\in A_0} \{\mathbb{E}_P[\overline{Y}_{s,a}] - \lambda C(a)\}. \quad (2)$$

The observed set $\widehat{L}_s^{\mathrm{obs}}$ estimates $A^\star(P)$ but can differ due to sampling noise and ties. The central question is whether agreement on $\widehat{L}_s^{\mathrm{obs}}$, the fact that independent runs select the same set of "winners," implies anything about the *quality* of those winners, that is, whether the actions in $A^\star(P)$ actually succeed.

Why ties dominate at small budgets. Consider three recovery actions each executed once (k=1) with binary outcomes. If all three fail (a common scenario on hard checkpoints), every action scores zero and the observed best-action set is A_0 itself. A second independent run that also produces all failures yields the same set, giving perfect agreement despite zero information about relative action

quality. As k grows, empirical ties become rarer, agreement drops, but action quality is unchanged. This mechanism, driven by ties, is the main source of inflated agreement in our experiments.

Record provenance. Beyond outcomes, the assignment of outcomes to checkpoints matters. Archived results may be grouped by action without retaining which checkpoint produced which execution. We show in Section 5.3 that permuting these bindings, while preserving every action's outcome distribution, changes 8 to 13% of cell-level conclusions. Formal definitions of record eligibility and the complete audit contract are in Appendix A.

4 Why Set Agreement Alone Is Insufficient

4.1 Best-action sets do not identify success probabilities

We formalize the core problem. Knowing which action "wins" does not, in general, reveal whether recovery actually works. Consider two equal-cost recovery actions with independent Bernoulli outcomes $Y_{a,t} \sim \text{Bernoulli}(p_a)$, $p_a \in (0,1)$, across actions $a \in \{1,2\}$ and rounds t. After n rounds, we retain only the observed best-action set (as is standard in set-agreement reporting):

$$S_n = \arg\max_{a\in\{1,2\}} \sum_{t=1}^{n} Y_{a,t}, \qquad n \geq 1. \tag{3}$$

The following result shows that the entire sequence $(S_1, S_2, \ldots)$ cannot distinguish between high-quality and low-quality recovery.

Proposition (set-path symmetry). Let $P = (p_1, p_2)$ and $P^\dagger = (1-p_2, 1-p_1)$. The full labeled path $(S_n)_{n\geq 1}$ has the same distribution under P and $P^\dagger$. For an action selected based only on this path, its success probability on a fresh execution shifts by $1 - p_1 - p_2$ between the two laws.

Proof. Set $Y^\dagger_{1,t} = 1 - Y_{2,t}$ and $Y^\dagger_{2,t} = 1 - Y_{1,t}$. These are independent Bernoulli with law $P^\dagger$ and satisfy $Y^\dagger_{1,t} - Y^\dagger_{2,t} = Y_{1,t} - Y_{2,t}$. Since cumulative differences are equal at every round, the labeled paths coincide, including ties. □

Example. Take $(p_1, p_2) = (0.9, 0.8)$ and $(p_1^\dagger, p_2^\dagger) = (0.2, 0.1)$. In both cases, Action 1 is uniquely best with the same gap of 0.1. Yet $p_1 = 0.9$ in the first case and $p_1^\dagger = 0.2$ in the second. The minimax estimation error for the best success probability from the path alone is $|1 - p_1 - p_2|/2$ (Eq. (54)). Flipping all outcomes (swapping success and failure) preserves which action is better, since an action that succeeds more often also fails less often. However, this flip transforms a high-success regime into a low-success one. Any statistic that depends only on relative performance is blind to this flip.

Identified set. To characterize all parameters compatible with the path law, let $u = \Pr(S_1 = \{1\})$, $v = \Pr(S_1 = \{2\})$, $d = u - v$, and $h = \sqrt{1 - 2(u+v) + d^2}$. Solving yields exactly two compatible parameter vectors:

$$\Theta(u,v) = \left\{ \left(\frac{1 + d + \eta h}{2}, \frac{1 - d + \eta h}{2} \right) : \eta \in \{-1, 1\} \right\}. \tag{4}$$

Both share the gap $d = p_1 - p_2$ and the expected best-action set, but differ by $h = |1 - p_1 - p_2|$ in each coordinate.

Resolution. If the pooled success probability $m = (p_1 + p_2)/2$ is known, both means are identified: $p_1 = m + d/2$, $p_2 = m - d/2$. Thus the missing information is a single scalar, the overall success level, which requires observing outcomes, not just winners. Appendix G.3 treats finite-sample estimation.

4.2 Certification limits: exact ties and separated margins

Even with outcomes in hand, certifying the expected best-action set faces fundamental limits.

Exact ties cannot be certified. If two or more actions are exactly tied at the population level, an arbitrarily small perturbation changes the set. Any procedure that controls false certification at level δ

under all admissible laws can certify such a tie in finite time with probability at most δ, even with unlimited sampling (Eq. (35); proof in Appendix C.4).

Unique winners can be certified with large margins. When a unique best action exists, meaning one action's expected score strictly exceeds all others, it can in principle be certified from finite data. For n executions per action, simultaneous Hoeffding bounds (Hoeffding, 1963) yield a confidence set. Let $\widehat{q}_n(a)$ be the sample mean of action a minus its cost penalty $\lambda C(a)$. For $|A_0| \geq 2$ actions and error level $0 < \delta < 1$:

$$\rho_n(\delta) = \sqrt{\frac{\log(2|A_0|/\delta)}{2n}}, \quad W^{\mathrm{rob}}_{\delta,n} = \{a \in A_0 : \widehat{q}_n(a) > \max_{b \neq a} \widehat{q}_n(b) + 2\rho_n(\delta)\}. \tag{5}$$

With probability $\geq 1-\delta$, any action in $W^{\mathrm{rob}}_{\delta,n}$ is the unique expected best action. The rule abstains when $W^{\mathrm{rob}}_{\delta,n}$ is empty, which happens whenever margins are too small relative to n. In practice, a gap of 0.40 reaches near-complete certification at $n = 128$, but a gap of 0.06 remains unresolved at $n = 512$ (Section 5.4).

4.3 Record assignments change checkpoint conclusions

A checkpoint comparison depends not only on outcomes but on which outcomes are assigned to which checkpoints. Archived results may be grouped by action without retaining their checkpoint bindings. This creates ambiguity even when every action's outcome multiset is fully known.

How pairing matters. Suppose two outcome laws P^+, P^- have distinct best-action sets, and two execution records z^+, z^- have different eligibility (one passes provenance checks, the other does not). The same laws and records can be paired in two ways: *diagonally* (P^+ with z^+, P^- with z^-) or *crosswise* (P^+ with z^-, P^- with z^+). Both pairings are consistent with the same marginal observations, but they yield opposite conclusions about which best-action set is backed by an eligible record. The assignment of outcomes to checkpoints therefore carries information beyond what marginal outcome distributions contain. We test this prediction experimentally in Section 5.3. The formal construction (compatible law–record relations and their target images) is in Appendix A.

Statistical certification. For execution blocks without checkpoint–action identifiers, we construct simultaneous confidence regions using Clopper–Pearson intervals (Clopper & Pearson, 1934) with a Bonferroni correction (Bonferroni, 1936) over the action set. The certified best-action set is the intersection of the confidence region with the set of possible maximizers (Appendix C.4).

5 Experiments

We evaluate exact-set agreement alongside held-out success on a frozen RecoveryBench panel of failed checkpoints and two planning cohorts. Execution records come from Qwen3-8B (Yang et al., 2025) and the GPT-5.6 family (OpenAI, 2026) (Luna, Terra, Sol). The central goal is to determine whether high set agreement predicts successful recovery. All experiments use frozen checkpoints in which the agent state is replayed from a saved snapshot, ensuring that the only source of variation is the recovery action and decoding randomness. Each experiment splits draws into a selection block (used to determine the best-action set) and a held-out block (used to evaluate the selected action's quality). Full protocols, statistical units, and uncertainty analyses are provided in Appendix B.

5.1 RecoveryBench: High Agreement, Low Success

We apply the complete-set readout to 12 frozen failed Terminal-Bench (Merrill et al., 2026) checkpoints from RecoveryBench (Tan et al., 2025). Each checkpoint is evaluated with three GPT-5.6 models and the actions `retry`, `verify`, and `replace_agent`, with eight repeats per action (864 total episodes). Repeats 0–3 form the selection block and repeats 4–7 the held-out block. For each $k \in \{1, 2, 4\}$ draws per action, the selection block yields a complete native-reward maximizer set, which is compared with the held-out set. Success is defined as native verifier reward >0.

At k=1, Luna's agreement reaches 85.4% while held-out success is only 1.4% (Table 1). The all-zero pairs fraction equals the agreement for Luna at k=1 (Table 1), indicating that every matching comparison consists entirely of zero-reward attempts. The models agree perfectly on which action is best because none of them ever succeed.

Table 1: RecoveryBench validation on 12 failed checkpoints. Each model row reports agreement, all-zero pairs, and held-out success for $k = 1, 2, 4$ draws per action. Agreement compares native-reward maximizer sets across blocks; all-zero pairs is the fraction of matching comparisons where every action scores zero in both blocks. Held-out success evaluates the selected set; intervals are in Appendix Table 47.

	Agreement			All-zero pairs			Held-out success		
Model	$k=1$	$k=2$	$k=4$	$k=1$	$k=2$	$k=4$	$k=1$	$k=2$	$k=4$
Luna	0.854	0.750	0.667	0.854	0.750	0.583	0.014	0.014	0.014
Terra	0.776	0.688	0.583	0.776	0.688	0.583	0.045	0.035	0.014
Sol	0.734	0.627	0.417	0.708	0.590	0.417	0.083	0.075	0.063

The pattern holds across all three models and three budgets (Table 1), no cross-block pair has positive reward for every action in both blocks (the all-success frequency is exactly 0). At the checkpoint level, agreement and held-out success are negatively correlated, with Spearman estimates ranging from -0.796 to -0.443 at k=1 (Appendix Table 50). Checkpoints where recovery sometimes works produce less agreement, because stochastic successes break the universal-failure ties. Under the assumption that agreement signals quality, high agreement should predict high success, yet the opposite holds, and the checkpoints with the most agreement are those where no action ever succeeds.

Selection by agreement vs. by outcome. Selecting models and actions by agreement attains only 1.3% held-out success at k=1, compared with 6.5% when selecting by mean outcome (Appendix H). A practitioner who chooses actions by set agreement would therefore perform worse than one who simply picks the highest mean score, even when that mean is low. Of the 864 episodes, 364 have a native verifier observation, 487 carry a model-protocol failure, and 13 carry an invalid adapter attempt (Appendix Table 49). The prevalence of protocol failures explains why success is so low. The majority of episodes never reach the verifier because the model fails to follow the recovery protocol, yet these failures still generate zero-reward outcomes that contribute to agreement.

Increasing k resolves ties but lowers agreement. As k increases from 1 to 4, agreement drops for all three models, e.g., from 0.854 to 0.667 for Luna and from 0.734 to 0.417 for Sol (Table 1). As the set-path symmetry result predicts, more observations make empirical ties less likely, producing smaller and more variable best-action sets. Held-out success, however, is essentially unchanged across budgets. Luna stays at 1.4% regardless of k (Table 1), confirming that the agreement decline reflects tie resolution, not quality degradation.

5.2 Planning: Agreement Drops While Quality Holds

The RecoveryBench results demonstrate the disconnect in a regime of near-universal failure. The planning experiments below test whether it persists when actions have moderate to high success rates.

We evaluate two frozen 24-checkpoint planning cohorts with GPT-5.6 Terra, Luna, and Sol. Each model–cohort panel has three recovery actions and eight draws per action (576 responses per panel, 3,456 total across six panels). The study cohort was designed first, and the replication cohort was drawn independently to test whether findings reproduce. Draws are split into selection and evaluation blocks. For each $k \in \{1, 2, 4\}$, all size-k subsets in the selection block form a complete public-score maximizer set. Fixed verification, selected on a separate development panel, serves as the one-call baseline. Full protocols are detailed in Appendix G.

Across all six panels, exact agreement drops from k=1 to k=4 (Table 2), with declines ranging from 0.039 to 0.320. For Terra in the replication cohort, agreement falls from 40.4% to 8.3%, a five-fold reduction. Yet held-out strict success is largely stable. The selected-minus-fixed differences at k=1 range from -0.072 to -0.006, with every 95% bootstrap interval containing zero in the replication cohort. Agreement is thus primarily tracking tie structure, not action quality. The pattern also holds across models with different overall capabilities. Sol, the strongest model, shows the lowest agreement at every budget because its higher success rate creates more variation in outcomes, breaking ties that weaker models preserve through shared failure.

Table 2: GPT-5.6 planning results on two frozen cohorts (point estimates; intervals in Appendix Tables 39 and 40). Agreement compares public-score maximizer sets across blocks. Selected strict success evaluates the chosen action mixture on held-out draws.

		Set agreement			Selected strict success			
Cohort	Model	$k=1$	$k=2$	$k=4$	$k=1$	$k=2$	$k=4$	Fixed verify
Study	Terra	0.318	0.204	0.167	0.699	0.695	0.710	0.771
Study	Luna	0.234	0.150	0.083	0.575	0.568	0.538	0.635
Study	Sol	0.424	0.331	0.167	0.722	0.729	0.750	0.729
Replication	Terra	0.404	0.238	0.083	0.765	0.754	0.762	0.771
Replication	Luna	0.299	0.212	0.250	0.727	0.728	0.719	0.760
Replication	Sol	0.497	0.422	0.458	0.808	0.814	0.825	0.844

Sources of the agreement decline. Table 3 decomposes agreement on the original Terra panel into contributions from singleton matches (both blocks select a unique winner) and multi-action matches (both blocks retain the same tied set). At k=1, nearly all agreement comes from multi-action matches (0.310 of 0.318). The blocks agree because most checkpoints produce ties among all three actions. As k grows, the mean set size shrinks from 2.29 to 1.52, multi-action agreement drops from 0.310 to 0.125, and singleton agreement rises only slightly. Success is unchanged throughout. The tie resolution does not affect which actions work, only which ones are declared "best."

A Bernoulli control confirms the mechanism analytically. With three equal-cost actions each having success probability 0.9, one-draw agreement is 0.553, and at four draws it falls to 0.183. Yet expected success remains 0.9. With success probability 0.1, one-draw agreement is also 0.553, illustrating the set-path symmetry in action (Appendix G.3).

Table 3: Decomposition of set agreement on the original 24 planning checkpoints (Terra). Single and Multiple are disjoint contributions to Agreement by the size of the matching sets. Mean size averages both blocks; Success evaluates the selected action on the held-out block.

k	Agreement	Single	Multiple	Mean size	Success
1	0.318	0.008	0.310	2.286	0.699
2	0.204	0.039	0.164	1.913	0.695
4	0.167	0.042	0.125	1.521	0.710

The replication cohort independently confirms the agreement decline (Table 2), with the same pattern of stable success despite falling agreement. Recomputing benefit without requiring correct self-reported totals preserves the decline in both panels (Appendix G.4). The 12-checkpoint Qwen3-8B panel, spanning rank, path, and filter tasks, shows 84.0% strict agreement at k=1 with 37.9% selected-action success (Appendix Table 46). Here too, matching sets include both all-failure and all-success blocks. In this panel, multi-action matching sets contribute 0.831 of the 0.840 agreement at k=1, and 58.3% of agreeing pairs have maximum score zero in both blocks (Appendix G.7). Strict successes concentrate in the rank and filter families. The path family contributes none, yet all three families produce high agreement.

The role of task difficulty. Across both cohorts, the disconnect between agreement and success is most pronounced in two regimes: (i) easy tasks where all actions succeed (producing agreement via universal success) and (ii) hard tasks where all actions fail (producing agreement via universal failure). Only tasks of intermediate difficulty, where some actions succeed and others fail, generate informative disagreement. These are precisely the cases where agreement would be useful for selecting actions. Set agreement is therefore least informative where it is needed most.

Empirical near ties. A natural attempt to boost agreement is to soften the exact-set comparison by retaining all actions within a tolerance ϵU_s of the best mean score, where U_s is the score range. This converts near-winners into co-winners, absorbing small fluctuations. Table 4 evaluates this at k=4 for two tolerances alongside the exact comparison. The results are mixed. In one cohort agreement increases slightly (from 0.083 to 0.250), but in the other it actually decreases (from 0.167 to 0.125). Held-out success changes negligibly in both cases. The tolerance does not reliably improve

agreement, and when it does, the improvement reflects the absorption of near-ties rather than a better signal about action quality. All tolerances and budgets are reported in Appendix G.5.

Table 4: Empirical tolerance sensitivity at $k = 4$ on 24 checkpoints per cohort. Strict is selected-action success on the held-out block; Public is normalized feasible benefit.

	Original Terra panel			Independent Terra panel		
ϵ	Agreement	Strict	Public	Agreement	Strict	Public
0.000	0.167	0.710	0.668	0.083	0.762	0.695
0.020	0.167	0.707	0.669	0.208	0.748	0.689
0.050	0.125	0.714	0.670	0.250	0.748	0.689
0.100	0.125	0.714	0.670	0.250	0.748	0.689

5.3 Record Assignments Change Comparisons

We test the prediction from Section 4.3 that changing checkpoint assignments, while preserving each action's outcomes, can change conclusions. On 48 local archive cells (Qwen3-8B, 12 prefixes $\times$ 4 faults), we fix all outputs and permute only their checkpoint assignments.

Table 5: Permutation controls on 48 archive cells. "Changed" counts cells where the permuted relation differs from the exact keyed reference. Columns report median cell proportions.

	Cell relation				
Association	Direct	Overlap	Discordant	Changed	Mean set size
Independent permutation	0.042	0.417	0.542	0.125	1.813
Shared permutation	0.042	0.438	0.521	0.083	1.875
Exact keyed	0.042	0.458	0.500	0.000	1.875

The shared and independent permutations change 4 and 6 of 48 cell conclusions at the median (Table 5), despite preserving every action's outcome multiset. The result confirms the prediction from Section 4.3, namely that the bindings between checkpoints and actions carry information beyond what the marginal outcome distributions contain. Only 4.2% of cells retain a singleton joint target under permutation (Table 5), meaning that for most cells, multiple checkpoint–action assignments are consistent with the same outcomes, and different assignments lead to different conclusions. If an evaluation pipeline fails to log which checkpoint produced which execution (e.g., due to batched processing or asynchronous dispatch), the resulting permuted bindings silently corrupt 8 to 13% of cell-level conclusions, exceeding typical measurement noise in agent evaluations and remaining undetectable from the outcome data alone.

In a separate retrieval audit, EC^2 (Golovin et al., 2010) requires 26 lookups to identify all action–eligibility conclusions, while eligibility-first settling completes in 19 (Appendix F.9). A restricted candidate class covers only 14 of 192 canonical tuples, causing all three selectors to stop incorrectly in the same three cells (Appendix F.8).

5.4 Certification Requires Large Margins

We validate the Hoeffding certificate (Eq. (5)) on controlled distributions with known margins (Figure 2). When the margin is large (0.40), valid certification rises from 40.5% at 64 branches to 97.9% at 128. In contrast, when the margin is small (0.06), certification stays below 0.1% even at 512 branches per action. No false certificates were observed across all 36 nested cells and 10 budgets (Appendix Table 11). At the margins typically observed in agent recovery experiments (success differences of 5 to 10 percentage points), certifying a unique best action would require hundreds or thousands of executions per action, far beyond current evaluation budgets. In a 12-checkpoint development pilot with 32 draws per action, only one pointwise certificate was positive, and simultaneous correction left all unresolved (Appendix G.8). Most current evaluations therefore operate in a regime where statistical certification is infeasible, reinforcing the importance of reporting outcome levels alongside agreement. The gap between what certification requires and what current

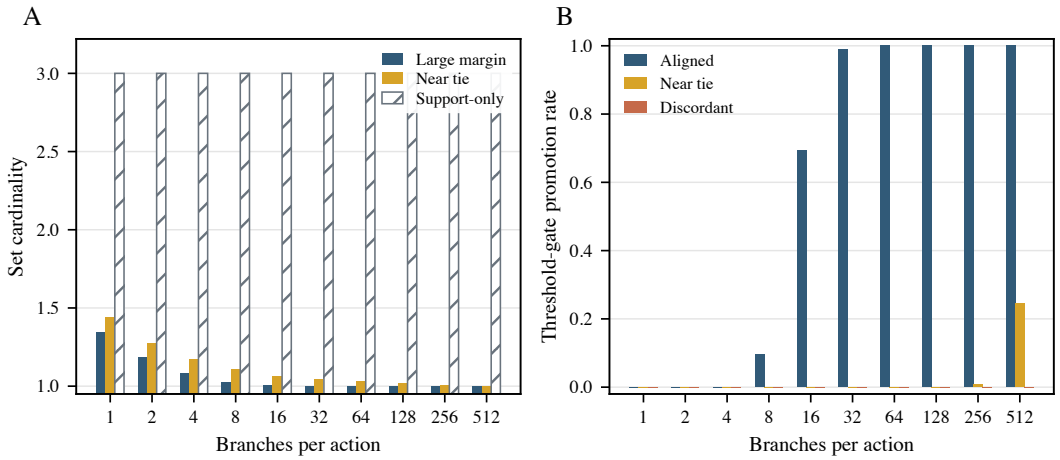


Figure 2: **(A)** Mean certified-set cardinality as the sample budget grows. With a large margin (0.40), the set shrinks to a singleton by $\sim$32 branches; a near tie (0.06) remains unresolved at 512. The support-only envelope (hatched) always retains all three actions. **(B)** Threshold-gate promotion rate. Large margins reach near-complete certification by 128 branches; near ties stay below 0.1% even at 512. Zero false certificates were observed across all conditions.

budgets provide also explains why practitioners default to set agreement in the first place. It is the only metric that returns a definitive answer at small sample sizes, even though that answer is uninformative about quality.

