CERTIFY, REFUTE, OR ABSTAIN: CLAIM-AWARE POLICY AUDITING IN MULTI-AGENT REINFORCEMENT LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

High task return does not establish that a policy satisfies a declared protocol or deviation-gap claim. *Claim-Aware Protocol Auditing* (CAPA) distinguishes the evidence needed to certify a claim from the evidence sufficient to refute it: an upper bound supports certification, a lower bound supports refutation, and insufficient evidence requires abstention. We formulate unconditional false-decision and false-selection guarantees and show that passive rollouts cannot improve a worst-case gap bound on an indistinguishable bounded-game pair. A constructed soft leader–follower example separates return ranking from protocol-map conformity: the return-10 candidate violates the map tolerance, whereas the selected conforming candidate has return 3.956. Synthetic boundary calibration exposes undercoverage of Gaussian plug-in bounds; matched Student-t bounds require iid Gaussian gains. On fixed learned MPE candidates, a seeded neural-deviation audit reports all 45 probes with paired returns and explicit Gaussian-approximation limits. No learned candidate has an unrestricted upper-bound certificate. The resulting evaluation contract makes policy performance, evidence of a violation, and the stronger evidence needed for certification separate objects of assessment.

Keywords: protocol auditing, selective prediction, abstention, multi-agent reinforcement learning, claim certification

1 INTRODUCTION

Structured multi-agent reinforcement learning (MARL) uses ordering, hierarchy, or factorization to define learners and policy classes (11; 14; 17; 21; 23). A high-return policy need not satisfy a separately requested sequential-response, surrogate-transfer, or equilibrium-style claim. We study this logical distinction through fixed candidate tables. Our reward-only selectors are illustrative rules to which a protocol claim is explicitly attached; the experiments do not show that published IPPO, MAPPO, or HAPPO evaluations made that unsupported inference.

Best-response evaluation already separates achieved payoff from worst-case deviation. In particular, approximate exploitability is a lower bound on true exploitability (19), so unsuccessful response search cannot establish a small unrestricted gap. CAPA does not introduce this asymmetry. It makes the audited policy transformation, objective, violation functional, tolerance, evidence assumptions, and error budget explicit, and uses them to license a decision before ranking candidates. This matters when one table mixes policy-map residuals, value-evaluation residuals, and searched deviations: these quantities support different claims even when all are called semantic evidence.

For a declared claim, CAPA certifies with an admissible upper bound, refutes with an admissible lower bound, and otherwise abstains. Surrogate fit and useful return are supplementary diagnostics; they cannot override a known violation. The statistical guarantee bounds the probability of a false decision over audit data and auditor randomization. It does not bound the error rate conditional on making a decision, and it does not protect retrospectively selected audit families unless their selection is included in the stated error budget.

We formalize this contract, derive its evidence and selection guarantees, and illustrate its behavior on an exact constructed example and a seeded neural audit of fixed MPE candidates. Synthetic

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.

boundary tests distinguish the guarantees of valid bounds from the behavior of the Gaussian approximation used in the learned audit. Exploratory learned measurements use a defective training-reward collector and incomplete RNG seeding; we report them separately in Appendix C.2 rather than pool them with the seeded audit.

2 Related Work

Exploitability, NashConv, and approximate response search already quantify unilateral improvement (7; 8; 19). Timbers et al.'s approximate-exploitability work explicitly establishes the lower-bound interpretation. Adversarial-policy learning supplies related search methods (5); best-response-dynamics metrics concern other solution objects (20). Selective prediction studies abstention in a different setting (3).

Structured MARL has also considered distributed learners, value factorization, and hierarchical game decompositions (9; 10; 17; 22). These methods define learning architectures or restricted policy classes; they do not by themselves establish that a candidate delivered by a finite response search has the claimed unilateral identity. Non-stationarity and optimization stability introduce additional failure modes in learned multi-agent systems (4; 12; 15). Classical stochastic-game learning and equilibrium refinements provide the broader game-theoretic context (2; 16). We do not claim novelty for approximate best responses, Bonferroni correction, residual bounds, or abstention itself.

The incremental object is the explicit claim contract across policy transformations, residual and gap objectives, surrogate transfer assumptions, and candidate selection. For the *same* deviation claim, an exploitability evaluator with valid bounds and honest uncertainty/abstention can make the same decisions as CAPA. Our exact examples illustrate specification and reporting; our learned records do not establish an additional statistical or computational capability over that comparator. Nor do they document failures of published learner evaluations.

3 Claims and Audit Decisions

Let $\mathcal{G}$ be a game, T a declared protocol, and $\Pi = \{\pi^{(1)}, \ldots, \pi^{(M)}\}$ a fixed candidate set. Subscripts s and u distinguish the semantic and surrogate operators, respectively. A claim specifies the policy actually evaluated, the information available to a deviator, the initial-state distribution, reward and entropy conventions, and the horizon or discount. These choices define J_i; they are not interchangeable across suites. For deviations b_i in the declared admissible space, write

$$G_i(\pi) = \sup_{b_i} J_i(b_i, \pi_{-i}) - J_i(\pi_i, \pi_{-i}), \qquad G_s(\pi) = \max_i G_i(\pi). \tag{1}$$

The candidate policy belongs to its deviation space, so $G_i \geq 0$. This is an existing unilateral-deviation criterion, related to exploitability and NashConv (7; 8).