Changing the release target. On 10,000 controlled releases (Appendix Table 22), requiring a replay action to be among the expected maximizers (the overlap target) admits population ties, while requiring it to be the unique winner excludes them. Status-aware set membership admits more positives than the full gate with no observed errors under the overlap target, while the full gate makes fewer errors under the unique-winner target. Record checks eliminate constructed provenance-gap errors in both rules (Appendix E.3). The choice of release target thus has substantial practical consequences, as it determines which conclusions an evaluation can support and what evidence is needed.

6 Discussion

A consistent pattern emerges across benchmarks, models, and sample budgets. Set agreement can be high even when recovery performance is near zero, because shared failure produces agreement just as effectively as shared success. The set-path symmetry (Section 4.1) makes this precise, showing that Bernoulli parameters (p_1, p_2) and $(1-p_2, 1-p_1)$ yield identical best-action-set distributions. The RecoveryBench experiments confirm it empirically, with 85.4% agreement coexisting alongside 1.4% success. The planning experiments extend this finding to a regime where recovery does work. Agreement drops from 40% to 8% as sample budgets grow, yet held-out success remains stable.

These results motivate several concrete recommendations. Every set-agreement report should be accompanied by the fraction of all-zero comparisons, held-out success, and pooled success probability, all of which can be computed from data already collected. Agreement should be decomposed by outcome quality to reveal whether it is driven by shared failure, shared success, or informative differentiation. Declines in agreement across sample budgets should not be interpreted as quality drops, since they may reflect tie resolution rather than performance degradation. Finally, evaluation pipelines should log checkpoint-to-action bindings, as our permutation experiments show that losing them silently corrupts 8 to 13% of cell-level conclusions.

Limitations. Our experiments use only binary outcomes. For continuous-valued metrics, the symmetry argument requires adaptation, though the qualitative insight that ordinal rankings do not determine absolute levels still holds. The RecoveryBench experiments cover 12 checkpoints with three recovery actions, and richer action spaces would strengthen the empirical conclusions. We do

not propose or evaluate more sophisticated action-selection algorithms (e.g., Thompson sampling or UCB variants), which remain a natural direction for future work. Finally, our analysis assumes independent Bernoulli outcomes across rounds. In practice, LLM outputs may exhibit autocorrelation due to temperature or seed reuse, which could affect both agreement and certification rates.

Reproducibility Statement

All experiments use frozen checkpoint states and fixed random seeds to ensure reproducibility. The appendices provide formal definitions, complete proofs, experimental protocols with all hyperparameters, statistical units, scoring contracts, and source-specific reproduction details. The accompanying scientific artifact indexes the frozen execution records, analysis programs, and shell-reproducible commands organized by table and section. All 864 RecoveryBench episodes and 3,456 planning responses are included in the artifact with their full outcome vectors and checkpoint bindings.

AI Use Statement. Generative AI tools were used to assist with manuscript editing and to provide feedback during discussions of experimental design. All experimental protocols, analyses, implementations, and reported results were independently determined or verified by the authors. The authors reviewed all AI-assisted content and take responsibility for the final manuscript.

References

Miguel Angel Alvarado Gonzalez, Michelle Bruno Hernandez, Miguel Angel Peñaloza Perez, Bruno Lopez Orozco, Jesus Tadeo Cruz Soto, and Sandra Malagon. Do repetitions matter? strengthening reliability in LLM evaluations. *arXiv preprint arXiv:2509.24086*, 2025. URL https://arxiv.org/abs/2509.24086.

Jean-Yves Audibert, Sébastien Bubeck, and Rémi Munos. Best arm identification in multi-armed bandits. In *Proceedings of the 23rd Conference on Learning Theory*, pp. 41–53, 2010.

Carlo E. Bonferroni. *Teoria statistica delle classi e calcolo delle probabilità*, volume 8. Pubblicazioni del R. Istituto Superiore di Scienze Economiche e Commerciali di Firenze, 1936.

Leo Breiman, Jerome H. Friedman, Richard A. Olshen, and Charles J. Stone. *Classification and Regression Trees*. Wadsworth International Group, Belmont, CA, 1984. ISBN 0-534-98054-6.

A. Colin Cameron, Jonah B. Gelbach, and Douglas L. Miller. Bootstrap-based improvements for inference with clustered errors. *The Review of Economics and Statistics*, 90(3):414–427, 2008. doi: 10.1162/rest.90.3.414.

Timothy Christensen, Hyungsik Roger Moon, and Frank Schorfheide. Optimal decision rules when payoffs are partially identified, 2022. URL https://arxiv.org/abs/2204.11748.

C. J. Clopper and E. S. Pearson. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, 26(4):404–413, 1934. doi: 10.1093/biomet/26.4.404.

DeepSeek. DeepSeek chat completions api. API documentation, 2026. URL https://api-docs.deepseek.com/api/create-chat-completion/. Accessed September 7, 2026.

Rémy Degenne and Wouter M. Koolen. Pure exploration with multiple correct answers. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/hash/60cb558c40e4f18479664069d9642d5a-Abstract.html.

Bradley Efron. Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7(1): 1–26, 1979. doi: 10.1214/aos/1176344552.

Shengda Fan, Xin Cong, Yuepeng Fu, Zhong Zhang, Shuyan Zhang, Yuanwei Liu, Yesai Wu, Yankai Lin, Zhiyuan Liu, and Maosong Sun. WorkflowLLM: Enhancing workflow orchestration capability of large language models. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2411.05451.

Yanqin Fan, Emmanuel Guerre, and Dongming Zhu. Partial identification of functionals of the joint distribution of potential outcomes. *Journal of Econometrics*, 197(1):42–59, 2017. doi: 10.1016/j.jeconom.2016.10.005.

Laura Fuentes-Vicente, Mathieu Even, Gaëlle Dormion, Antoine Chambaz, Uri Shalit, and Julie Josse. Set-valued policy learning, 2026. URL https://arxiv.org/abs/2605.19830.

Aurelien Garivier and Emilie Kaufmann. Optimal best arm identification with fixed confidence. In *Proceedings of the 29th Annual Conference on Learning Theory*, pp. 998–1027, 2016.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems*, 2017.

Daniel Golovin, Andreas Krause, and Debajyoti Ray. Near-optimal bayesian active learning with noisy observations. In *Advances in Neural Information Processing Systems 23*, pp. 766–774, 2010. URL https://proceedings.neurips.cc/paper/2010/hash/1e6e0a04d20f50967c64dac2d639a577-Abstract.html.

Ashritha Gonuguntla. The replay gap: Static evaluation of model switching in LLM agents scores the wrong world, 2026.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963. doi: 10.1080/01621459.1963.10500830.

Steven R. Howard, Aaditya Ramdas, Jon D. McAuliffe, and Jasmeet Sekhon. Time-uniform, nonparametric, nonasymptotic confidence sequences. *The Annals of Statistics*, 49(2):1055–1080, 2021. doi: 10.1214/20-AOS1991.

Kevin Jamieson, Matthew Malloy, Robert Nowak, and Sébastien Bubeck. lil' ucb: An optimal exploration algorithm for multi-armed bandits. In *Proceedings of the 27th Conference on Learning Theory*, volume 35, pp. 423–439, 2014.

Emilie Kaufmann and Shivaram Kalyanakrishnan. Information complexity in bandit subset selection. In *Proceedings of the 26th Annual Conference on Learning Theory*, volume 30, pp. 228–251. PMLR, 2013. URL https://proceedings.mlr.press/v30/Kaufmann13.html.

Emilie Kaufmann, Olivier Cappé, and Aurélien Garivier. On the complexity of best-arm identification in multi-armed bandit models. *Journal of Machine Learning Research*, 17(1):1–42, 2016.

Tor Lattimore and Csaba Szepesvári. *Bandit Algorithms*. Cambridge University Press, 2020.

Lucien Le Cam. *Asymptotic Methods in Statistical Decision Theory*. Springer Series in Statistics. Springer-Verlag, New York, NY, 1986. doi: 10.1007/978-1-4612-4946-7.

Junchi Liao. Auditing provenance sensitivity in LLM agent action selection, 2026. URL https://arxiv.org/abs/2607.20827.

Xiao Liu et al. AgentBench: Evaluating LLMs as agents. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2308.03688.

Xun Liu, Yi Evie Zhang, Vira Kasprova, Parisa Rabbani, Pardis Sadat Zahraei, Tianyu Zhang, Ali Ebrahimpour-Boroojeny, and Varun Chandrasekaran. AgentAbstain: Do LLM agents know when not to act?, 2026. URL https://arxiv.org/abs/2607.10059.

Charles F. Manski. *Partial Identification of Probability Distributions*. Springer, 2003.

Mike A. Merrill, Alexander G. Shaw, Nicholas Carlini, et al. Terminal-Bench: Benchmarking agents on hard, realistic tasks in command line interfaces, 2026. URL https://arxiv.org/abs/2601.11868.

OpenAI. GPT-5.6: Frontier intelligence that scales with your ambition. Official model release announcement, July 2026. URL https://openai.com/index/gpt-5-6/. Published July 9, 2026; accessed September 8, 2026.

M. S. Pinsker. *Information and Information Stability of Random Variables and Processes*. Holden-Day, San Francisco, CA, 1964.

Jiaxing Qi, Zhongzhi Luan, Hongyu Zhang, Shaohan Huang, Carol Fung, Yongxin Tong, Hailong Yang, and Depei Qian. Can LLMs really recover microservice failures? a recovery-aware evaluation of diagnosis-to-action reasoning, 2026. URL https://arxiv.org/abs/2607.04623.

Parker Riley, Daniel Deutsch, George Foster, Viresh Ratnakar, Ali Dabirmoghaddam, and Markus Freitag. Finding replicable human evaluations via stable ranking probability. In *Proceedings of NAACL-HLT*, pp. 4908–4919, 2024. doi: 10.18653/v1/2024.naacl-long.275. URL https://aclanthology.org/2024.naacl-long.275/.

Jaineet Shah. Causal agent replay: Counterfactual attribution for LLM-agent failures, 2026. URL https://arxiv.org/abs/2606.08275.

Yining She, Yiliang Liang, and Eunsuk Kang. Safeguarding LLM agents from misalignment through provenance analysis, 2026. URL https://arxiv.org/abs/2607.01236.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 36, 2023. doi: 10.52202/075280-0377.

Shangyin Tan, Kevin Lin, Koushik Sen, and Matei A. Zaharia. Recovery-Bench: Evaluating agentic recovery from mistakes. In *NeurIPS 2025 Workshop: Evaluating the Evolving LLM Lifecycle*, 2025. URL https://openreview.net/pdf?id=8FZRnDgDxq.

Lan V. Truong. Optimal best-arm identification under fixed confidence with multiple optima, 2025. URL https://arxiv.org/abs/2505.15643.

Zexun Wang. CAVA: Canonical action verification and attestation for runtime governance of agentic AI systems, 2026a. URL https://arxiv.org/abs/2607.13716.

Zexun Wang. Proof-carrying agent actions: Model-agnostic runtime governance for heterogeneous agent systems, 2026b. URL https://arxiv.org/abs/2606.04104.

Edwin B. Wilson. Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, 22(158):209–212, 1927. doi: 10.1080/01621459.1927.10502953.

Qingyun Wu et al. Autogen: Enabling next-gen LLM applications via multi-agent conversation, 2023.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, et al. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

Shunyu Yao et al. React: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2210.03629.

Kohei Yata. Optimal decision rules under partial identification, 2021. URL https://arxiv.org/abs/2111.04926.

Yusheng Zheng, Xiaoyu Song, Yanpeng Hu, Lebin Cheng, Yuxi Huang, and Wei Zhang. When can agents safely checkpoint, fork, restore, and merge? exact checking for execution edits, 2026. URL https://arxiv.org/abs/2608.22928.

Dongsheng Zhu, Xuchen Ma, Yucheng Shen, Xiang Li, Yukun Zhao, Shuaiqiang Wang, Lingyong Yan, and Dawei Yin. When tools fail: Benchmarking dynamic replanning and anomaly recovery in LLM agents, 2026. URL https://arxiv.org/abs/2606.05806.

Peiying Zhu and Sidi Chang. ClaimReceipt: Verifying evidence sufficiency and coverage in agent evaluations, 2026. URL https://arxiv.org/abs/2609.01992.

Guide to the supplement. Definitions and proofs are collected in Appendix A; replay experiments and statistical units are specified in Appendix B. Appendix C develops identification boundaries and comparisons with prior work. Controlled evidence studies appear in Appendix D, with additional relation audits in Appendix E. Execution archives and binding retrieval are documented in Appendix F; the fresh repeated-branch study is reported in Appendix G; the frozen RecoveryBench panel and status-stratified analysis are presented in Appendix H.

A DEFINITIONS AND PROOFS

This appendix formalizes the compatible law–record relation used by the checkpoint audit and proves the certification claims. We first define the contract and compatible pairs, then derive the target images used in the main text.

A.1 JOINT IDENTIFICATION OF PREFERENCE AND RELEASE

Audit contract. An audit contract κ specifies a checkpoint state space $\mathcal{S}$, a recovery-action set A_0 with costs $C : A_0 \to \mathbb{R}_{\geq 0}$ (weighted by $\lambda \geq 0$), a class of outcome laws $\mathfrak{P}_s$, and admissible law–record associations $\mathcal{K}^{\mathrm{joint}}_{\kappa,s}$. Each state s has a selector-visible signature $o = \phi(s)$. The contract fixes the observable inputs; evaluator-side fault labels, answer keys, and unexecuted counterfactual outcomes are excluded.

Compatible pairs. Given observed outcomes y and record fields x, the set of outcome laws consistent with y is $\mathcal{P}_s(y) := \{P \in \mathfrak{P}_s : y \in \mathrm{supp}(P)\}$, and the records consistent with x are $\mathcal{Z}_{\kappa,s}(x) := \{z \in \mathcal{Z}^{\mathrm{all}}_{\kappa,s} : \psi_\kappa(z) = x\}$. The *compatible relation* intersects their product with the admissible associations:

$$\mathcal{H}_{\kappa,s}(y,x) := \big(\mathcal{P}_s(y) \times \mathcal{Z}_{\kappa,s}(x)\big) \cap \mathcal{K}^{\mathrm{joint}}_{\kappa,s}. \tag{6}$$

When this equals the full product $\mathcal{P}_s(y) \times \mathcal{Z}_{\kappa,s}(x)$ (i.e., the association imposes no additional constraint), we say the relation is *raw-rectangular*.

Source keys and candidate bridges. A source key $\ell^\star$ (e.g., a model identifier) can restrict the relation before any target is computed: $\mathcal{B}^{[\ell^\star]}_{\kappa,s}$ retains only pairs whose record maps to key $\ell^\star$. A *candidate bridge* $\mathcal{B}$ is a declared subset of law–record pairs representing a hypothesized data-generating process. The set of checkpoint states consistent with observed data under bridge $\mathcal{B}$ is

$$\mathcal{S}^{\mathcal{B}}_{\kappa|o}(\ell^\star; y, x) := \{s : \phi(s) = o,\ \mathcal{B}^{[\ell^\star]}_{\kappa,s} \cap \big(\mathcal{P}_s(y) \times \mathcal{Z}_{\kappa,s}(x)\big) \neq \varnothing\}. \tag{7}$$

Two bridge systems are *observationally equivalent* when they yield the same set of compatible state–law–record tuples. The collection of all candidate bridges consistent with the data is $\mathfrak{B}_{\kappa|o}(y, x, \ell^\star) := \{\mathcal{B} \in \mathcal{B}^{\mathrm{cand}}_\kappa : \mathcal{S}^{\mathcal{B}}_{\kappa|o}(\ell^\star; y, x) \neq \varnothing\}$.

A.2 POPULATION RECORDS AND FIXED-STATE TARGETS

From single checkpoints to populations. When the evaluation covers a population of checkpoints $\mathcal{D} \subseteq \mathcal{S}$, the record and law spaces are products over individual checkpoints: $\mathfrak{Z}^{\mathrm{all}}_{\kappa,\mathcal{D}} := \prod_{s\in\mathcal{D}} \mathcal{Z}^{\mathrm{all}}_{\kappa,s}$ and $\mathfrak{P}_{\kappa,\mathcal{D}}(\mathbf{y}) := \prod_{s\in\mathcal{D}} \mathcal{P}_s(y_s)$. The population-level compatible relation requires each per-checkpoint pair to be admissible, and its target image collects the per-checkpoint overlap sets alongside a population-level eligibility bit. For a single checkpoint ($|\mathcal{D}| = 1$) this reduces to the fixed-state case; for larger populations it is a vector-valued target that cannot be replaced by a scalar summary.

Record eligibility. Three binary checks determine whether a record supports reliable conclusions: *completeness* P_{rec} (all required fields present), *linkage* P_{link} (traceable to the generating model), and *observability* O_κ (scoring contract can be recomputed). We write $C_\kappa := P_{\mathrm{rec}} P_{\mathrm{link}} O_\kappa$ for the combined check. The per-checkpoint overlap status $v_s(z) \in \{A, B, D, \mathrm{NA}\}$ classifies whether the observed and reference best-action sets are equal (A), overlapping (B), disjoint (D), or missing (NA). Direct-transfer eligibility requires all checks to pass and no disjoint or ambiguous cells:

$$D^{\mathcal{D}}_\kappa(z) := \mathbf{1}\{P_{\mathrm{rec}}(z) = P_{\mathrm{link}}(z) = O_\kappa(z) = 1,\ d_{\mathcal{D}}(z) = b_{\mathcal{D}}(z) = 0\}, \tag{8}$$

where $d_{\mathcal{D}}(z) = \mathbf{1}\{\exists s : v_s(z) = D\}$ flags any disjoint cell and $b_{\mathcal{D}}(z)$ flags ambiguity in the absence of disjointness. An evidence update is *truth-separating* when records with different eligibility can always be distinguished by some observable evidence.

Target images. The joint target of the compatible relation maps each surviving pair (P, z) to its overlap set $I_s(P) = R_s \cap A^\star(P)$ and eligibility $D_\kappa(z)$:

$$\Gamma^{\mathcal{I}}_{\kappa,s}(y,x) := \{(I_s(P), D_\kappa(z)) : (P,z) \in \mathcal{H}_{\kappa,s}(y,x)\}. \tag{9}$$

The binary version $\Gamma_{\kappa,s}(y,x)$ replaces $I_s(P)$ with $T_s(P) = \mathbf{1}\{I_s(P) \neq \varnothing\}$ (whether the overlap is nonempty). We also write the coordinate projections as Γ^{pref}_s (preference images) and $\Gamma^{\mathrm{rel}}_{\kappa,s}$ (eligibility images).

A.3 PAIRING, SINGLETON CERTIFICATION, AND CANDIDATE COVERAGE

Exact image vs. marginal envelope. For a relation $\mathcal{B}$ between laws and records, the *exact* joint image $(I_s, D_\kappa)(\mathcal{B})$ collects only actually paired overlap-set/eligibility values, while the *marginal envelope* $I_s(\mathrm{pr}_P\,\mathcal{B}) \times D_\kappa(\mathrm{pr}_Z\,\mathcal{B})$ takes the Cartesian product of coordinate projections and can therefore include unsupported combinations. The exact image is always contained in the envelope; when equality holds the relation is *target-rectangular*. However, singleton certification is immune to this distinction.

$$|\Gamma^{\mathcal{I}}(\mathcal{B})| = 1 \iff |I_s(\mathrm{pr}_P\,\mathcal{B})| = |D_\kappa(\mathrm{pr}_Z\,\mathcal{B})| = 1. \tag{10}$$

Proof. A singleton joint image forces both coordinates to singletons; conversely, singleton coordinates produce exactly one Cartesian pair, which the nonempty relation attains. □

Why the pairing matters. The distinction is sharp beyond the singleton case. Two relations with identical coordinate projections but different law–record pairings (diagonal vs. crossed) produce different joint images:

$$\Gamma^{\mathcal{I}}(\mathcal{B}_{\mathrm{diag}}) = \{(I^+, 1), (I^-, 0)\}, \qquad \Gamma^{\mathcal{I}}(\mathcal{B}_{\mathrm{cross}}) = \{(I^+, 0), (I^-, 1)\}. \tag{11}$$

Any readout that sees only coordinate projections returns the same answer for both relations and cannot distinguish these cases (Table 6).

Table 6: Exact joint-image queries on finite constructions. N is the number of queries; Correct, FP, and Miss are counts for each readout.

		Joint readout			Marginal readout		
Query	N	Correct	FP	Miss	Correct	FP	Miss
Possible image	15	15	0	0	13	2	0
Singleton	15	15	0	0	15	0	0
Hidden pairing	2	2	0	0	1	1	0

Full-set certification. A gate that returns CONTINUE$(I^\dagger)$ is *sound* if and only if the robust candidate image $\Gamma^{\mathcal{I},\mathrm{rob}}_{\kappa|o}$ (the joint image over all surviving candidates, possibly restricted by a source key $\ell^\star$) equals the singleton $\{(I^\dagger, 1)\}$. The canonical full-set query combines this singleton check with the record eligibility check:

$$\mathsf{JointImage}^{\mathrm{rob}}_\kappa(I^\dagger \mid y, x) \Longleftrightarrow \Gamma^{\mathcal{I},\mathrm{rob}}_{\kappa|o}(y, x; \ell^\star) = \{(I^\dagger, 1)\} \ \wedge \ C_\kappa = 1. \tag{12}$$

If the image equals the displayed singleton, every surviving candidate shares the same nonempty overlap set and has an eligible record, so promotion is sound. Any additional value in the image (an ineligible record, empty overlap, or different overlap set) makes promotion unsound.

Active-bridge coverage. The certification above is relative to the declared candidate class. To extend it to the *active* relation (the true generating process), we require that every admissible active pair is covered by some candidate bridge: Coverage ensures that a positive candidate singleton is also sound for the active relation. A source key $\ell^\star$ that separates observable bridge classes ($\mathrm{Sep}^{\mathrm{obs}}_\kappa = 1$) further refines the candidate fiber, so that a singleton image under the key implies the canonical full-set gate returns CONTINUE$(I^\dagger)$.

A.4 ACTION-SET BRIDGE WITNESS

This minimal example shows why the bridge (law–record pairing) must be observable. Consider two actions $\{r, v\}$ with equal costs, both yielding success with probability 0.90, but differing in which action wins. Under law P_r action r is the unique expected winner; under P_v it is v:

$$P_r = 0.90\,\delta_{(1,1)} + 0.10\,\delta_{(1,0)}, \qquad P_v = 0.90\,\delta_{(1,1)} + 0.10\,\delta_{(0,1)}. \tag{13}$$

Each law is paired with its own eligible record (z_r or z_v), producing two candidate bridges $\mathcal{B}_r = \{(P_r, z_r)\}$ and $\mathcal{B}_v = \{(P_v, z_v)\}$. Their joint images are singletons pointing to opposite winners: $\Gamma^{\mathcal{I}}(\mathcal{B}_r) = \{(\{r\}, 1)\}$ and $\Gamma^{\mathcal{I}}(\mathcal{B}_v) = \{(\{v\}, 1)\}$. Without a key that identifies which record generated

the observation, the robust image contains both positive action sets, so the gate cannot authorize Continue. A source key resolves the ambiguity only if the candidate contract associates it injectively with the corresponding law.

A.5 Safety boundary and evidence requests

The dispatcher chooses among Promote, Hold, and Reject by minimizing worst-case loss over the surviving target image Γ. With $\theta = (T_s(P), D_\kappa(z))$ encoding whether the overlap is nonempty and the record is eligible, the loss penalizes false promotion more heavily than missed promotion ($c_{\mathrm{FP}} > c_H > 0$):

$$\ell(a, \theta) = c_{\mathrm{FP}} \mathbf{1}\{a = \textsc{Promote}, \theta \neq (1,1)\} + c_H \mathbf{1}\{a \neq \textsc{Promote}, \theta = (1,1)\}. \tag{14}$$

The minimax decision $\mathcal{G}^\star(\Gamma) := \arg\min_d \sup_{\theta \in \Gamma} \ell(d, \theta)$ contains only Promote when $\Gamma = \{(1,1)\}$ and excludes Promote whenever any nonpositive state survives.

When the image is not yet a singleton, the dispatcher requests additional evidence along each unresolved coordinate:

$$\mathsf{U}_{\kappa,s}(\mathcal{H}) = \{\mathsf{Repeat} : \Gamma_s^{\mathrm{pref}} = \{0,1\}\} \cup \{\mathsf{Collect} : \Gamma_{\kappa,s}^{\mathrm{rel}} = \{0,1\}\}. \tag{15}$$

Repeat requests another task attempt to resolve the preference coordinate; Collect requests record evidence to resolve the eligibility coordinate. Truth-separating evidence (Eq. (A.2)) collapses the corresponding coordinate to a singleton on the updated fiber.

B Experimental protocols and replay controls

B.1 Evaluation populations and statistical units

The coordinate-stress audit draws 2,000 independent outcome vectors in each of four regimes (eligible, near-tie, missing-linkage, and insufficient-observability). The separate release comparison has a fifth, discordant regime. The acquisition study uses 4,000 controlled states with truthful informative responses and 64 paired seeds; calibrated selection uses 32 calibration and 32 disjoint evaluation seeds. The preference study uses 10,000 independent vectors per operating point over 36 nested cells. Finite relation queries use enumerated truth, and the dependence study reuses 4,000 latent specifications under product and coupled laws.

The replay population has development, in-distribution (ID), and structural out-of-distribution (OOD) splits. The Qwen3-8B archive (Yang et al., 2025) contains 12 prefixes, four faults, three actions, and two repetitions (48 cells and 288 branches); assignment controls and uncertainty cluster by prefix. The strict score-aware filter retains 13 of 24 development prefixes, 28 of 48 ID prefixes, and 27 of 48 OOD prefixes (68 prefixes and 220 held-out prefix–fault cells).

B.2 Workflow actions and execution records

Replay compares fixed, status-only, and learned trace-only policies under a shared objective; transfer freezes the selector while the replay kernel is shifted, and completion, replay utility, and intervention cost are reported separately. Luna, Terra, and Sol each provide two executions per action on the same 48-cell surface (288 branches per model); Qwen3-8B provides a separate local archive with the same two-repetition budget. All 144 branches in each GPT-5.6 second execution satisfy the strict predicate. The pooled score averages the two strict outcomes per action before subtracting cost. Strict validity requires semantic correctness and public-score agreement; the local key binds outputs to checkpoint/action assignments, provider linkage is separate, and source groups remain distinct for artifact validity, record eligibility, and observed relations. The provider archive has 12 held-out prefixes, nine satisfying the strict score-aware contract, and a 108-branch evaluation population; its records lack complete linkage.

B.3 Comparison rules and outcome measures

Release comparisons fix branches and, for status-aware rules, supplied record information. The full gate combines observed action relations with record checks; A-only omits record checks; set

membership and serialized top-1 differ in their tie handling; Robust-95 requires statistical separation. Acquisition fixes the budget cap and final eligibility criterion while comparing exact/calibrated selection with fixed local/joint request patterns; request cost is reported with coverage and excludes calibration acquisition from calibrated evaluation cost.

Three controls isolate associations: finite constructions vary law–record pairings while preserving coordinate projections; the dependence study varies joint outcome distributions while preserving marginal action probabilities; archive controls reassign retained outputs to fault cells while preserving each action-specific within-prefix multiset. Promotion and acquisition coverage use eligible cases; errors and misses use each table's stated denominator. Exact queries report correct, unsupported, and missed counts, archive comparisons report observed cell-relation changes, and statistical certification reports valid and false certificates at fixed budgets.

B.4 Task construction and split composition

The benchmark contains ranking, path-selection, and constrained-filtering tasks on typed directed acyclic graphs (DAGs), varying depth, breadth, information horizon, and parallelism. Workers produce structured artifacts evaluated against deterministic answer keys. For a task graph $D = (V, E, \tau, \delta)$, τ maps nodes to finite types and δ stores typed payloads and dependency annotations. A faulted checkpoint is $s = \Psi_f(s_0)$, and for action a and generator noise $u \sim \nu_{s,a}$,

$$\omega = (\mathsf{B}, g) = \mathsf{Branch}(s, a, u), \qquad K_{\mathrm{rep}}^{Y,\kappa}(\cdot \mid s, a) = (Y_{s,a}^{\mathrm{rep},\kappa})_{\#} K_{\mathrm{rep}}^{\mathrm{raw}}(\cdot \mid s, a). \tag{16}$$

Here $\mathcal{F}$ contains four injected fault surfaces; Branch covers typed transition, artifact materialization, and propagation draws, while Clean checks schema, evidence, and semantics. The contract fixes the shared checkpoint/action interface while keeping replay law and endpoint record distinct.

B.5 Trace representation and action contract

The selector receives the task question, transcript prefix, intermediate artifact, dependency context, validator-visible checks, and public candidate table. Artifact checks cover status/schema/version, role compatibility, candidate membership, evidence validity, score presence, and score recomputation; structural features record depth, breadth, information horizon, and parallelism. Injected faults, answer keys, counterfactual branches, and future outcomes remain evaluator-side.

The action vocabulary is `continue`, `retry`, `verify`, and `replace_agent`, with protocol costs 0, 1, 1, and 2; materialization uses the three recovery actions. Replay utility subtracts the cascade penalty $0.01G$, while materialization stops at the recovered artifact and uses the shared-cost comparator. Strict score-aware success requires semantic cleanliness and public-score agreement within the declared tolerances. The frozen trace selector is a dependency-free CART (Breiman et al., 1984) with depth 5, balanced class weights, and seed 2026. It is fit on one row per development checkpoint–fault after collapsing action rows, then frozen before held-out replay and endpoint materialization; invalid or low-confidence traces (maximum class probability < 0.60) route to verification. The split manifest, counterfactual rows, objective-alignment record, and backend remain in the release bundle. Semantic-clean sensitivity and endpoint family composition remain in the live archive.

B.6 Primary replay uncertainty

Declared reference selects the action with highest expected utility under the original replay kernel, using a fixed order to resolve ties. In the shifted-kernel comparisons, Oracle selects the highest-utility action under the shifted kernel. Both use evaluator knowledge of their respective kernels; Trace-only chooses from the visible checkpoint features.

All replay intervals below are central 95% prefix-cluster bootstrap intervals; prefixes, rather than fault cells, are the resampling unit (Efron, 1979; Cameron et al., 2008). The same frozen transition kernel supplies objective-alignment, replay-headroom, and one-shot-reference rows. Trace-only–Verify completion differences are 0.4196 ($[0.3571, 0.4732]$) on ID and 0.4537 ($[0.3889, 0.5185]$) on structural OOD; utility differences are 0.4620 ($[0.4537, 0.4704]$) and 0.5732 ($[0.5562, 0.5911]$), respectively (Table 7). The shifted-kernel audit changes only action-outcome probabilities; states, traces, actions, costs, fault surfaces, and score contract stay fixed.

Table 7: Central 95% prefix-cluster intervals for score-aware replay success rates. Prefix is the resampling unit; endpoints are rounded outward to three decimals.

Policy	ID success			Structural OOD success		
	Estimate	Lower	Upper	Estimate	Lower	Upper
Continue	0.000	0.000	0.000	0.000	0.000	0.000
Blind retry	0.268	0.232	0.304	0.250	0.213	0.287
Always verify	0.438	0.366	0.500	0.407	0.343	0.472
Always replace	0.893	0.839	0.938	0.889	0.824	0.944
Status-only	0.670	0.589	0.741	0.611	0.546	0.676
Declared reference	0.857	0.795	0.920	0.852	0.796	0.907
Trace-only	0.857	0.795	0.920	0.861	0.806	0.917

Utility/regret intervals. On the same prefix-cluster plane, each entry is Estimate $[\mathrm{Lower}, \mathrm{Upper}]$: Original Reference has ID utility $0.805\,[0.803, 0.808]$, OOD utility $0.779\,[0.774, 0.783]$, and OOD regret $0.000\,[0.000, 0.000]$; Original Trace-only has $0.805\,[0.803, 0.808]$, $0.778\,[0.773, 0.783]$, and $0.001\,[0.000, 0.002]$; shifted Oracle has $0.753\,[0.750, 0.756]$, $0.720\,[0.714, 0.725]$, and $0.000\,[0.000, 0.000]$; shifted Trace-only has $0.735\,[0.731, 0.738]$, $0.691\,[0.684, 0.698]$, and $0.029\,[0.027, 0.031]$ for the same three metrics.

The original-kernel Trace-only structural-OOD regret interval $[0.000, 0.002]$ crosses the predeclared 0.001 boundary; the OOD label is therefore a point-estimate status and does not support transfer beyond this split. The shifted Trace-only interval $[0.027, 0.031]$ remains above that boundary. On the original kernel, Trace-only minus Always replace utility is $+0.0250$ on both ID and OOD. Completion contrasts are -0.0357 with $[-0.0893, 0.0179]$ on ID and -0.0278 with $[-0.0926, 0.0278]$ on OOD. These are protocol-cost contrasts on the frozen replay surface and do not estimate endpoint quality.

Unit-matched statistical audit. Wilson intervals (Wilson, 1927) apply to independent draws; paired-seed results retain cluster ranges, and release-vector results use a release-vector cluster bootstrap (Efron, 1979; Cameron et al., 2008). The generating unit is retained in every row: seed for allocation, cell for joint-law controls, draw for coordinate stress, and vector for model-class certificates. In the near-tie certificate row the estimate is 2×10^{-5} before rounding, with original endpoints 1×10^{-5} and 4×10^{-5}; displayed 0.000 is not necessarily an observed zero.

Key adverse rows are coordinate near-tie $0.049\,[0.040, 0.060]$ at $N = 2000$, joint discordant FP $0.000\,[0.000, 0.001]$ at $N = 4000$, and near-tie certification $0.000\,[0.000, 0.001]$ at $n = 256$, $N = 10000$; the full row-level audit remains in the release record.

B.7 Selector-fit stability

Two hundred strict-development-prefix bootstrap refits leave action selection invariant on ID and structural OOD: held-out utility ranges are $[0.805, 0.806]$ and $[0.778, 0.778]$, regret is 0.000 on both splits, and action agreement with the frozen fit is 1.000.

B.8 Fixed-kernel replay controls

Table 8: Frozen replay policies evaluated with U_{pol}. C_{emp} is strict completion, U_{pol} expected replay utility, and c mean intervention cost over 112 ID and 108 OOD cells.

	ID			Structural OOD		
Policy	C_{emp}	U_{pol}	c	C_{emp}	U_{pol}	c
Blind retry	0.268	0.145	1.000	0.250	-0.039	1.000
Always verify	0.438	0.343	1.000	0.407	0.205	1.000
Always replace	**0.893**	0.780	2.000	**0.889**	0.753	2.000
Status-only	0.670	0.583	1.000	0.611	0.499	1.000
Trace-only	0.857	**0.805**	1.500	0.861	**0.778**	1.500

Trace-only has the highest replay utility, exceeding Always replace by 0.025 on each split at mean cost 1.50 versus 2.00, while Always replace has the highest completion rate (Table 8). The propagation-aware control is

$$U_{\mathrm{pol}}(s,a) = \mathbb{E}_{\omega\sim K_{\mathrm{rep}}^{\mathrm{raw}}(\cdot|s,a)}\left[Y^{\mathrm{rep},\kappa_{\mathrm{pol}}}(s,a;\omega) - \lambda C(a) - 0.01G(s,a;\omega)\right]. \tag{17}$$

It is a workflow control and does not alter the shared-cost preference or release estimands. Replay headroom has a minimum top-two margin of 0.010 on ID, structural OOD, and the combined strict test; replacement-ratio sensitivity is retained in the release bundle.

C Related work and identification boundaries

Best-arm identification supplies the repeated-arm reference (Audibert et al., 2010; Garivier & Kaufmann, 2016; Jamieson et al., 2014; Kaufmann et al., 2016; Truong, 2025); partial-identification work supplies compatibility-set language (Manski, 2003; Fan et al., 2017; Yata, 2021; Christensen et al., 2022); selective and set-valued prediction supply abstention and alternative-action interfaces (Fuentes-Vicente et al., 2026; Geifman & El-Yaniv, 2017; Liu et al., 2026). Replay and recovery studies measure scalar utility, repair, or trajectory validity (Gonuguntla, 2026; Shah, 2026; Qi et al., 2026); execution-edit safety work studies checkpoint, fork, restore, and merge semantics (Zheng et al., 2026); provenance and action-certificate studies verify source or execution objects (Liao, 2026; She et al., 2026; Wang, 2026a;b; Zhu & Chang, 2026). OrchLoop adds a contract-relative action-set relation paired with release status at one checkpoint.

The live endpoint receives task and candidate documents, the faulted artifact, dependency context, and retained prior-artifact score; validator-visible checks are available to the selector but are not additionally serialized to the endpoint. Fault labels, answer keys, future outcomes, and unexecuted counterfactual results are excluded from both inputs. The reconstructed-input manifest hashes all 108 recovery inputs; missing provider request identifiers, complete prompts, and source-complete response records set provider linkage to zero.

C.1 Serialization sensitivity of top-1 transfer

For a fixed action order σ, let $\mathrm{Pref}_\sigma(k)$ be its first k actions, $H_\sigma(s;L)$ the earliest-action readout, and $M_s(L) = \mathbf{1}\{\mathcal{R}_s^{\mathrm{sc}} \cap L \neq \varnothing\}$. Then

$$[\forall\, \varnothing \neq L \subseteq \mathcal{A}_{\mathrm{rec}} :\ H_\sigma(s;L) = M_s(L)] \iff \mathcal{R}_s^{\mathrm{sc}} = \mathrm{Pref}_\sigma(|\mathcal{R}_s^{\mathrm{sc}}|)\,. \tag{18}$$

Thus materialization is set-valued and order-invariant, while top-1 can change under serialization. The primary readout retains ties; unique-winner sensitivity is diagnostic.

For a nonempty live-best set L, the replay-relative status is

$$\nu_{\mathcal{R}}(L) = \begin{cases} A, & L = \mathcal{R}, \\ B, & L \cap \mathcal{R} \neq \varnothing,\ L \neq \mathcal{R}, \\ D, & L \cap \mathcal{R} = \varnothing. \end{cases} \tag{19}$$

For a nonempty cell population $\mathcal{D}$, write $d_{\mathcal{D}} = \mathbf{1}\{\exists s : v_s = D\}$, $b_{\mathcal{D}} = \mathbf{1}\{d_{\mathcal{D}} = 0, \exists s : v_s = B\}$, and let $P_{\mathrm{rec}}, P_{\mathrm{link}}, O_\kappa$ be record, linkage, and observability checks. On the declared three-status domain,

$$M_{\mathcal{D}} = \mathbf{1}\{\forall s : v_s \in \{A,B\}\} = 1 - d_{\mathcal{D}}. \tag{20}$$

The admissible interface class and its sufficient quotient are

$$\mathfrak{A}_\kappa(\mathcal{R};\mathcal{D}) = \mathcal{V}_{\mathcal{R}}^{\mathcal{D}} \times \{0,1\}^3 \times \mathrm{Bij}(\mathcal{A}_{\mathrm{rec}}, [m]). \tag{21}$$

$$Q_\kappa(\mathcal{D}) = (M_{\mathcal{D}}, d_{\mathcal{D}}, b_{\mathcal{D}}, P_{\mathrm{rec}}, P_{\mathrm{link}}, O_\kappa). \tag{22}$$

The reduced interface drops the redundant $M_{\mathcal{D}} = 1 - d_{\mathcal{D}}$ coordinate. A two-bit card summary is insufficient: Partial and Provider-gap cards can share $(\min M, \min H_\sigma) = (1,1)$ while the contract returns Hold-Ambiguity and Hold-Record-Completeness, respectively. The proof is by toggling one status or record coordinate at a time; serialization affects H_σ but none of the gate coordinates. If missing-information status NA is admitted, $M_{\mathcal{D}}$ must explicitly exclude both D and NA, and a required NA forces a completeness hold.

C.2 Observational non-identifiability of a trace-only gate

The masked probe replaces the five marker keys while recomputing public derived fields. Let $\mathcal{U}_T$ be the transformed observable classes and C_u the corresponding collision class. If a class contains different recovery labels, no deterministic selector using only that observable input can match all labels. With singleton replay labels,

$$B_{\mathrm{acc}} = \frac{1}{N_{\mathrm{obs}}} \sum_{u \in \mathcal{U}_T} \max_{a \in \mathcal{A}_{\mathrm{rec}}} n_{u,a}. \tag{23}$$

$$B_U = \frac{1}{N_{\mathrm{obs}}} \sum_{u \in \mathcal{U}_T} \max_{a \in \mathcal{A}_{\mathrm{rec}}} \sum_{z \in C_u} Q^{\mathrm{replay}}_{\kappa_{\mathrm{pol}}}(z, a). \tag{24}$$

where $n_{u,a}$ counts labels in C_u and

$$Q^{\mathrm{replay}}_{\kappa_{\mathrm{pol}}}(z, a) = \mathbb{E}_{\omega \sim K^{\mathrm{raw}}_{\mathrm{rep}}(\cdot|z,a)} \left[\Pi_{\mathrm{pol}}(z, a; \omega) \right]. \tag{25}$$

The five-key mask yields same-prefix collisions with distinct recovery-optimal labels in a held-out ID witness. The field vector and witness identifier remain in the release artifact; the construction establishes non-identifiability under the mask, while model-specific inference of the hidden mechanism remains unmeasured. The unit is a single-worker checkpoint; worker-pool independence and a persistent process are outside the measured object.

C.3 One-shot expected-ranking identification audit

Why one observation is never enough. Let the rich class $\mathfrak{P}_{\mathrm{rich}}$ contain all independent-Bernoulli product laws over recovery actions, with cost-adjusted utility $q_a(P) = p_a - \lambda C(a)$ and best-action set $A^\star(P) = \arg\max_a q_a(P)$. Under this class, every outcome vector y has positive probability under some law whose best actions exclude any given target set $\mathcal{R}_{\mathrm{sc}}$ (set the target actions' success rates near zero and competitors' near one). Therefore any gate with zero false-promotion rate must abstain on every single observation:

$$\sup_{P \in \mathfrak{P}_{\mathrm{rich}} : A^\star(P) \cap \mathcal{R}_{\mathrm{sc}} = \varnothing} P\{g(Y) = 1\} = 0 \quad \Longrightarrow \quad g \equiv 0. \tag{26}$$

The same argument extends to any fixed finite number of independent branches. Controlled-risk release requires either repeated evidence or structural restrictions on the law class.

Observable hidden-state mixtures. When multiple hidden checkpoints share the same observable signature o, the observed outcome is a mixture over checkpoint-specific kernels. The mixture utility set $\mathcal{U}^{\mathrm{obs}}_o(y)$ collects all possible expected-utility vectors consistent with y and the declared mixing family, and the transfer bit $\mathcal{T}^{\mathrm{obs}}_R(y)$ records whether the target action set R overlaps the best-action set for each such utility vector. A full-support bad law in the fiber restores the one-shot abstention boundary.

Repeated support and matched coupling. With repeated independent outcome vectors, a bad-law class whose support covers the entire sample space still leaves no safe promotion region. However, the support geometry depends on the *coupling* between actions within each vector. Consider two laws with identical per-action marginals $\mathrm{Bernoulli}(\theta)$ and $\mathrm{Bernoulli}(1-\theta)$, a shared-shock law (actions succeed or fail together via a common uniform draw U) and a product law (actions are independent).

$$P^{\mathrm{ss}}_\theta = \mathrm{Law}\big(\mathbf{1}\{U \le \theta\}, \mathbf{1}\{U \le 1-\theta\}\big), \quad P^{\times}_\theta = \mathrm{Bernoulli}(\theta) \otimes \mathrm{Bernoulli}(1-\theta). \tag{27}$$

The product law has full support on $\{0,1\}^2$, so its bad subfamily leaves no safe region. The shared-shock law for $\theta < 1/2$ cannot produce $(1,0)$, making that outcome a safe promotion witness for target $\{1\}$:

$$\mathcal{X}^{\mathrm{safe}}_{\{1\}}(\{P^{\times}_\theta\}) = \varnothing, \qquad \mathcal{X}^{\mathrm{safe}}_{\{1\}}(\{P^{\mathrm{ss}}_\theta\}) = \{10\}. \tag{28}$$

If the coupling is undeclared, the safe region is the intersection across admissible classes, so a gate that cannot distinguish couplings must be conservative wherever any class has a bad law.

C.4 FINITE-BRANCH CERTIFICATION AND STATISTICAL RELEASE

Hoeffding-based certificates. Given n branches per action, the empirical cost-adjusted utility $\widehat{q}_{a,n} = n^{-1}\sum_i Y_{a,i} - \lambda C(a)$ and Hoeffding radius $\varepsilon_n(\delta) = \sqrt{\log(2m/\delta)/(2n)}$ define a singleton-or-abstain certificate:

$$\widehat{\mathcal{R}}_{n,\delta} = \left\{ a : \widehat{q}_{a,n} - \max_{b\neq a} \widehat{q}_{b,n} > 2\varepsilon_n(\delta) \right\}, \quad \Pr\{\widehat{\mathcal{R}}_{n,\delta} \subseteq A^\star\} \geq 1-\delta. \tag{29}$$

The proof is a union bound; within-vector dependence is allowed. With a unique target gap $\Delta > 0$, certification succeeds with probability $\geq 1-\delta$ once $n > 8\log(2m/\delta)/\Delta^2$, and the misidentification probability decays exponentially:

$$\Pr\{\widehat{A}_n \neq \{a^\star\}\} \leq 2(m-1)\exp(-n\Delta^2/2). \tag{30}$$

The near-tie gap 0.06 remains largely unresolved at $n = 512$, while the large-margin certificate rises from 40.5% at $n = 64$ to 97.9% at $n = 128$ with no false certificates.

Statistical release. If the confidence relation covers the true law–record pair (failure $\leq \delta$) and the decoder is correct (failure $\leq \eta$), then a singleton release must equal the true target:

$$\Pr\{\text{STATISTICALCONTINUE outputs an incorrect target}\} \leq \delta + \eta. \tag{31}$$

An additional bridge-omission term α gives $\alpha + \delta + \eta$; a coverage bound already including bridge omission must not double-count. The conservative maximizer image from simultaneous utility intervals $[l_a, u_a]$ is:

$$\mathcal{S}(l,u) = \{\varnothing \neq S \subseteq A_0 : \max_{a\in S} l_a \leq \min_{a\in S} u_a,\ \max_{b\notin S} l_b < \min_{a\in S} u_a\}. \tag{32}$$

Calibration boundary. For a model class where the winner is unknown but the gap is at least Δ, the Le Cam–Pinsker lower bound (Le Cam, 1986; Pinsker, 1964) gives the minimax error for identifying the best action from n branches with gap 2ζ:

$$\inf_{\widehat{a}} \sup_{\theta\in\{+,-\}} \Pr_\theta\{\widehat{a} \neq a^\star_\theta\} \geq \tfrac{1}{2}\left(1 - \sqrt{2n\zeta \log \tfrac{1+2\zeta}{1-2\zeta}}\right)_+ . \tag{33}$$

This also lower-bounds set certification, since a correct singleton induces a correct selector. Fixed-budget resolution follows best-arm analyses (Audibert et al., 2010; Kaufmann et al., 2016; Jamieson et al., 2014; Lattimore & Szepesvári, 2020); stopping-aware evidence requires time-uniform confidence sequences (Howard et al., 2021).

Constructing a relation from anonymous execution blocks. Suppose m stored blocks of n observations each are available, one block per action, but the action bindings are hidden. Let $\mathcal{M}(x)$ enumerate all block-to-action bijections consistent with the visible metadata. For each block j, construct a Clopper–Pearson interval $[L_j, U_j]$ at level δ/m (Clopper & Pearson, 1934; Bonferroni, 1936). The confidence relation is:

$$\widehat{H}_\delta = \bigcup_{\beta\in\mathcal{M}(x)} \{(P, z_\beta) : L_j \leq \mathbb{E}_P[Y_{\beta(j)}] \leq U_j\ \forall j\}. \tag{34}$$

Coverage $\geq 1-\delta$ holds because only the true binding's m means need to be covered, and there is no multiplicity penalty for $|\mathcal{M}(x)|$. If metadata or a source receipt can exclude the true binding with failure probabilities α_M and α_R, coverage becomes $1-\delta-\alpha_M-\alpha_R$ by a union bound. This construction separates the checkable source-assignment contract from the sampling assumptions needed for preference inference.

Exact ties under controlled error. Full-set certification faces a fundamental limit at exact ties. If p_0 is a parameter where two or more actions are exactly tied and nearby parameters $p_k \to p_0$ break the tie, then any uniformly valid certifier satisfies:

$$\sup_p \Pr_p\{\text{release an incorrect set}\} \leq \delta \quad\Longrightarrow\quad \Pr_{p_0}\{\text{release } I_0 \text{ in finite time}\} \leq \delta. \tag{35}$$

Proof. Releasing I_0 is incorrect at every p_k, so $\Pr_{p_k}(E \cap \{T \leq N\}) \leq \delta$. Finite-horizon TV continuity gives the same bound at p_0; taking $N \to \infty$ proves the claim. □

Thus a controlled-error rule cannot reliably certify an exact multi-action tie from finite outcomes over the unrestricted class. A declared gap or tolerance-based target changes this problem; standard unique-best-arm formulations impose uniqueness explicitly (Garivier & Kaufmann, 2016).

Compatibility boundary. With one branch vector, the realized overlap is exact, but the rich class allows every action to be a possible unique expected winner. The expected-transfer bit therefore has support-compatible range $\{0, 1\}$ in every cell. Simultaneous 95% Clopper–Pearson intervals (Clopper & Pearson, 1934; Bonferroni, 1936) are reported as model-class diagnostics, separate from the realized readout.

D Controlled evidence and decision studies

D.1 Generators for the main release and request comparisons

Release population. Table 22 uses `independent_gate_decision_comparison.py`: five equally weighted regimes, 2,000 independent releases each, and four cells per release. The action order is retry, verify, replacement, with normalized costs $(1, 1, 2)$, $\lambda = 0.05$, and deductions $c_a = \lambda C(a)$. Each release chooses replay action r uniformly; b is the first other action in that order. In the order $(r, b, \text{remaining})$, the Bernoulli mean triples are $(0.76, 0.44, 0.44)$ for eligible and both record-gap regimes, $(0.62 + c_r, 0.62 + c_b, 0.44)$ for the tied regime, and $(0.46, 0.76, 0.44)$ for discordance. All four cells use that triple and 32 draws per action. Conditional on these choices, outcomes are independent across draws, actions, cells, and releases; every rule sees the same draws. The separate eight-draw output is not included in this table.

Completeness is zero only for provider gap, observability is zero only for observability gap, and linkage is fixed at one. Thus $h_i = P_{\mathrm{rec},i} O_{\kappa,i}$ incorporates all varying supplied checks. The eligible regime has $(u_i, o_i) = (1, 1)$, the tied regime $(0, 1)$, and the other three $(0, 0)$, using the definitions in Section E.2. These are design weights rather than estimated task frequencies. Each observed set maximizes the empirical mean minus cost, with numerical tie tolerance 10^{-12}. Membership requires r in every cell's observed set; Full gate requires each set to be exactly $\{r\}$ and supplied checks to pass. Top-1 uses the fixed action order. Robust-95 requires the replay lower bound to exceed every competitor upper bound with radius $\sqrt{\log(6/0.05)/(2 \cdot 32)}$. The generator seed is 20260825.

Request population and responses. Table 15 uses 4,000 states: 250 per pair of family (rank, path, filter, delegation) and unresolved-coordinate regime (preference only, release only, both, neither). Within each family–regime group, index states by $k = 0, \ldots, 249$. An unresolved coordinate starts at $\{0, 1\}$ and a resolved coordinate at $\{1\}$. Truths (t_P, t_R) are $(k \bmod 2, 1)$, $(1, k \bmod 2)$, $(\lfloor k/2 \rfloor \bmod 2, k \bmod 2)$, and $(1, 1)$ in those four regimes. The population contains 2,248 eligible and 1,752 ineligible states, including 248 eligible states with both coordinates unresolved.

The five requests are Repeat, Collect, JointCertificate, RepeatReplica, and CollectReplica, with costs $(0.700, 0.800, 1.100, 0.850, 0.650)$ and coverage sets $\{P\}, \{R\}, \{P, R\}, \{P\}, \{R\}$ in that order. Budget is 1.50. Every one of the 32 subsets is considered, subject to its total cost and coverage. Family-specific (q_R, q_C) are $(0.90, 0.84)$, $(0.82, 0.88)$, $(0.86, 0.84)$, and $(0.78, 0.92)$. The remaining informative probabilities are $q_J = q_R q_C + 0.04$, $q_{R'} = \max(0.05, q_R - 0.06)$, and $q_{C'} = \max(0.05, q_C - 0.08)$. The five response indicators are independent conditional on family/state; the joint request has its own indicator. An informative response reveals the true covered coordinates; otherwise the fibers remain unchanged. Exact selection maximizes complete informative coverage over feasible subsets.

Potential responses keyed by seed, state, and request are shared across policies. Exact and fixed rules use 64 seeds; calibrated selection uses 32 calibration seeds and 32 disjoint evaluation seeds, choosing from response coverage without truth labels. Evaluation cost in Table 15 excludes calibration; its separate cost analysis is in Section D.7. The source files and frozen response summaries are listed by table label in the supplementary code index.

Observable contract and utility. The observable perturbation fixes the strict score-aware replay states, action kernel, labels, prefix-cluster bootstrap, and eight nested deterministic noise draws per prefix. It changes only the trace. The five masked keys are `status_timeout`, `version_current`, `role_extractor`, `schema_v1`, and `score_matches_answer`, with neutral values $(0, 1, 1, 1, 1)$; derived status and score fields are recomputed. The unmasked view

has no cross-fault collision, whereas the five-key view creates the declared conflicting classes and supplies the $O = 1$ versus $O = 0$ audit. The public scheduler bundle contains 90 task contexts and 392 opaque decision states. Codebook controls remap only the four fault-signature codes, evaluating all 24 within-prefix bijections with frozen and transformed-view refits (Table 9).

Table 9: Observable-codebook shift on held-out replay. Agreement is with the held-out replay action; fixed and refit selectors receive the same transformed cells.

Split	Original	Fixed	Refit	Fixed regret
ID	**1.000**	0.375	**1.000**	0.293
Structural OOD	0.880	0.405	0.880	0.361

The collision audit uses the full transformed selector input and permits dependence across fault cells. Its best common-action accuracy and utility ceilings are reported in Table 10. The frozen selector has shifted-law regret 0.019 on 112 ID cells and 0.029 on 108 structural-OOD cells; its choices agree with the primary selector on all 220 cells. Interpolating the original and transfer kernels as $p_\alpha = (1-\alpha)p_{\text{original}} + \alpha p_{\text{transfer}}$ for $\alpha \in \{0, .25, .50, .75, 1\}$ gives frozen regret $(0.000, 0.001, 0.004, 0.008, 0.019)$ on ID and $(0.001, 0.004, 0.008, 0.015, 0.029)$ on OOD. Condition-refit regret is zero at every displayed interpolation point except 0.001 at $\alpha = 0$ on OOD.

Table 10: Collision-class ceilings and observed agreement under the five-key view. Rates and counts use separate columns; B_U is the best common-action replay-utility ceiling and Floor is its collision-regret floor.

Split	Obs. rate	Obs. n	B_{acc}	$B_{\text{acc}}\ n$	B_U	Floor
ID	0.500	56	0.750	84	0.804	0.002
Structural OOD	0.380	41	0.870	94	0.779	0.000

For checkpoint s, action a, and replay branch ω, define the semantic outcome $Y^{\text{sem}}(s, a; \omega)$ and

$$U^{\kappa_{\text{pol}}}(s,a) = \mathbb{E}_{\omega \sim K^{\text{raw}}_{\text{rep}}(\cdot|s,a)}[\Pi_{\text{pol}}(s,a;\omega)], \qquad \Pi_{\text{pol}} = Y^{\kappa_{\text{pol}}} - 0.05C(a) - 0.01G(s,a;\omega). \tag{36}$$

Here $Y^{\kappa_{\text{pol}}} = Y^{\text{sem}} S_{\text{score}}$ requires semantic validity and score agreement; score comparison uses relative tolerance 10^{-4} and absolute tolerance 0.05. The replay readout evaluates this predicate on $K^{\text{raw}}_{\text{rep}}$, while the live readout evaluates retained endpoint artifacts without a propagation coordinate. The charge G uses the frozen `cascade_mass` draw. The policy reference is

$$\mathcal{R}_{\text{pol}}(s) = \arg\max_{a \in \mathcal{A}_{\text{full}}} U^{\kappa_{\text{pol}}}(s,a). \tag{37}$$

The one-shot score-aware relation removes G and restricts to $\mathcal{A}_{\text{rec}}$; the collision audit uses the recovery-only projection. All 462 semantic-clean counterfactual branches are score-consistent. The 90-prefix semantic-clean run is a separate sensitivity analysis, and its live utility omits cascade mass because execution stops after recovery.

D.2 Contract-level decision rule

The release gate maps a six-dimensional binary input (population disjointness $d_{\mathcal{D}}$, ambiguity $b_{\mathcal{D}}$, record completeness P_{rec}, linkage P_{link}, observability O_κ, and membership $M_{\mathcal{D}}$) to a disposition via the first applicable case.

$$\widetilde{G}_\kappa = \begin{cases} \mathsf{H}_R, & P_{\text{rec}} = 0, \\ \mathsf{H}_L, & P_{\text{rec}} = 1,\ P_{\text{link}} = 0, \\ \mathsf{H}_O, & P_{\text{rec}} = P_{\text{link}} = 1,\ O_\kappa = 0, \\ \mathsf{R}_D, & P_{\text{rec}} = P_{\text{link}} = O_\kappa = 1,\ d_{\mathcal{D}} = 1, \\ \mathsf{H}_A, & C_\kappa = 1,\ d_{\mathcal{D}} = 0,\ (b_{\mathcal{D}} = 1 \vee M_{\mathcal{D}} = 0), \\ \mathsf{P}_{\text{obs}}, & C_\kappa = M_{\mathcal{D}} = 1,\ d_{\mathcal{D}} = b_{\mathcal{D}} = 0. \end{cases} \tag{38}$$

A repeat request is issued only for target-separated ambiguity with positive margins and available budget; it cannot bypass record or observability holds.

Coordinate necessity. Each coordinate is necessary. Starting from the all-pass state $(0, 0, 1, 1, 1) =$ PROMOTE, flipping any single coordinate changes the disposition (to rejection, ambiguity hold, record hold, linkage hold, or observability hold respectively). No coordinate can be removed without collapsing distinct dispositions.

An endpoint scalar seeing only the replay-best set and observed best set cannot reproduce the gate, since two records with identical scalars but $P_{\mathrm{rec}} = 0$ vs. 1 receive HOLD vs. PROMOTE.

D.3 MODEL-CLASS BRANCH-BUDGET AUDIT

The finite-branch audit varies repeated evidence while the evaluator knows the unique expected winner and cost-adjusted gap: 0.40 for the large-margin class, 0.06 for the near tie, and 0.36–0.40 for the heterogeneous class. Each of 36 nested cells uses budgets $n \in \{1, 2, 4, 8, 16, 32, 64, 128, 256, 512\}$ and 10,000 independent releases (seed 20260822). The same Hoeffding rule in Eq. (5) is used at every budget. The sufficient gap thresholds $4\rho_n(0.05)$ at $n = 32, 64, 128, 256, 512$ are $1.094, 0.774, 0.547, 0.387, 0.274$.

Table 11: Finite-branch preference certification at $\delta = 0.05$. Valid is the fraction of 360,000 cells with a correct unique-winner certificate; False counts incorrect certificates. Small nonzero rates may round to 0.0000; exact counts follow below.

	$n = 32$		$n = 64$		$n = 128$		$n = 256$		$n = 512$	
Class	Valid	False	Valid	False	Valid	False	Valid	False	Valid	False
Near tie	0.0000	0	0.0000	0	0.0000	0	0.0000	0	0.0004	0
Heterogeneous	0.0124	0	0.3084	0	0.9353	0	1.0000	0	1.0000	0
Large margin	0.0174	0	0.4054	0	0.9785	0	1.0000	0	1.0000	0

Certification reaches one for the large-margin and heterogeneous classes by 256 branches, while the near tie remains at 0.0004 at 512. The exact near-tie valid counts at $n = 32, 64, 128, 256, 512$ are $0, 1, 0, 7, 136$ out of 360,000, and every class–budget condition has zero false certificates. At 32 branches, set membership, serialized top-1, robust-95, and Contract-A-only have respectively $(\text{overlap}, \mathrm{FP}_{\mathrm{all}}, \mathrm{Miss}_{\mathrm{all}})$ equal to $(0.493, 0.090, 0.045)$, $(0.481, 0.083, 0.051)$, $(0.001, 0.000, 0.448)$, and $(0.469, 0.077, 0.057)$ on 12,000 draws.

D.4 DETERMINISTIC GATE CONTROL AND JOINT-LAW CHECKS

The joint-law check compares observed compatibility across product and common-shock families with matched action marginals (Table 12).

Table 12: Observed compatibility under joint outcome laws with matched action marginals. Within each regime, both families use the same $N = 4{,}000$ latent-law specifications. The two families share replay actions, costs, marginal success probabilities, and branch budgets per action. Branch outcomes are sampled separately for the two families. All rates are proportions of these specifications. OC denotes observed compatibility; FP and Miss denote false promotion and missed overlap, respectively.

	Product						Coupled					
	1 branch			32 branches			1 branch			32 branches		
Regime	OC	FP	Miss	OC	FP	Miss	OC	FP	Miss	OC	FP	Miss
Discordant	0.372	0.372	0.000	0.002	0.002	0.000	0.427	0.427	0.000	0.000	0.000	0.000
Near tie	0.514	0.269	0.101	0.472	0.255	0.128	0.624	0.311	0.033	0.470	0.220	0.095
Aligned	0.646	0.000	0.354	0.992	0.000	0.008	0.748	0.000	0.252	0.999	0.000	0.001

The known-by-construction control uses 36 cells, singleton replay labels cycling over three actions, five scenarios, and 20,000 Monte Carlo draws at one and eight branches per action (Table 13). Strict requires alignment, complete provenance, and no collision; $\tau = 0.95$, set intersection, and earliest serialization are baselines. In Table 13, the strict implementation has $0/5$ held-card label disagreements and $1/1$ aligned promotion; replay-only promotion disagrees on $5/5$ held cards.

Table 13: Known-by-construction gate control. Acceptance rates use the predeclared truth; a displayed zero is $0/20{,}000$.

		1 branch				8 branches			
Scenario	Truth	Strict	τ	Set	Top-1	Strict	τ	Set	Top-1
Discordant	hold	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Ambiguous	hold	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Provider	hold	0.000	0.000	0.000	0.000	0.000	0.000	0.656	0.553
Codebook	hold	0.000	0.000	0.000	0.000	0.000	0.000	0.654	0.551
Aligned	promote	0.000	0.000	0.000	0.000	0.466	0.622	0.651	0.549

Repeated projection can accept missing-provenance or collision cards, which is why the release gate retains both record and observability coordinates.

The joint-law check reuses 4,000 fixed-seed latent-law specifications under a product family and a common-shock family with probability 0.80 plus an independent component of probability 0.20. Per-action parameters lie in $(0, 1)$, so both families have full support; repeated vectors are independent while within-vector actions may be dependent. Matched marginals, release signatures, budgets, and replay identities are held fixed. The mechanism-separation audit uses 8,000 paired cells per condition: $\gamma_i = (\text{family}, \text{depth bin})$ is shared, the link condition uses $a^{\star}_{i,\text{link}} = g(\gamma_i)$, and the permuted condition uses a balanced fold-wise action permutation. Cost-adjusted utilities use $\lambda = 0.05$ and $C = (1, 1, 2)$, with margins 0.16 or 0.03 and budgets $1, 8, 32$. A leave-fold-out codebook predictor and set-valued, serialized, and robust-95 readouts use only the visible signature; their paired results are reported below.

D.5 Evidence allocation

The coordinate-local audit crosses four families (rank, path, filter, delegation) with four requirement regimes (preference-only, release-only, joint, resolved), using 250 states per cell, 4,000 states, and 64 paired seeds. An informative response replaces its target fiber by truth; an uninformative response leaves it unchanged. The same potential response is paired across policies.

Table 14: Budget feasibility for coordinate-targeted requests. TP, FP, and Miss use eligible, ineligible, and eligible denominators; Joint TP is restricted to eligible states with both coordinates outstanding. Cost and Rounds are means over all 4,000 states.

Rule	Budget	TP	FP	Miss	Joint TP	Cost	Rounds
Scalar projection	0	0.667	0.285	0.333	0.000	0.000	0.000
Typed-greedy	1	0.826	**0.000**	0.174	0.000	0.750	0.750
Typed-parallel	2	0.911	**0.000**	0.089	0.774	1.000	0.750

Two evidence channels are available, namely *outcome evidence* (repeat the task to observe new outcomes) and *record evidence* (collect provenance or observability information). Each channel is modeled as a paired Bernoulli draw in which an informative response reveals the true coordinate value and an uninformative response leaves the fiber unchanged. **Evidence update.** New evidence restricts the compatible relation to pairs consistent with the observation; a *separating* response collapses the corresponding coordinate image to a singleton.

Unresolved coordinates and request vocabulary. Three coordinates may be unresolved: bridge identity, preference overlap, and record eligibility. Each maps to a request type, respectively VERIFY (bridge), REPEAT (preference), and COLLECT (eligibility). Pending record checks (completeness σ_R, linkage σ_L, observability σ_O, each $\in \{\perp, 0, 1\}$) augment the request set: COLLECT is added for pending completeness or observability, and VERIFY for pending linkage.

Dispatcher logic. The dispatcher follows a three-way rule. If the full-set certificate holds, it returns CONTINUE$(I^{\dagger})$; if no failure is confirmed and requests remain, it returns REQUEST(Q_{all}); otherwise

it returns HOLD.

$$\pi^{\mathcal{I}}_{\mathfrak{C}}(\Gamma, \sigma, \mathsf{U}) = \begin{cases} \textsc{Continue}(I^{\dagger}), & e = 1, \\ \textsc{Request}(Q_{\mathrm{all}}), & e = 0,\ F = 0,\ Q_{\mathrm{all}} \neq \varnothing, \\ \textsc{Hold}, & \text{otherwise.} \end{cases} \tag{39}$$

The budgeted version substitutes HOLD when the total request cost exceeds the remaining budget B. A confirmed failure ($F = 1$) is terminal; a missing check is pending, not terminal. Truthful restriction can only shrink the target image, so Continue retains the original certificate, and a class with no positive target cannot acquire one. With initial budget B_0 and minimum request cost $c_{\min}$, at most $\lfloor B_0/c_{\min} \rfloor$ rounds can occur.

Coordinate-local guarantee. Under independent informative probabilities $q_{\mathrm{rep},c}$ and $q_{\mathrm{rec},c}$ for the two coordinates, the typed-parallel rule promotes an eligible case with probability equal to the product over its unresolved coordinates:

$$\Pr(\widehat{E}_c = 1) = \mathbf{1}\{\theta_c = (1,1)\}\, q_{\mathrm{rep},c}^{u_{\mathrm{rep},c}}\, q_{\mathrm{rec},c}^{u_{\mathrm{rec},c}}, \tag{40}$$

where $u_{\mathrm{rep},c}, u_{\mathrm{rec},c} \in \{0,1\}$ indicate unresolved coordinates. False promotion is exactly zero because an ineligible coordinate cannot be resolved to $\{1\}$ by truthful evidence.

Shared budget and coupled evidence. When the two evidence channels are positively correlated (joint informative probability $q^{\mathrm{rep}}q^{\mathrm{rec}} + \kappa$ with coupling $\kappa = 0.04$), the typed-parallel rule still achieves zero false promotion:

$$\Pr(\widehat{E}_c = 1 \mid \theta_c = (1,1)) = p_{U_c^{\mathrm{req}}}, \qquad \Pr(\widehat{E}_c = 1 \mid \theta_c \neq (1,1)) = 0, \tag{41}$$

where $p_{U_c^{\mathrm{req}}}$ is the joint informative probability over unresolved coordinates. When the budget cannot cover both requests, a one-round rule must hold because one fiber remains unresolved. A preference-only projection that omits the release coordinate has a positive false-promotion floor.

D.6 EVIDENCE ACQUISITION UNDER A SHARED BUDGET

The acquisition comparison uses a catalog containing local requests for each coordinate, alternative local requests with different costs and reliabilities, and a request covering both coordinates. Local-coordinate requests each outstanding coordinate separately; Joint-coordinate uses the joint request whenever any requirement remains. Exact typed uses the declared request law to maximize feasible coverage; Calibrated typed estimates coverage from separate response observations. These four typed rules share the same final promotion condition. Under the common budget cap $B = 1.50$, Exact typed reaches 93.2% eligible coverage, compared with 90.6% for Local-coordinate and 87.2% for Joint-coordinate (Table 15). Exact typed uses more of the permitted budget than Local-coordinate; relative to Joint-coordinate, it improves coverage while reducing average expenditure.

Table 15: Evidence acquisition under a shared request budget of $B = 1.50$. Coverage and false promotion use eligible and ineligible states as denominators, respectively; mean cost is averaged over all states. The audit covers 4,000 states with 64 paired response vectors per state for exact and fixed rules; calibrated selection uses 32 calibration and 32 independent evaluation vectors per state. Informative updates are truthful, coverage and false promotion are proportions, and evaluation cost excludes calibration acquisition.

Rule	Coverage	False promotion	Mean cost
Scalar projection	0.667	0.285	0.000
Joint-coordinate	0.872	0.000	0.825
Local-coordinate	0.906	0.000	0.750
Calibrated typed	0.924	0.000	0.856
Exact typed	0.932	0.000	0.813

Calibrated selection reaches 92.4% coverage on independent evaluation draws, compared with 93.3% for the exact selector on those same draws (Appendix Table 16). In a separate simulation with independent responses and a request law shared within each case family, the cost including calibration is 9.034 per evaluation case, above the fixed policies' costs (Appendix Table 17). The exact and calibrated selectors and the local and joint controls record zero false promotions, as expected from the

truthful-update model and shared final gate. Scalar projection, which omits the release requirement, promotes 28.5% of ineligible states despite issuing no request.

Among the 248 eligible cases with both coordinates outstanding, Exact typed and Joint-coordinate each reach 77.0% coverage, compared with 73.3% for Local-coordinate. The aggregate comparison also includes cases with only one outstanding coordinate, where local requests can avoid the cost of a joint request. The separate budget-feasibility comparison is reported in Appendix Table 14.

D.7 Evidence coverage and calibration

This audit uses a coordinate-local interface in which each binary coordinate fiber updates independently, with no cross-coordinate inference. Each request r in the catalog has cost $c_r > 0$, a coverage set S_r of coordinates it can resolve, and an informative indicator Z_r. A coordinate is *covered* when at least one informative request targeting it returns truth; uncovered fibers stay unchanged:

$$F_{c,u}^{A,Z} = \begin{cases} \{\theta_{c,u}\}, & \exists r \in A:\ u \in S_r,\ Z_r = 1, \\ V_{c,u}, & \text{otherwise.} \end{cases} \tag{42}$$

Theorem (budgeted evidence coverage). Under truthful updates and a one-round non-adaptive rule choosing requests A within budget B, the optimal true-positive rate and false-positive rate satisfy:

$$\mathrm{TP}_c^\star(B) = \max_{A \in \mathcal{F}_B(U_c)} p_c^{\mathrm{cov}}(A) \quad (c \text{ eligible}), \qquad \mathrm{FP}_c(B) = 0 \quad (c \text{ ineligible}), \tag{43}$$

where $p_c^{\mathrm{cov}}(A)$ is the probability that all unresolved coordinates receive informative responses. When the budget is below the minimum cover cost, $\mathrm{TP} = 0$.

Proof. An eligible case promotes exactly when every unresolved coordinate is covered by an informative response. Truthful updates cannot make an ineligible coordinate appear positive. Maximizing over feasible covers gives the optimum; randomization over covers yields a convex combination and cannot improve it. For two coordinates, the optimum is computed by exact enumeration. □

The coverage event is a conjunction over coordinates of a disjunction over requests: With independent indicators, inclusion–exclusion gives the exact coverage probability. For two outstanding coordinates, local requests have joint success $p_L = \Pr(Z_1 = Z_2 = 1)$ and a coupled request has success $p_J = \Pr(Z_{12} = 1)$. Compare only feasible covers: choose the local pair when $p_L > p_J$, the joint certificate when $p_J > p_L$, and hold when neither fits. If all three requests fit under independent indicators, coverage is

$$p_c^{\mathrm{cov}} = 1 - (1 - q_{12})(1 - q_1 q_2) = q_{12} + q_1 q_2 - q_1 q_2 q_{12} \geq q_1 q_2 q_{12}. \tag{44}$$

The audit enumerates all 32 subsets of five declared request types; at $B = 1.50$, expected and empirical coverage are 0.933 and 0.932.

Exact and calibrated selection. Under the same coordinate-local transition, let $\Pi_B(c)$ index a finite family of feasible request sets $A_\pi \in \mathcal{F}_B(U_c)$ fixed before calibration, and write $p_c(\pi) = p_c^{\mathrm{cov}}(A_\pi)$. If $\Pi_B(c) = \varnothing$, return Hold without issuing a request. Otherwise, exact selection uses the declared request law to choose $\pi_c^\star \in \arg\max_{\pi \in \Pi_B(c)} p_c(\pi)$. Enumerating all feasible sets recovers Eq. (43).

For a nonempty family, calibrated selection uses $m \geq 1$ independent response vectors $Z_{c,1}, \ldots, Z_{c,m}$ from the conditional request law given c and $\theta_c = (1, 1)$. Each vector evaluates every candidate set and may contain dependent responses. Such data can come from an independently labelled eligible calibration population. Unlabelled calibration instead requires the informative-response law to be independent of eligibility conditional on the observed context; otherwise it estimates a mixture, and the bound below need not hold. Our finite simulation satisfies this independence by construction, with response probabilities determined only by the observed family. Write $S_Z(A) = \bigcup_{r \in A: Z_r = 1} S_r$ for the coordinates revealed by informative responses, so the coverage event is $\mathsf{C}_c(A, Z) = \mathbf{1}\{U_c \subseteq S_Z(A)\}$. The empirical coverage and selected request index are

$$\widehat{p}_{m,c}(\pi) = \frac{1}{m} \sum_{i=1}^{m} \mathsf{C}_c(A_\pi, Z_{c,i}), \qquad \widehat{\pi}_c \in \arg\max_{\pi \in \Pi_B(c)} \widehat{p}_{m,c}(\pi). \tag{45}$$

The selector issues $A_{\widehat{\pi}_c}$, resolving ties by a fixed rule. On new responses from the same conditional law, its coverage loss satisfies, for $0 < \delta < 1$,

$$\Pr_{\mathrm{cal}}\left\{ p_c(\pi_c^\star) - p_c(\widehat{\pi}_c) \le \sqrt{\frac{2\log(2|\Pi_B(c)|/\delta)}{m}} \right\} \ge 1 - \delta. \tag{46}$$

The simulated initial fibers are $\{1\}$ or $\{0, 1\}$, so excluding known-negative states from the conditional objective changes none of the reported initial request choices. The audit uses 32 calibration seeds and 32 disjoint evaluation seeds; the largest candidate family contains seven request sets.

Table 16: Finite-sample calibration of evidence selection. The bound uses $\delta = 0.05$ and the largest nonempty candidate family; TP and FP use eligible and ineligible denominators.

Rule	m	$\lvert\Pi_B\rvert$	Bound	TP	Oracle TP	Regret	FP
Calibrated coverage	32	7	0.593	0.924	0.933	0.009	0.000

The separate-case transfer audit uses 1,000 calibration and 3,000 evaluation cases, with 248 calibration cases requiring no request and 752 supplying five response indicators. The estimator sees family and responses but no eligibility labels. The evaluation set has 1,675 eligible and 1,325 ineligible cases, each with 32 new response vectors. Fixed local and joint policies achieve eligible TP 0.908 and 0.873 at request costs 0.750 and 0.824; the declared-law oracle reaches 0.935 at cost 0.812. With eight calibration draws per case, the family estimator uses 1,504 observations per request and family and matches the oracle's subset on every evaluation case. The pooled estimator reaches the same TP at request cost 0.974. The 8-draw by-family total 9.034 is above both fixed policy costs, so the transfer result does not establish a cost advantage. The family-constant independent-response law is an assumption of this simulation; the split and response records permit direct recomputation.

Table 17: Calibration transfer to separate synthetic cases. TP is over 1,675 eligible cases; Request is mean selected-request cost over all 3,000 evaluation cases. Total adds full calibration cost divided by 3,000, with each response vector costing 4.100; the deployment cap is 1.500 and offline calibration is separately funded.

	1 draw			8 draws			32 draws		
Estimator	TP	Request	Total	TP	Request	Total	TP	Request	Total
Pooled	0.930	0.912	1.940	0.935	0.974	9.196	0.935	0.812	33.699
By family	0.932	0.902	1.930	0.935	0.812	9.034	0.935	0.812	33.699

D.8 Kernel-path and mechanism controls

The frozen selector is trained under U_{pol} and evaluated under an independently specified shifted replay law with the action contract and cost objective fixed (Table 18). It agrees with the primary selector on all 220 cells; shifted-law regret is 0.019 on 112 ID cells and 0.029 on 108 structural-OOD cells. The collision decomposition above separates the best common-action ceiling from the irreducible floor. Removing answer-, evidence-content-, and score-derived trace fields leaves ID agreement 1.000 and OOD agreement 0.859, with regret 0.000 and 0.001, respectively (Table 19).

Table 18: Evaluation of a frozen selector under the shifted replay law. N counts checkpoint cells; Trace regret is shifted-law oracle utility minus Trace-only utility.

Split	N	Trace U_{pol}	Oracle U_{pol}	Trace regret
ID	112	0.735	0.753	0.019
Structural OOD	108	0.691	0.720	0.029

Table 19: Outcome-blind replay control. Answer-, evidence-content-, and score-derived fields are removed before refitting; the reference and blind selector use the same labels and held-out cells.

Split	Agreement	Same	Regret
ID	1.000	1.000	0.000
Structural OOD	0.859	1.000	0.001

D.9 Matched-signature mechanism separation

The mechanism audit retains the public signature and action-wise replay marginals while preserving or permuting their link to the latent winner (Table 20). Each condition contains 8,000 paired cells. Codebook accuracy is visible-signature winner accuracy; one-shot FP is the one-branch set false-promotion rate; robust FP and miss use the 32-branch robust-95 readout. The denominator is 8,000 paired cells per row. The visible link can be perfectly predictable while one-shot preference transfer still has false promotions; the 32-branch robust readout has zero false promotions and misses about one third of positives in both conditions.

Table 20: Matched-signature mechanism separation.

Condition	Codebook acc.	One-shot FP	Robust FP	Robust miss
Link permuted	0.333	0.297	**0.000**	0.333
Link preserved	**1.000**	0.411	**0.000**	0.332

E Supplementary decision and relation audits

E.1 Source-key mechanism witness

The source-key fixture holds the public signature and release status fixed while a key separates two eligible singleton action-set images (Table 21). Before a key response, the current unresolved set contains both bridge and preference coordinates, so the typed dispatcher returns VERIFY+REPEAT under the existing request vocabulary. A key response can alone collapse the action-set image, but the pre-response dispatcher reports all unresolved evidence operations and is not a minimum-cover optimizer.

Table 21: A separating source coordinate changes the singleton decision. The key-absent row reports the full pre-response request set.

Key view	Action-set image	Next evidence	Disposition
Key absent	$\{(\{\text{RETRY}\},1),(\{\text{VERIFY}\},1)\}$	VERIFY+REPEAT	HOLD
Key ℓ_r	$\{(\{\text{RETRY}\},1)\}$	None	PREFERENCE-CERT
Key ℓ_v	$\{(\{\text{VERIFY}\},1)\}$	None	PREFERENCE-CERT

E.2 Paired release contrasts

All release rules use the same 2,000 independent outcome vectors per regime and the same supplied records. Linkage is fixed at one in this source abstraction. Let $h_i = P_{\mathrm{rec},i} O_{\kappa,i}$ and let r_s be the replay action. The overlap estimand is

$$o_i = h_i \prod_{s \in \mathcal{D}_i} \mathbf{1}\{r_s \in A^\star(P_s)\}, \tag{47}$$

and the unique-winner estimand is

$$u_i = h_i \prod_{s \in \mathcal{D}_i} \mathbf{1}\{A^\star(P_s) = \{r_s\}\}. \tag{48}$$

Thus the eligible regime has $(u_i, o_i) = (1, 1)$, the tied regime $(0, 1)$, and discordant, provider-gap, and observability-gap regimes $(0, 0)$. Paired differences use the same draws and equal-tailed Clopper–Pearson intervals (Clopper & Pearson, 1934) on discordant-release counts.

E.3 Release targets on the same outcome draws

We evaluate the same observed release decisions under two requirements. The overlap requirement accepts a replay action among the expected maximizers in every cell; the stricter unique-winner requirement excludes ties. Both require the supplied record checks to pass. The generator supplies the expected utilities for these evaluator labels. Its regime named near tie has exactly equal expected utilities for the replay action and a competitor.

Set membership admits a nonempty observed intersection, serialized top-1 uses a fixed action order, and Full gate requires each observed best set to equal the replay reference. Status-aware variants of the first two rules use the same record checks as Full gate. These rules test observed outcomes against the two preference requirements; the candidate certificate is defined separately in Eq. (12).

Table 22: Post hoc reanalysis of the same 10,000 releases at 32 branches per action: 2,000 per eligible, tied, discordant, provider-gap, and observability-gap regime. Overlap and unique-winner labels have 4,000 and 2,000 positives, respectively, and include the same supplied-status checks. Correct and Error count admitted positive and negative releases; Miss counts withheld positives.

	Overlap			Unique winner		
Rule	Correct	Error	Miss	Correct	Error	Miss
Replay-only	4000	6000	0	2000	8000	0
Set membership	2100	3847	1900	1924	4023	76
Serialized top-1	2059	3830	1941	1919	3970	81
Status-aware set	2100	0	1900	1924	176	76
Status-aware top-1	2059	0	1941	1919	140	81
Status-aware Robust-95	0	0	4000	0	0	2000
A-only	2009	3820	1991	1911	3918	89
Full gate	2009	0	1991	1911	98	89

For the overlap target, Status-aware set admits more positives than Full gate with no observed errors in either rule (Table 22). Requiring a unique winner changes the tied admissions from correct to incorrect. Full gate then reduces near-tie error by 3.90 percentage points relative to Status-aware set, with a 0.65-point loss in eligible promotion. The pointwise paired intervals exclude zero; Per-regime rates and contrasts are reported in Appendix Table 23.

The record checks account for zero errors in the provider and observability gaps across all status-aware rules. Removing them gives A-only, whose gap error rates exceed 95%. At this budget, Status-aware Robust-95 makes no errors and admits no eligible releases.

E.4 Replicate stability and outcome composition

The paired local-replicate audit contains 144 matched action branches and 48 checkpoint cells. Replicate strict success agrees on 134/144 (0.931), semantic success on 134/144 (0.931), and schema on 140/144 (0.972). The strict contingency has 49 joint failures, six failures unique to replicate 0, four failures unique to replicate 1, and 85 joint successes. Across cells, the pooled relation is direct 2/48 (0.042), overlap 22/48 (0.458), and discordant 24/48 (0.500); the discordant cells include five with partial strict success (Table 25).

Table 23: Unique-winner release operating characteristics with 32 branches per action and 2,000 independent releases per regime. Rates and paired differences are proportions. T_{elig} denotes eligible promotion; E_{tie}, E_{prov}, and E_{obs} denote error rates in the tied, provider-gap, and observability-gap regimes. ΔT and ΔE subtract each row from Full gate for eligible promotion and tied-regime error. Pointwise exact 95% interval endpoints are rounded outward.

	Observed rates				ΔT			ΔE		
Rule	T_{elig}	E_{tie}	E_{prov}	E_{obs}	Estimate	Lower	Upper	Estimate	Lower	Upper
Replay-only	1.000	1.000	1.000	1.000	-0.045	-0.055	-0.035	-0.951	-0.961	-0.940
Set membership	0.962	0.088	0.959	0.965	-0.007	-0.012	-0.003	-0.039	-0.049	-0.030
Serialized top-1	0.960	0.070	0.953	0.962	-0.004	-0.008	-0.001	-0.021	-0.029	-0.015
Status-aware set	0.962	0.088	0.000	0.000	-0.007	-0.012	-0.003	-0.039	-0.049	-0.030
Status-aware top-1	0.960	0.070	0.000	0.000	-0.004	-0.008	-0.001	-0.021	-0.029	-0.015
Status-aware Robust-95	0.000	0.000	0.000	0.000	0.956	0.945	0.965	0.049	0.039	0.060
A-only	0.956	0.049	0.952	0.959	0.000	0.000	0.000	0.000	0.000	0.000
Full gate	0.956	0.049	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000

Table 24: Paired local-replicate strict-success contingency on 144 matched branches.

	Replicate 1 fail	Replicate 1 success	Total
Replicate 0 fail	49	4	53
Replicate 0 success	6	85	91
Total	55	89	144

Table 25: One-shot endpoint relation crossed with strict action quality. Counts use all three actions per cell.

Relation	Cells	Fraction	All success	Partial	All fail
Direct	1	0.021	0	1	0
Overlap	23	0.479	15	0	8
Discordant	24	0.500	13	4	7

F Execution archives and record reconstruction

F.1 Artifact Validity and Action Agreement

Each model's action branches cover four fault scenarios per workflow prefix. We evaluate semantic correctness and public-score agreement, then compare the cost-adjusted action sets with the replay reference (Section 3). Retry, verify, and replacement have normalized costs 1, 1, and 2, respectively, with $\lambda = 0.05$. The replay reference is evaluated separately in simulation, including fixed-policy comparisons and a shifted replay law (Appendices B.8 and D.8).

Luna and Sol succeed with every action, yet their cost-adjusted sets are disjoint from the replay reference in half the cells (Table 26). Equal success favors the cheaper retry and verify actions; discordance occurs wherever replay selects replacement. Pooling two executions leaves every GPT-5.6 cell's best set unchanged. Record eligibility follows the source-specific conditions in Appendix B.2.

For Qwen3-8B, matching sets contain both all-failure and all-success cells (Appendix Table 24). Cost adjustment favors retry and verify when all actions fail, so stable observed sets still require an outcome-quality check before they support a recovery claim.

The version-locked provider pool uses one DeepSeek API endpoint (DeepSeek, 2026), 12 prefixes (six filter, three rank, three path), and protocol `live-recovery-calibration/2` with model `deepseek-chat`, temperature 0, and a 240-token cap. Nine score-consistent prefixes enter the strict audit (six filter, three rank). Four fault mechanisms and three recovery actions give 36 cells and 108 one-branch-per-action rows. Each cell retains deterministic `continue`; retry, verify, and replace use fresh prompt branches. Provider artifacts retain normalized outputs and token fields but omit

Table 26: Retained executions on 48 checkpoint–fault cells per source. Each source provides two executions per action (288 branches): $b = 1$ uses the first execution, and $b = 2$ pools both. Strict_{144} is the fraction of the 144 $b = 1$ branches satisfying the strict predicate. D, O, and X count cells with equal singleton sets, other nonempty overlaps, and disjoint sets, respectively.

		$b = 1$				$b = 2$			
Source	Strict_{144}	D	O	X	Mean $\lvert\widehat{L}_s^{\text{obs}}\rvert$	D	O	X	Mean $\lvert\widehat{L}_s^{\text{obs}}\rvert$
Luna	1.000	0	24	24	2.000	0	24	24	2.000
Terra	0.958	1	22	25	1.896	1	22	25	1.896
Sol	1.000	0	24	24	2.000	0	24	24	2.000
Qwen3-8B	0.632	1	23	24	1.917	2	22	24	1.875

request identifiers, complete prompts, and source-complete response records, so $P^{(1)}_{\text{link}} = P^{(1)}_{\text{fields}} = 0$ (Table 27).

Table 27: Score-scrubbed local-model materialization on 12 prefixes. Branches/action is the action-column denominator (24 per fault, 96 overall); Schema uses all action rows.

	Sample	Strict recovery			Structure
Fault	Branches/action	Retry	Verify	Replace	Schema
Transient tool failure	24	0.667	0.667	0.667	0.986
Semantic error	24	0.583	0.583	0.583	0.944
Stale message	24	0.583	0.583	0.583	0.931
Delegation mismatch	24	0.708	0.583	0.708	0.944
All branches	96	0.635	0.604	0.635	0.951

The terminal release audit on 48 source-complete cells gives replay-only release on 48/48, set membership on 24/48, serialized top-1 on 14/48, record-erased gate hold on 48/48, and exact keyed gate decisions of release 2, hold 22, reject 24 with agreement 1.000 (Table 28). Extra authorization is 0 for the exact keyed gate and 0.958 for replay-only.

Table 28: Terminal release decisions on 48 source-complete cells. Agreement is relative to the exact keyed observed-compatibility gate.

Readout	Release	Hold	Reject	Agreement	Extra auth.
Replay-only	48	0	0	0.042	0.958
Set membership	24	0	24	0.542	0.458
Serialized top-1	14	0	34	0.542	0.250
Record-erased gate	0	48	0	0.458	0.000
Exact keyed gate	2	22	24	1.000	0.000

F.2 Supplementary materialization controls

On 16 held-out non-injected-error cells under semantic cleanliness, blind retry succeeds 0/16, always replace 3/16 (0.188), always verify 16/16 (1.000), and trace-only with verify fallback 16/16 (1.000) (Table 29). Clopper–Pearson 95% intervals (Clopper & Pearson, 1934) are $(0.000, 0.206)$ for blind retry and $(0.040, 0.456)$ for always replace; the two perfect rows have lower endpoint 0.794 and upper endpoint 1.000. The evaluator-key verification row is privileged.

The score-scrubbed Qwen3-8B archive uses bfloat16 weights with 36 layers, hidden size 4096, 32 attention heads, 8 key–value heads, maximum context 40,960, temperature 0.7, sampling, and a 128-token cap. Seeds produce $12 \times 4 \times 3 \times 2 = 288$ branches. A keyed join over `task_id`, `fault`, `action`, and `replicate` separates runtime metadata from embedded configuration and supports direct regeneration of relation tables. The local archive has complete keyed receipts, so $P^{\text{loc}}_{\text{rec}} = P^{\text{loc}}_{\text{link}} = 1$; 14 schema-invalid completions remain decoder outcomes. The full unmasked trace has $O^{\text{loc}}_{\kappa} = 1$, while the five-key perturbation supplies the $O_{\kappa} = 0$ control.

Table 29: Held-out non-injected error fallback under semantic cleanliness. Counts and rates use 16 cells per row; interval endpoints are rounded outward.

	Sample	Success		95% interval	
Action	Cells	Count	Rate	Lower	Upper
Blind retry	16	0	0.000	0.000	0.206
Always replace	16	3	0.188	0.040	0.456
Always verify	16	16	1.000	0.794	1.000
Trace-only with verify fallback	16	16	1.000	0.794	1.000

Table 30: One-shot partial-identification audit on strict live cells. Counts and rates use 36 cells unless stated otherwise.

Quantity	Cells	Count	Rate
Observed live set is singleton	36	5	0.139
Observed live set is tied	36	31	0.861
All actions confidence-compatible	36	36	1.000
Cell has a robust expected winner	36	0	0.000
Expected-transfer rate lower endpoint	1	0	0.000
Expected-transfer rate upper endpoint	1	1	1.000

The one-shot endpoint set is therefore usually tied and every cell is confidence-compatible (Table 30). No cell has a robust expected winner, and the support-compatible aggregate expected-transfer rate spans $[0, 1]$ at this sample size.

F.3 ARCHIVE-COMPLETE LOCAL RELEASE GATE

The archive-complete gate applies the same release coordinates after retaining one or two branches per action. Setting only $P^{\mathrm{loc}}_{\mathrm{rec}}$ or $O^{\mathrm{loc}}_{\kappa}$ to zero sends every cell to its corresponding hold. On 16 held-out cells whose initial records fail source-complete schema, the dispatcher requests verification; all 16 generated artifacts remain schema-incomplete, producing 16 terminal holds and zero unsafe promotions. Branch-only completion rates for verify, retry, and replace are 1.000, 0.000, and 0.188, with mean costs 1.000, 1.000, and 2.000.

F.4 SEQUENTIAL EVIDENCE-REQUEST CLOSURE

The closure audit starts with incomplete records and requests verification before reading the response. All 16 generated artifacts remain schema-incomplete, so the gate returns 16 terminal holds. The source-key fixture starts with both bridge and preference coordinates unresolved; the dispatcher requests VERIFY+REPEAT under the existing vocabulary, a separating key yields a singleton action-set image and continuation, and an uninformative response remains HOLD-AMBIGUITY.

F.5 KEYED-BRIDGE SENSITIVITY

Within each prefix, shared and action-wise independent fault permutations erase the fault key while preserving each action-specific multiset of four fault-indexed outcomes. Across 10,000 deterministic permutations (seed 20260903), decision agreement with the exact keyed gate has empirical intervals $(0.854, 0.979)$ for the shared family and $(0.813, 0.938)$ for action-wise independent permutations; false-promotion intervals are $(0.000, 0.063)$ and $(0.000, 0.104)$. These bridge intervals are empirical quantiles over the fixed archive, not uncertainty about newly generated endpoint outcomes.

F.6 EMPIRICAL JOINT-IMAGE AUDIT

For a prefix with four fault coordinates, an assignment $\beta = (\beta_a)_{a \in \mathcal{A}_{\mathrm{rec}}}$ is a per-action permutation; both retained replicates use the same assignment. The key family contains identity only, the shared family one common permutation, and the independent family all action-wise tuples. For n retained branches,

$$\widehat{y}^{\beta,n}_{c,a} = n^{-1} \sum_{r=1}^{n} s^{\beta}_{c,a,r} - \lambda C(a), \qquad \widehat{L}^{\beta,n}_{c} = \arg\max_{a \in \mathcal{A}_{\mathrm{rec}}} \widehat{y}^{\beta,n}_{c,a}. \tag{49}$$

With replay set R_c, retain the action-set image before binary projection:

$$\widehat{\Gamma}^{\mathcal{J},n}_{c,\mathcal{I}} = \{(R_c \cap \widehat{L}^{\beta,n}_{c}, 1) : \beta \in \mathcal{J}\}. \tag{50}$$

The literal receipt coordinate is one for every retained branch and does not assert provider linkage or masked-probe observability. Exact keyed images are singleton; control families contain identity, so a keyed positive pair remains in the robust union while extra assignments can remove singleton status. Applying the same gate gives the bridge table above. The two-replicate audit has strict and semantic agreement of 134/144 pairs and schema agreement of 140/144 pairs. Pooling one execution per action gives direct, overlap, and discordant relations of 1/48, 23/48, and 24/48; pooling both gives 2/48, 22/48, and 24/48. A semantic-clean sensitivity on the same frozen source pool gives 22/48 materialized cells (0.458), with 41 live ties; by family, filter, rank, and path materialization rates are 0.458, 0.417, and 0.500 on 24, 12, and 12 cells, respectively. This sensitivity retains the prior-artifact score in provider prompts and changes only the evaluator outcome predicate.

F.7 AGREEMENT ACROSS RETAINED EXECUTION SOURCES

Each source contributes 48 prefix–fault cells, with two retained executions under each of the three recovery instructions (Table 31). The comparison retains all maximizers, including ties. Qwen3-8B has matching multi-action sets in both all-failure and all-success cells, whereas the GPT sources are almost entirely all-success.

Qwen3-8B uses a score-scrubbed prompt and a different 12-prefix panel from the GPT-5.6 sources, whose prompts retain the prior-artifact score. The common prefix intersection contains nine tasks. The intervals below are source-internal descriptive 95% percentile intervals. Each bootstrap resample

draws the 12 workflow prefixes with replacement and retains all four fault conditions and all branches within a sampled prefix; the analysis uses 10,000 resamples with seed 742031. "Cost-adjusted strict" uses the complete maximizer set of strict success minus $0.05C(a)$ with $C = (1, 1, 2)$; "Raw strict" uses the unadjusted strict-success maximizer set. Selected success is the uniform-tie expected strict success on the other execution, averaged in both directions.

Table 31: Retained-source diagnostics on 48 cells and 288 branches per source. All-failed and All-passed count equal-set cells with all six branches failing or passing. Lower and Upper delimit descriptive 95% bootstrap intervals; resampling preserves the 12 prefix clusters.

		Equal-set cells		Exact agreement			Branch success			Selected success		
Source	Rule	All-failed	All-passed	Estimate	Lower	Upper	Estimate	Lower	Upper	Estimate	Lower	Upper
Qwen3-8B	Adjusted	15	26	0.875	0.750	0.979	0.625	0.368	0.858	0.620	0.365	0.870
Terra	Adjusted	0	42	0.896	0.771	1.000	0.979	0.955	1.000	0.974	0.943	1.000
Luna	Adjusted	0	48	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
Sol	Adjusted	0	48	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
Qwen3-8B	Raw	15	26	0.875	0.771	0.979	0.625	0.378	0.865	0.625	0.377	0.875
Terra	Raw	0	42	0.875	0.729	1.000	0.979	0.955	1.000	0.979	0.955	1.000
Luna	Raw	0	48	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
Sol	Raw	0	48	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

F.8 Retrieval of retained checkpoint bindings

The retained Qwen3-8B, Luna, Terra, and Sol archives each contribute 48 cells within 12 prefixes and two executions per action. The selector sees four groups per action, two strict outcomes and a cost-adjusted mean per group, plus the replay reference; model, checkpoint, provenance, prompts, fault labels, and the canonical binding remain evaluator-side. Each cell starts with $4^3 = 64$ action-wise assignments. A lookup retrieves the stored group for one action and restricts the candidate class. Fixed order, score range, target-directed minimax, and EC^2 use the same masked view and stop on a singleton overlap image or budget exhaustion (Table 32). EC^2 uses a uniform prior over 64 bindings and edges between hypotheses with different overlap targets; costs count record lookups only.

Table 32: Checkpoint binding retrieval on 48 cells within 12 prefixes per model. Resolved counts singleton overlap images, including the empty set; Positive counts resolved nonempty targets after two queries.

	Luna			Terra			Sol		
	Resolved		Positive	Resolved		Positive	Resolved		Positive
Policy	1	2	2	1	2	2	1	2	2
Fixed order	48	48	24	45	48	23	48	48	24
Score range	48	48	24	45	48	23	48	48	24
EC^2	48	48	24	47	48	23	48	48	24
Target-directed	48	48	24	47	48	23	48	48	24

The Luna and Sol scores are constant across candidate groups, so the assignment cannot change their target. In the Qwen3-8B archive, fixed order resolves 40 and 44 cells at budgets one and two, while EC^2 resolves 44 and 47; the corresponding positive counts at budget two are 21 and 23 (Table 33). All policies resolve the full 48 cells by budget three. A separate full-set readout is not substituted for $R_c \cap \widehat{L}_c$.

Candidate-class omission. A shared-slot stress class retains four of the 64 assignments and contains the canonical binding in 14/192 cells. With one truthful lookup, all three selectors make the same 3 wrong stops, reporting two nonempty targets and one empty target. In a second control, twelve classes each omit one action–slot value, excluding each canonical binding three times. At budget three, EC^2 produces one incorrect empty target and 27 empty images over 2,304 dependent cell–class evaluations; fixed order and minimax produce no incorrect targets and 50 and 47 empty images,

Table 33: Additional retrieval comparison on the Qwen3-8B archive, with 48 cells within 12 prefixes. Resolved counts singleton overlap images, including the empty set; Positive counts resolved nonempty targets after two queries.

	Resolved		Positive
Policy	1	2	2
Fixed order	40	44	21
Score range	41	43	20
EC^2	44	47	23
Target-directed	42	44	21

respectively. Empty images trigger Hold. These are candidate-class consequences, not evidence that a retrieval policy has identified an unknown outcome law.

F.9 JOINT OBSERVED TARGETS AND LOCAL RELEASE ELIGIBILITY

Table 34: Retrieval on all 48 Qwen3-8B cells. Image counts identified complete $(I_s^{\mathrm{obs}}, D_\kappa)$ values; Decision counts identified Continue or ineligible dispositions. An unresolved Hold is not an identified disposition. Headers 1 and 2 are per-cell query budgets; Queries totals lookups at budget two. Initially, 32 images and 34 dispositions are identified.

		Image		Decision		Queries
Query rule	Stop target	1	2	1	2	2
Fixed order	Image	36	44	39	46	28
Minimax, joint	Image	41	44	43	46	23
EC^2, eligibility first	Image	40	45	44	47	24
EC^2, joint	Image	41	45	44	47	23
EC^2, joint	Decision	39	42	44	47	18
EC^2, eligibility first	Decision	38	41	44	47	18

The joint-target extension uses all 48 local Qwen3-8B cells and 288 retained branches (Table 34). Local receipt and linkage checks are fixed at one, with $O_\kappa^{\mathrm{loc}} = 1$. For assignment β, the target is $(I_c^{\mathrm{obs}}(\beta), D_c^{\mathrm{loc}}(\beta))$, where $D_c^{\mathrm{loc}} = 1$ means $\widehat{L}_c^{\beta,2} = R_c$. All 64 assignments are initially possible; the canonical tuple is included and truthful replies preserve it. Enumerating all eight reply subsets gives 384 covered cell–history pairs. Initially, 16 joint images and 14 eligibility coordinates are unresolved. Products of coordinate images add unsupported pairs in 12 cells; imposing $D_c^{\mathrm{loc}} = 1 \Rightarrow I_c^{\mathrm{obs}} = R_c$ leaves six inexact images. After one and two replies, the stronger joint construction remains inexact in 9/144 and 4/144 histories.

The main joint-retrieval table shows that eligibility-first and joint EC^2 issue 24 and 23 lookups at budget two and identify 45/48 full images and 47/48 eligibility values. The decision-stopping variants use 18 lookups and identify 42/48 and 41/48 full images while identifying 47/48 dispositions. The canonical Qwen3-8B archive contains two positive cells at this budget. Sequential controls and joint EC^2 use the same uniform prior and action-order tie breaks; the exact-coordinate representation preserves the association between the action-membership vector and eligibility bit. A richer full-image target therefore incurs additional resolution work without establishing a representation advantage.

A complete lookup example. At two branches per action, the Qwen3-8B archive has two positive canonical cells: task `id-1004` under the semantic-error fault with verification as the observed and replay winner, and task `id-1004` under the stale-message fault with replacement as both winners. The remaining 46 cells are known ineligible or unresolved under the stated target. These outcomes illustrate that the lookup target is the joint action–eligibility conclusion; it does not estimate an outcome law.

Post hoc target-stop audit. This comparison reuses the same 48 cells, 64-assignment query view, and truthful lookups; it makes no model calls. It distinguishes an unresolved image, which remains operational HOLD, from an identified ineligible disposition. At budget two, the full-target joint rule used 23 lookups, identifying 45 full images and 47 release dispositions; the joint disposition rule used 18 lookups, identifying 42 full images and 47 dispositions, while eligibility-first disposition stopping used 18 lookups and identified 41 full images and 47 dispositions. At budget three, the full-target rule used 26 lookups and identified all 48 full images; the two disposition-stopping rules used 19 lookups each and identified 43 and 42 full images, respectively, while identifying all 48 dispositions. Each disposition rule yielded two correct Continue decisions, 46 known-ineligible dispositions, and zero incorrect dispositions. This same-archive comparison shows that a richer target requires more resolution work when the release decision can be certified earlier.

G PROSPECTIVE REPEATED-BRANCH STUDY

Original Terra reference panel. On 24 constrained planning checkpoints, one draw per action gives held-out strict success 0.699 versus 0.771 for fixed verification (Table 37). More selection draws do not close this gap and increase call cost. Public best-set agreement between selection and evaluation

blocks is 0.318 at $k = 1$ and falls to 0.204 and 0.167 at $k = 2, 4$; strict-set agreement is 0.378, 0.262, 0.250 (Table 36). The primary contrast is selected minus fixed verification on held-out strict success.

The panel contains 12 tasks with 12 items and 12 with 16 items. Items have integer benefit, cost, time, prerequisite, conflicting-pair, budget, and subset-size fields. A deterministic greedy pass constructs a feasible suboptimal current plan; mean current benefit divided by exact optimum is 0.650. Retry, verify, and replace receive the same public task and current plan, and replacement uses the same model and permissions.

GPT-5.6 Terra runs at medium reasoning effort in fresh ephemeral contexts with tools prohibited. The 576-call collection uses eight draws for each of three actions at every checkpoint; draws 0–3 are selection and 4–7 evaluation under a fixed interleaved schedule with four concurrent requests. Task identifiers, draw indices, block assignments, and optimum values are absent from prompts. Two capacity failures succeed on their permitted second attempt. Of 576 completed outputs, 557 have feasible selected subsets and 522 have matching totals; these checks can fail on overlapping rows. No tool invocation occurs. Fresh contexts do not imply iid provider sampling. The Luna and Sol extensions use the same frozen prompts and schedule and are summarized below.

The public score is recomputed benefit when the subset is feasible, reported totals agree, and no tool is invoked; strict success additionally requires equality to the offline exact optimum. For each $k \in \{1, 2, 4\}$, all size-k subsets of the four selection indices produce an argmax action set; uniform tie-breaking defines the selected mixture, which is evaluated on four held-out draws per action. The six-task, 36-call development panel preselected fixed verification. Fixed retry, fixed replacement, and uniform action selection are additional baselines. A fixed action costs one final call; selection costs $3k + 1$ calls. Performance and tokens are estimated from held-out draws rather than an online deployment. The complete panel uses 12,376,363 input tokens (11,631,104 cached) and 592,590 output tokens; cached tokens remain in the input total.

The checkpoint is the statistical unit. Descriptive 95% percentile intervals use 10,000 bootstrap resamples within the two task-size strata and preserve complete policy comparisons; repeated draws and overlapping subsets are not independent tasks. Exact set agreement is tie-sensitive and need not increase with k. Across 192 three-action draw vectors, 20 have no strictly successful action and 90 have all three, so tied sets occur at both extremes. At 12 items, selected-minus-verification strict-success differences are $-0.131, -0.133, -0.104$ for $k = 1, 2, 4$; at 16 items they are $-0.012, -0.019, -0.017$. The deficit is concentrated in the smaller-task stratum. This panel tests branch replication and action selection on planning checkpoints (Table 35).

Table 35: Held-out strict success and normalized public benefit on the original Terra planning cohort. Calls include selection and one final action; each metric has a descriptive 95% checkpoint-bootstrap interval.

		Strict success			Normalized benefit		
Policy	Calls	Estimate	Lower	Upper	Estimate	Lower	Upper
Fixed retry	1	0.677	0.552	0.802	0.848	0.763	0.922
Fixed replacement	1	0.688	0.562	0.802	0.887	0.825	0.945
Uniform	1	0.712	0.608	0.806	0.882	0.841	0.920
Fixed verification	1	0.771	0.656	0.875	0.912	0.853	0.963
Selected, $k = 1$	4	0.699	0.591	0.798	0.872	0.824	0.913
Selected, $k = 2$	7	0.695	0.583	0.797	0.870	0.812	0.917
Selected, $k = 4$	13	0.710	0.594	0.816	0.880	0.807	0.939

Table 36: Exact best-set agreement across selection and evaluation blocks on the original Terra planning cohort. Public uses the observable score; Strict uses evaluator-only optimality labels.

	Public			Strict		
k	Estimate	Lower	Upper	Estimate	Lower	Upper
1	0.318	0.216	0.424	0.378	0.268	0.492
2	0.204	0.119	0.303	0.262	0.161	0.377
4	0.167	0.042	0.333	0.250	0.083	0.417

Table 37: Selected-policy contrast and expected input-plus-output tokens per decision on the original Terra planning cohort. Differences are relative to fixed verification; intervals are paired strict-success differences.

			Strict-success difference		
k	Calls	Mean tokens	Estimate	Lower	Upper
1	4	89997.928	-0.072	-0.142	-0.001
2	7	157445.139	-0.076	-0.156	0.001
4	13	292310.155	-0.061	-0.160	0.035

G.1 Independent planning replication

We fixed a new 24-checkpoint panel, prompts, request order, scoring, two execution blocks, and the primary $A_1 - A_4$ contrast before collection. The complete cross of 12 seeds (98231 through 99342 in increments of 101) and task sizes 12 and 16 uses generator seed seed $+ n$, avoiding nested random streams across sizes. All 576 completed Terra medium executions are retained. The public task and current artifact are shared across recovery instructions; evaluator optima, task identifiers, seeds, and block indices are excluded from prompts. Calls use fresh contexts and a shuffled order, which reduces order confounding but does not establish independent provider sampling. Transport failures permit one retry; completed invalid or suboptimal outputs are never retried. The scientific artifact retains the protocol, call order and timing, usage, outputs, and offline scoring inputs.

The primary contrast is 0.320 (95% paired interval [0.240, 0.406]). Intervals resample the 24 checkpoints within size strata 10,000 times with seed 98419. The difference is computed inside each paired resample; intervals for separate means are not subtracted. This implementation correction was made before the completed panel was analyzed. Results for all three budgets are reported in Table 38. Matching multi-action contributions are 0.391, 0.205, and 0.083; matching-singleton contributions are 0.013, 0.034, and 0.000. Mean set sizes are 2.411, 2.122, and 1.833. These decompositions are descriptive.

Table 38: Independent replication on 24 new planning checkpoints. Agreement uses public-score sets; Success evaluates uniform selection from those sets on the held-out block. Each metric has a descriptive 95% checkpoint-bootstrap interval.

	Agreement			Success		
k	Estimate	Lower	Upper	Estimate	Lower	Upper
1	0.404	0.307	0.503	0.765	0.677	0.844
2	0.238	0.148	0.341	0.754	0.665	0.835
4	0.083	0.000	0.208	0.762	0.665	0.847

Common-action overlap is 0.940, 0.863, and 0.750 at $k = 1, 2, 4$. Jaccard similarity is 0.693, 0.557, and 0.375 at $k = 1, 2, 4$. Fixed verification achieves 0.771 strict success. Selected-minus-verification differences are -0.006 [-0.049, 0.037], -0.016 [-0.064, 0.031], and -0.009 [-0.054, 0.036] in the same budget order. Each interval contains zero, so this panel supports the agreement decline without establishing a selection deficit. Selection costs remain $3k + 1$ calls versus one for a fixed action. Collection used 12,401,744 input tokens (11,590,656 cached) and 599,798 output tokens; cached tokens are included in the input total.

G.2 GPT-5.6 family extension

The family extension reuses both frozen cohorts, public task files, current plans, action instructions, block assignments, and scoring code. Terra contributes the two completed reference panels above; Luna and Sol were collected later with the same eight-draw schedule, fresh contexts, medium reasoning effort, and tools prohibited. Each model–cohort panel contains 576 responses, giving 3,456 responses across six panels. Terra and the later runs are therefore matched on tasks and prompts, while their collection windows are distinct. The panel-level counts of feasible and strict rows are

retained in the run manifest; they are diagnostics for the response contract and are not additional statistical units.

Table 39: GPT-5.6 family exact public-score set agreement. Each entry is Estimate, Lower, or Upper from 10,000 checkpoint-bootstrap resamples stratified by task size. The six panels contain 24 checkpoints each.

		$k = 1$			$k = 2$			$k = 4$		
Cohort	Model	Estimate	Lower	Upper	Estimate	Lower	Upper	Estimate	Lower	Upper
Study	Terra	0.318	0.221	0.424	0.204	0.122	0.306	0.167	0.042	0.333
Study	Luna	0.234	0.164	0.312	0.150	0.109	0.199	0.083	0.000	0.208
Study	Sol	0.424	0.341	0.503	0.331	0.242	0.428	0.167	0.042	0.333
Replication	Terra	0.404	0.307	0.503	0.238	0.148	0.343	0.083	0.000	0.208
Replication	Luna	0.299	0.224	0.388	0.212	0.139	0.303	0.250	0.083	0.417
Replication	Sol	0.497	0.427	0.570	0.422	0.326	0.520	0.458	0.292	0.625

Table 40: GPT-5.6 family selected strict success on held-out draws. Each entry is Estimate, Lower, or Upper from the same checkpoint-bootstrap procedure. Fixed verification is a one-call baseline within the corresponding model–cohort panel.

		$k = 1$			$k = 2$			$k = 4$			Fixed		
Cohort	Model	Estimate	Lower	Upper	Estimate	Lower	Upper	Estimate	Lower	Upper	Estimate	Lower	Upper
Study	Terra	0.699	0.588	0.795	0.695	0.581	0.794	0.710	0.590	0.816	0.771	0.656	0.865
Study	Luna	0.575	0.469	0.674	0.568	0.460	0.669	0.538	0.417	0.660	0.635	0.510	0.750
Study	Sol	0.722	0.655	0.790	0.729	0.668	0.794	0.750	0.684	0.819	0.729	0.646	0.823
Replication	Terra	0.765	0.678	0.844	0.754	0.666	0.835	0.762	0.665	0.847	0.771	0.667	0.865
Replication	Luna	0.727	0.644	0.803	0.728	0.646	0.804	0.719	0.630	0.802	0.760	0.656	0.854
Replication	Sol	0.808	0.745	0.867	0.814	0.739	0.881	0.825	0.743	0.903	0.844	0.760	0.917

The six panels all show a lower exact agreement at $k = 4$ than at $k = 1$, although the magnitude is model- and cohort-dependent. Selected strict success remains near the fixed model–cohort verifier in the matched task panels; the corresponding point contrasts and paired intervals are retained in the analysis manifest. A post hoc model-selection audit uses repeats 0–3 for stability selection and repeats 4–7 for held-out evaluation; it is reported as a descriptive sensitivity analysis rather than a prespecified model ranking.

G.3 Exact-set agreement and changing empirical ties

Table 41: Post hoc decomposition of observed public-score set agreement on the original 24 planning checkpoints. Single and Multiple are disjoint contributions to Agreement, according to the size of the matching sets. Mean size averages both blocks. Success evaluates the sampled-score action mixture on the held-out block; fixed verification achieves 0.771 with one call. Selection requires $3k + 1$ calls.

k	Agreement	Single	Multiple	Mean size	Success	Calls
1	0.318	0.008	0.310	2.286	0.699	4
2	0.204	0.039	0.164	1.913	0.695	7
4	0.167	0.042	0.125	1.521	0.710	13

The decomposition in Table 41 uses the original 24 planning checkpoints and public-score subsets described in Appendix G. For each k, we classify a pair of estimated maximizer sets as equal singletons, equal multi-action sets, unequal sizes, or unequal members at equal size. Their checkpoint-averaged proportions sum to one. For $k = 1, 2, 4$, the singleton fractions among individual estimated sets are 0.193, 0.368, and 0.604; unequal-size pairs contribute 0.544, 0.557, and 0.292, while unequal-member pairs of the same size contribute 0.138, 0.240, and 0.542. These are dependent subset comparisons within checkpoints, not additional independent observations.

A finite model shows why exact-set agreement alone has no monotone relationship with action quality. Let $m \geq 2$ equal-cost actions have independent Bernoulli outcomes of the same success probability

$p \in (0,1)$, with n independent draws per action. Let $f_n(k)$ and $F_n(k)$ denote the probability mass and cumulative distribution of $\mathrm{Binomial}(n,p)$, with $F_n(-1) = 0$. A specified set of j actions is exactly the empirical maximizer set when its counts equal some k and every other count is smaller. Its probability is therefore

$$a_{n,j} = \sum_{k=0}^{n} f_n(k)^j F_n(k-1)^{m-j}, \qquad j = 1, \ldots, m. \tag{51}$$

The convention $0^0 = 1$ applies when $j = m$. Two independent blocks have exact-set agreement A_n, while F_n^{set} is the probability of recovering the full population maximizer set:

$$A_n = \sum_{j=1}^{m} \binom{m}{j} a_{n,j}^2 \longrightarrow \frac{1}{m}, \qquad F_n^{\mathrm{set}} = a_{n,m} = \sum_{k=0}^{n} f_n(k)^m \longrightarrow 0. \tag{52}$$

To see the limits, the maximal binomial mass tends to zero, so the probability of any pair of equal counts tends to zero by a union bound. The empirical maximizer is consequently unique with probability tending to one, and symmetry gives each action limiting probability $1/m$. Independence of the two blocks gives the agreement limit. Also, $\sum_k f_n(k)^m \le (\max_k f_n(k))^{m-1} \to 0$. Every action selected from a block nevertheless has expected success p on a fresh independent execution.

At one draw, the complete tie set occurs when all actions succeed or all fail, with probability $p^m + (1-p)^m$. Each proper nonempty subset of size j occurs with probability $p^j(1-p)^{m-j}$. Summing their squared probabilities yields

$$A_1(p) = \big(p^2 + (1-p)^2\big)^m + 2\big[p(1-p)\big]^m = A_1(1-p). \tag{53}$$

Thus even exact knowledge of this agreement does not distinguish fresh-action success p from $1-p$ in the equal-mean family. The cross term counts a block of all successes paired with a block of all failures; both have the same full maximizer set.

The full set path can hide quality even with a unique winner. Consider two actions with independent outcomes $Y_{a,t} \sim \mathrm{Bernoulli}(p_a)$ across actions and rounds. Each round supplies one draw per action, and only $S_n = \arg\max_a\{n^{-1}\sum_{t=1}^{n} Y_{a,t} - \lambda C(a)\}$ is retained. Write $p = (p_1, p_2)$ and define $\widetilde{p} = (1-p_2, 1-p_1)$. The coupling $\widetilde{Y}_{1,t} = 1 - Y_{2,t}$ and $\widetilde{Y}_{2,t} = 1 - Y_{1,t}$ has the required independent Bernoulli coordinates under $\widetilde{p}$, and preserves every difference $Y_{1,t} - Y_{2,t}$. Hence it preserves the entire labeled path $(S_n)_{n\ge1}$, including ties, for the same fixed costs. Every set-only stopping rule or randomized action selector has the same transcript law under these two parameters.

Both success probabilities shift by $c = 1 - p_1 - p_2$, so the population preference gap is unchanged. For any action selected from this path, the expected success on an independent fresh execution shifts by c. At equal costs, $(p_1, p_2) = (0.9, 0.8)$ and $\widetilde{p} = (0.2, 0.1)$ share the same unique best action and the same set-path law, while their best-action success rates are 0.9 and 0.2.

Let $\theta_{\max} = \max_a \theta_a$ denote the best absolute success rate. For an estimator $\widehat{v}$ using only the set path and independent randomization, the sharp two-point bound is

$$\inf_{\widehat{v}} \max_{\theta \in \{p, \widetilde{p}\}} \mathbb{E}_\theta |\widehat{v} - \theta_{\max}| = \frac{|1 - p_1 - p_2|}{2}. \tag{54}$$

The estimator has the same distribution under both parameters, so the triangle inequality bounds the sum of its risks below by $|1 - p_1 - p_2|$. The constant midpoint of the two values of $\theta_{\max}$ attains equality. For $p_1 = p_2 = p$, this reduces to the common-quality ambiguity p versus $1-p$ and risk $|1-2p|/2$. Raw counts or absolute scores provide information absent from this set-only observation channel.

Absolute information and finite sampling. At equal costs, the first-step law determines the distribution of the iid increments $Y_{1,t} - Y_{2,t}$: their probabilities at $1, -1, 0$ are $u, v, 1-u-v$. It therefore determines the complete path law. Solving for the means gives Eq. (4), so a population statistic $r(p)$ removes the remaining ambiguity precisely when $r(p) \ne r(1-p_2, 1-p_1)$ for every distinct pair. Pooled success satisfies this condition. Retaining whether both actions succeeded also suffices: if $t = \Pr(Y_1 = Y_2 = 1)$, then $p_1 = u + t$ and $p_2 = v + t$. The variance of total success is invariant under the transformation and does not suffice.

For a finite-sample construction, retain B independent one-round set observations and N independent paired total-success observations. Let $\widehat{d}$ average $\mathbf{1}\{S_1 = \{1\}\} - \mathbf{1}\{S_1 = \{2\}\}$ over the former and

let $\widehat{s}$ average $Y_1 + Y_2$ over the latter. The two collections may overlap; independence is required within each collection. For $0 < \alpha_d, \alpha_s < 1$, set $\varepsilon_d = \sqrt{2\log(2/\alpha_d)/B}$ and $\varepsilon_s = \sqrt{2\log(2/\alpha_s)/N}$. Hoeffding's inequality (Hoeffding, 1963) and a union bound (Bonferroni, 1936) give

$$\Pr\{|\widehat{d} - d| \le \varepsilon_d,\ |\widehat{s} - (p_1 + p_2)| \le \varepsilon_s\} \ge 1 - \alpha_d - \alpha_s. \tag{55}$$

Consequently, $\widehat{p}_1 = (\widehat{s} + \widehat{d})/2$ and $\widehat{p}_2 = (\widehat{s} - \widehat{d})/2$ both have absolute error at most $(\varepsilon_d + \varepsilon_s)/2$ on this event. These bounds use independent restarted observations. Successive sets along one cumulative path are dependent and cannot be substituted for those samples. The midpoint attaining Eq. (54) is optimal on the specified two-point parameter class; it does not assert that the unknown gap can be estimated without sampling error.

For $m = 3$ and $p = 0.9$, $A_1 = 0.552826$ and $A_4 = 0.182657$, with unchanged expected success 0.9. This is not a universal decreasing trend. For $p = 0.5$, $A_1 = 0.156250$ and the same limit is $1/3$. The calculation provides a controlled explanation for changes in an exact-set metric. The Terra decomposition measures the associated empirical set-size changes without asserting equality of its unknown action means.

G.4 Sensitivity to self-reported totals

The planning contract checks both the selected items and the reported benefit, cost, and time. To separate solution quality from reporting accuracy, we recompute benefit from the selected item identifiers and public task fields. The sensitivity score retains this benefit whenever the selection satisfies every task constraint and the tool contract, regardless of its reported totals. Semantic success requires a feasible optimal selection; strict success also retains the original reporting requirement. Both are evaluated on the held-out block after selection with the sensitivity score. This posthoc analysis uses all 576 executions and 24 checkpoints in each cohort, with the same blocks, subset comparisons, and checkpoint bootstrap as the primary analysis.

Reported totals disagree with their recomputed values in 53 original and 48 independent-cohort executions. Removing that requirement changes the selection score in 35 and 24 executions, respectively; infeasible selections remain invalid. At $k = 4$, the selected maximizer set changes at nine original and three independent checkpoints. Exact agreement still declines from $k = 1$ to $k = 4$ in both cohorts (Table 42). The effect on selected-action quality differs: under semantic scoring, semantic success changes from 0.725 to 0.700 in the original cohort and from 0.796 to 0.799 in the independent cohort.

Table 42: Selection using recomputed feasible benefit. Agreement compares complete maximizer sets; Semantic and Strict evaluate the selected action mixture on the held-out block. Each cohort contains 24 checkpoints. Lower and Upper are the endpoints of descriptive 95% checkpoint-bootstrap intervals.

		Agreement			Semantic success			Strict success		
Cohort	k	Estimate	Lower	Upper	Estimate	Lower	Upper	Estimate	Lower	Upper
Original	1	0.359	0.250	0.479	0.725	0.614	0.827	0.697	0.592	0.796
Original	2	0.221	0.126	0.339	0.713	0.600	0.819	0.686	0.577	0.788
Original	4	0.125	0.000	0.250	0.700	0.580	0.812	0.679	0.562	0.786
Independent	1	0.477	0.365	0.594	0.796	0.706	0.877	0.764	0.678	0.843
Independent	2	0.325	0.203	0.459	0.786	0.691	0.871	0.753	0.664	0.834
Independent	4	0.208	0.042	0.375	0.799	0.698	0.889	0.764	0.668	0.849

G.5 Empirical near-tie sensitivity

This post hoc analysis retains all 24 checkpoints in each Terra panel. For checkpoint s, let m_s be the public limit on the number of selected items, and let $b_{s,(j)}$ be the jth largest public item benefit. Ignoring the remaining feasibility constraints gives the public upper bound

$$U_s = \sum\nolimits_{j=1}^{m_s} b_{s,(j)}. \tag{56}$$

Let $v_{s,a,r}$ be the recomputed feasible public benefit of action a in repetition r, with zero assigned to invalid responses. For a subset J of k repetitions, the empirical tolerance set is

$$A_{s,J,\epsilon} = \left\{ a \in A_0 : \sum\nolimits_{r \in J} v_{s,a,r} \geq \max\nolimits_{a' \in A_0} \sum\nolimits_{r \in J} v_{s,a',r} - \epsilon k U_s \right\}. \tag{57}$$

The four tolerances are $\epsilon \in \{0, 0.02, 0.05, 0.10\}$ and $k \in \{1, 2, 4\}$ (Table 43). We retain every pair of within-block subsets from the original four-draw selection and four-draw evaluation blocks. The selected action mixture is uniform over $A_{s,J,\epsilon}$ and evaluated on all four held-out draws. Strict labels and the true optimum enter only evaluation; the selection scores and U_s use public task information. Integer score sums and rational thresholds preserve exact ties. These are empirical sets with no claim of population coverage.

Table 43: Complete empirical-tolerance sensitivity. Agree is exact-set agreement, Common is nonempty-intersection frequency, Jacc. is Jaccard overlap, and Size averages set cardinality across both blocks. Strict and Public evaluate the selected mixture on the held-out block. Each cohort contributes 24 checkpoints.

		Original Terra panel						Independent Terra panel					
ϵ	k	Agree	Common	Jacc.	Size	Strict	Public	Agree	Common	Jacc.	Size	Strict	Public
0.000	1	0.318	0.898	0.619	2.286	0.699	0.662	0.404	0.940	0.693	2.411	0.765	0.699
0.000	2	0.204	0.755	0.472	1.913	0.695	0.661	0.238	0.863	0.557	2.122	0.754	0.695
0.000	4	0.167	0.458	0.299	1.521	0.710	0.668	0.083	0.750	0.375	1.833	0.762	0.695
0.020	1	0.474	0.948	0.727	2.510	0.702	0.665	0.458	0.982	0.759	2.578	0.756	0.695
0.020	2	0.358	0.919	0.660	2.361	0.704	0.663	0.346	0.958	0.680	2.438	0.758	0.695
0.020	4	0.167	0.875	0.556	2.083	0.707	0.669	0.208	0.917	0.535	2.229	0.748	0.689
0.050	1	0.576	0.984	0.810	2.677	0.710	0.668	0.578	0.992	0.829	2.714	0.761	0.697
0.050	2	0.377	0.954	0.704	2.476	0.708	0.667	0.411	0.965	0.733	2.538	0.758	0.695
0.050	4	0.125	0.917	0.569	2.188	0.714	0.670	0.250	0.917	0.562	2.271	0.748	0.689
0.100	1	0.596	0.984	0.826	2.714	0.709	0.668	0.609	0.992	0.844	2.740	0.761	0.697
0.100	2	0.395	0.954	0.714	2.493	0.710	0.667	0.434	0.965	0.741	2.552	0.758	0.695
0.100	4	0.125	0.917	0.569	2.188	0.714	0.670	0.250	0.917	0.562	2.271	0.748	0.689

For uncertainty, 10,000 checkpoint bootstrap samples resample the 12 size-12 and 12 size-16 tasks within strata, preserving every paired observation (seed 98519). Table 44 reports the paired $k = 4$ minus $k = 1$ contrasts. Intervals summarize each cohort separately. Small differences from the original zero-tolerance interval endpoints reflect the independent bootstrap stream used for this sensitivity analysis.

Table 44: Paired changes with descriptive 95% percentile intervals. Each contrast compares $k = 4$ with $k = 1$ at the indicated empirical tolerance.

		Agreement change			Strict-success change			Public-benefit change		
Cohort	ϵ	Estimate	Lower	Upper	Estimate	Lower	Upper	Estimate	Lower	Upper
Original	0.000	-0.151	-0.281	-0.005	0.011	-0.033	0.057	0.006	-0.021	0.034
Original	0.020	-0.307	-0.451	-0.156	0.005	-0.020	0.028	0.004	-0.014	0.020
Original	0.050	-0.451	-0.573	-0.315	0.004	-0.015	0.023	0.002	-0.009	0.013
Original	0.100	-0.471	-0.589	-0.341	0.004	-0.014	0.023	0.003	-0.008	0.013
Independent	0.000	-0.320	-0.406	-0.237	-0.003	-0.029	0.023	-0.005	-0.025	0.013
Independent	0.020	-0.250	-0.378	-0.115	-0.008	-0.024	0.006	-0.006	-0.018	0.004
Independent	0.050	-0.328	-0.448	-0.208	-0.013	-0.029	0.000	-0.008	-0.021	0.003
Independent	0.100	-0.359	-0.484	-0.232	-0.013	-0.029	0.000	-0.008	-0.021	0.003

At zero tolerance, every empirical action set agrees with the original integer-score argmax, reproducing agreement, tie structure, and selected strict success. At positive tolerances, changes in the reported sets are determined by the public score gaps. All thresholds and checkpoint-level outputs are retained in the scientific artifact.

G.6 AGREEMENT METRICS AND EMPIRICAL TIES

This post hoc analysis uses the same 24 checkpoints and all within-block subsets as Section 5.2. For two nonempty maximizer sets A, B, common-action agreement is $\mathbf{1}\{A \cap B \neq \varnothing\}$, Jaccard similarity is $|A \cap B|/|A \cup B|$, and independent uniform choices agree with probability $|A \cap B|/(|A||B|)$. Each metric is averaged over subset pairs within a checkpoint, then equally over checkpoints. All four agreement measures decline in this panel (Table 45).

A descriptive reference retains each block's empirical set-size distribution and independently randomizes action identities. Writing $p_{b,j}$ for the fraction of size-j sets in block b, its exact-set agreement is $\sum_{j=1}^{3} p_{1,j}p_{2,j}/\binom{3}{j}$. Excess agreement subtracts this reference from observed agreement. This relabelling reference removes persistent action identity; it is not an assumption that the actual action means are equal. The interval for the change in excess agreement includes zero, so the panel does not isolate how much of the decline is attributable to set size versus action identity.

Table 45: Alternative agreement measures on 24 checkpoints. Change is $k = 4$ minus $k = 1$; Lower and Upper are descriptive 95% percentile bounds from 10,000 bootstrap samples stratified by checkpoint size. All comparisons remain paired within checkpoint.

	Draws per action			Change		
Metric	1	2	4	Estimate	Lower	Upper
Exact set	0.318	0.204	0.167	-0.151	-0.281	-0.005
Common action	0.898	0.755	0.458	-0.440	-0.607	-0.271
Jaccard	0.619	0.472	0.299	-0.320	-0.428	-0.197
Uniform choice	0.318	0.300	0.194	-0.123	-0.206	-0.027
Size reference	0.336	0.230	0.264	-0.072	-0.180	0.024
Excess agreement	-0.018	-0.026	-0.097	-0.079	-0.189	0.049

Matching singleton sets contribute 0.008, 0.039, and 0.042 to exact agreement at $k = 1, 2, 4$. Matching multi-action sets for which both blocks attain the exact optimum contribute 0.294, 0.159, and 0.083; other matching multi-action sets contribute 0.016, 0.006, and 0.042. No matching multi-action pair has zero public score in both blocks. Thus successful empirical ties dominate the one-draw matches. These categories describe observed outputs; the optimum is used only for this evaluation and never supplied to the selector.

G.7 SET AGREEMENT ACROSS OUTCOME CONTRACTS

This post hoc analysis reuses all 1,152 Qwen3-8B executions from the 12-checkpoint development pilot below, with four rank, four path, and four filter tasks (Table 46). Each action's 32 draws form four consecutive eight-draw groups; the first and last four draws in each group define paired blocks. For $k \in \{1, 2, 4\}$, all size-k subsets are compared across each block pair. Averages weight subset pairs equally within a group, the four groups equally within a checkpoint, and the 12 checkpoints equally. Checkpoint bootstraps resample within task family. These intervals describe this fixed panel; the analysis does not convert the pilot into a confirmatory study.

Table 46: Complete-set agreement on 12 Qwen3-8B checkpoints. Strict requires both the correct answer and a matching reported score; Answer requires only the correct answer. Cost-adjusted rows subtract the fixed penalties 0.05, 0.05, and 0.10 for retry, verify, and replacement. Change is $k = 4$ minus $k = 1$, with separate 95% bootstrap bounds.

	Agreement			Change		
Outcome	$k = 1$	$k = 2$	$k = 4$	Estimate	Lower	Upper
Strict	0.840	0.811	0.854	0.014	-0.135	0.167
Strict, cost-adjusted	0.889	0.859	0.896	0.007	-0.117	0.112
Answer	0.840	0.811	0.854	0.014	-0.135	0.167
Answer, cost-adjusted	0.889	0.859	0.896	0.007	-0.117	0.112

Under strict scoring, matching multi-action sets contribute 0.831, 0.764, and 0.708 to agreement at $k = 1, 2, 4$. The share whose two blocks both have maximum score zero is 0.583 at every budget, while the all-success share falls from 0.247 to 0.125. Common-action agreement is 0.999, 0.997, and 1.000, and Jaccard similarity is 0.930, 0.910, and 0.927. Strict successes come from the rank and filter families; the path family has none.

Selecting uniformly among the first block's empirical maximizers and evaluating on the second block gives strict success 0.379, 0.389, and 0.401. These values integrate over all tied choices analytically. The answer-only counterpart is 0.462, 0.472, and 0.485; exact-set agreement is unchanged because the answer-only contract changes outcomes without changing the selected-set partitions in this panel. The artifact retains every checkpoint, cost sensitivity, both evaluation directions, and the originally sampled tie-breaking diagnostic. No population preference or recovery improvement is inferred from these post hoc comparisons.

G.8 Confidence-relation development pilot

A separate development pilot instantiated Eq. (34) on 12 new checkpoints, four each from rank, path, and filter tasks, with 32 independent seed draws per recovery action and the fixed reference $R = \{\text{verify}\}$. Qwen3-8B generated all 1,152 branches at temperature 0.7 with a 128-token cap. The public checker used only the question and documents; its outputs agreed with the evaluator on 120 preflight cases and with the stored evaluation on every collected branch. Six possible action bindings preceded a simulated source receipt that recovered the correct binding. Here eligibility required a complete receipt and request/message linkage; an incorrect or unparseable answer counted as a failed outcome while its complete execution record remained eligible. At one fixed look, three Bonferroni-adjusted two-sided Clopper–Pearson intervals supplied the candidate mean box. The sampling interpretation is conditional on the fixed model, software, and independently sampled seed law.

Of the 1,152 branches, 396 passed strict evaluation, 586 had incorrect answers, 96 had correct answers but incorrect scores, and 74 were unparseable. Every receipt passed. Eleven post-receipt overlap images remained $\{\varnothing, \{\text{verify}\}\}$; one was the singleton $\{\text{verify}\}$, yielding one positive statistical certificate. The prespecified development threshold was three positive checkpoints, so this panel remains a feasibility result. Total usage was 622,272 input and 103,199 output tokens. A post hoc answer-only check yields 492 correct answers. Two checkpoints have 96 correct answers out of 96 branches, six have none, and four have 91, 88, 67, and 54 correct answers. Pointwise confidence intervals exclude verification only in the checkpoint with the singleton image; simultaneous intervals across both outcome contracts and all 12 checkpoints leave every overlap unresolved. The frozen protocol, raw responses, and evaluation are retained separately from the earlier execution archive.

H Frozen RecoveryBench validation

Protocol and statistical unit. We froze the RecoveryBench source version at commit `c5f83f2ba4f882a9b544c7bf0fa9be1bc3859c78` and the Terminal-Bench 2.0 task source (Merrill et al., 2026) at commit `69671fbaac6d67a7ef0dfec016cc38a64ef7a77`. The panel follows the failed-checkpoint recovery setting of RecoveryBench (Tan et al., 2025) and contains 12 fixed checkpoints, the actions `retry`, `verify`, and `replace_agent`, and GPT-5.6 Luna, Terra, and Sol. Each model–checkpoint–action combination has eight repeats, yielding $12 \times 3 \times 3 \times 8 = 864$ episodes. The merged audit contains 864 expected and 864 observed keys, with no missing or duplicate episode key. The checkpoint is the statistical unit for all primary summaries.

Readout and split. Repeats 0–3 are the selection block and repeats 4–7 are the held-out block. For $k \in \{1, 2, 4\}$, every size-k subset in each block produces a complete maximizer set of the native verifier reward. Agreement is the fraction of cross-block subset pairs with equal sets. All-zero pairs is the fraction of all cross-block subset pairs for which every action in both compared subsets has zero reward. Held-out success averages the binary native-reward outcome over the actions in the selection maximizer set and uses no outcome from repeats 4–7 during selection. Binary success is defined by native verifier reward > 0; the raw continuous reward is retained as a descriptive quantity.

The complementary all-success-pair frequency is 0 in all nine model–budget cells. A checkpoint-level rank diagnostic is reported below; its intervals condition on non-constant bootstrap resamples and are descriptive rather than a model-ranking claim.

Table 47: RecoveryBench set-readout intervals. Each entry is Estimate, Lower, or Upper from 10,000 paired checkpoint-bootstrap resamples. Bootstrap resampling is over the 12 checkpoints and preserves all model, action, and repeat observations within a checkpoint. Held-out success is the binary native-reward outcome averaged over the selected action set.

		Agreement			All-zero pairs			Held-out success		
Model	k	Estimate	Lower	Upper	Estimate	Lower	Upper	Estimate	Lower	Upper
Luna	1	0.854	0.729	0.958	0.854	0.729	0.958	0.014	0.000	0.035
Luna	2	0.750	0.583	0.917	0.750	0.583	0.917	0.014	0.000	0.035
Luna	4	0.667	0.417	0.917	0.583	0.333	0.833	0.014	0.000	0.035
Terra	1	0.776	0.583	0.943	0.776	0.583	0.943	0.045	0.007	0.090
Terra	2	0.688	0.458	0.896	0.688	0.458	0.896	0.035	0.007	0.069
Terra	4	0.583	0.333	0.833	0.583	0.333	0.833	0.014	0.000	0.035
Sol	1	0.734	0.563	0.891	0.708	0.510	0.885	0.083	0.000	0.196
Sol	2	0.627	0.438	0.815	0.590	0.375	0.799	0.075	0.000	0.182
Sol	4	0.417	0.167	0.667	0.417	0.167	0.667	0.063	0.000	0.167

Decision-value analysis. To keep model and action choice separate from held-out evaluation, all choices use repeats 0–3. Agreement-guided selection first chooses the model with the larger agreement between two selection sub-blocks and then chooses the action with the largest selection-block mean, averaging tied choices uniformly. Outcome-mean selection uses the same split but chooses the model–action pair by its selection-block outcome mean. The direct nine-arm and two-stage outcome rules coincide on this panel. The resulting task-level held-out success intervals are reported in Table 48.

Table 48: Descriptive decision-value analysis on the frozen RecoveryBench panel. Selection uses repeats 0–3 and evaluation uses repeats 4–7. Each entry is the held-out success Estimate, Lower, or Upper from 10,000 task-bootstrap resamples.

Policy	Estimate	Lower	Upper
Agreement-guided, $k = 1$	0.013	0.000	0.028
Agreement-guided, $k = 2$	0.030	0.003	0.066
Outcome-mean direct	0.065	0.000	0.171
Fixed verify, Luna	0.000	0.000	0.000
Fixed verify, Terra	0.021	0.000	0.063
Fixed verify, Sol	0.146	0.000	0.313

Execution provenance. The official reward is available for every merged row, while the execution status records how that reward was obtained. Native verifier observations account for 364 episodes and contain 46 positive rewards; model-protocol failures account for 487 episodes and contain one positive reward; invalid adapter attempts account for 13 episodes and contain no positive reward. These categories are provenance fields rather than additional statistical units. The primary table follows the declared reward contract, and the status breakdown below keeps semantic outcome evidence separate from execution-protocol evidence.

Table 49: Execution provenance for all 864 RecoveryBench episodes. Positive rate is the number of reward-positive official receipts divided by the number of episodes in the row.

Execution status	Episodes	Reward-positive	Positive rate
Native verifier observed	364	46	0.126
Model protocol failure	487	1	0.002
Invalid adapter attempt	13	0	0.000
All episodes	864	47	0.054

Rank concordance. We correlate checkpoint-level agreement with held-out success for each model and budget. Spearman estimates are negative in all nine cells; Kendall's τ_b has the same sign. The table reports the Spearman intervals.

Table 50: Checkpoint-level rank concordance between exact-set agreement and held-out success on the 12 RecoveryBench checkpoints. Each entry is Estimate, Lower, or Upper from 10,000 checkpoint-bootstrap resamples; resamples with a constant vector are omitted.

Model	k	Estimate	Lower	Upper
Luna	1	-0.443	-1.000	-0.094
Luna	2	-0.443	-1.000	-0.094
Luna	4	-0.632	-1.000	-0.357
Terra	1	-0.783	-1.000	-0.283
Terra	2	-0.741	-1.000	-0.266
Terra	4	-0.529	-1.000	-0.255
Sol	1	-0.796	-0.936	-0.535
Sol	2	-0.776	-0.916	-0.535
Sol	4	-0.376	-0.693	-0.174