A different claim can concern a value residual

$$R_s(V, \pi) = \left\| V - \mathcal{T}_s^{\pi} V \right\|_{\infty}, \qquad R_u(V, \pi) = \left\| V - \mathcal{T}_u^{\pi} V \right\|_{\infty}. \tag{2}$$

A small policy-evaluation residual controls value error under a suitable operator assumption; it does not establish policy optimality or equilibrium. Our exact experiment instead uses a *policy-map* residual, defined in Section 5. When V is part of an audited record, it is fixed along with π.

Definition 3.1 (Claim contract and unconditional errors)**.** A claim q consists of a violation functional $\Delta_q(\pi)$, tolerance ϵ_q, and binary certification and refutation predicates $c_q(C)$ and $r_q(C)$ on an audit log C. For every fixed admissible game and candidate, valid predicates satisfy

$$\Pr\{c_q(C) = 1,\ \Delta_q(\pi) > \epsilon_q\} \leq \delta_q^{\mathrm{C}}, \tag{3}$$

$$\Pr\{r_q(C) = 1,\ \Delta_q(\pi) \leq \epsilon_q\} \leq \delta_q^{\mathrm{R}}. \tag{4}$$

Probability is over audit data and internal randomization, with the game and candidate fixed. The audit log records the evidence assumptions and the scope of each error budget.

One may set both budgets to a common δ_q. The inequalities are *unconditional error-event bounds*, not posterior probabilities of claim truth or guarantees conditional on certification/refutation. For

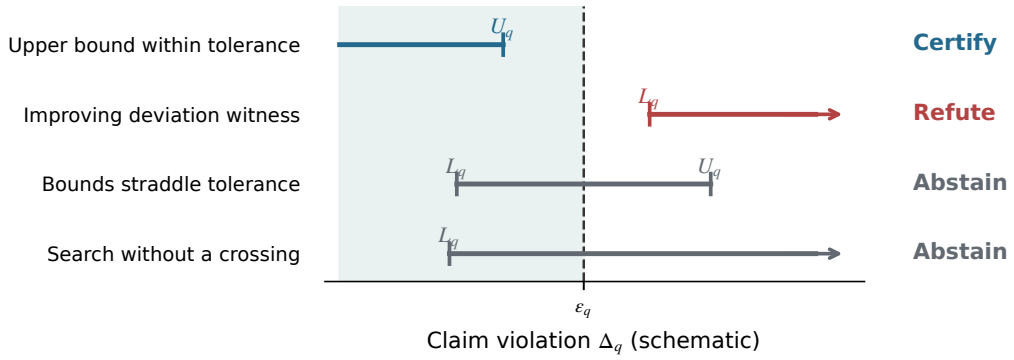

Figure 1: **Directional evidence in the claim contract.** An upper bound within ϵ_q licenses certification, while a lower-bound crossing gives refutation evidence. A search with no crossing remains Abstain: it does not become an upper bound on the unsearched deviation class. Horizontal positions are schematic and encode only the stated threshold tests.

Algorithm 1 CAPA: claim-aware auditing and selection

Require: Fixed candidates Π, contract q, evidence budget

for each candidate $\pi^{(m)}$ **do**
  Record the claim-matched backend, assumptions, estimates, and error allocation in C_m.
  Apply the certification/refutation predicates; record a decision and separate diagnostic tags.
end for
Form Π_q from candidates whose certification predicate alone passes.
if $\Pi_q = \emptyset$ **then**
  return $\perp$, audit table.
else
  return the candidate with smallest certified violation upper bound, breaking ties by return and then a fixed index.
end if

example, an invalid candidate certified with probability δ_q^{C}, with the selector otherwise abstaining, satisfies (3), although every selected return is invalid. Conditioning on independently fitted candidates and probes is permissible when the bounds hold for every such fit; using evaluation outcomes to fit or select probes requires a simultaneous or adaptive guarantee.

Decisions and diagnostics. A valid upper bound U_q permits certification when $U_q \leq \epsilon_q$; a valid lower bound L_q permits refutation when $L_q > \epsilon_q$. With exact Δ_q, these tests give Certified or Refuted without abstention. If neither predicate passes, the decision is Abstain. If inconsistent evidence makes both predicates pass, the auditor abstains and flags the conflict. Tags such as SURROGATE-ONLY, PRACTICAL, OBSTRUCTED, and UNCERTIFIED record surrogate fit, performance, or missing assumptions separately. In particular, a Refuted candidate may also fit a surrogate. A promotion is *unsupported* when a selector attaches q without a passing certification predicate, and *false* when $\Delta_q > \epsilon_q$. Unsupported does not imply known false; a statistical refutation also retains its stated error risk.

Figure 1 makes the decision asymmetry explicit: one admissible improving probe can refute a candidate, whereas an absent crossing supplies no certification.

Exact models, verified response oracles, and validated finite-cover bounds can supply upper bounds. A finite deviation search supplies lower evidence for the unrestricted gap, but upper evidence only for its searched class unless a transfer bound is available. Without coverage or model assumptions, passive return evaluation does not in general supply an informative unrestricted upper bound or a counterexample. The full residual, response-oracle, and finite-cover contracts and proofs are in Appendices A and B; no such learned upper-bound backend is instantiated by the retained experiments.

4 Evidence and Selection Guarantees

The results below expose evidence requirements rather than new concentration or dynamic-programming methods. In particular, a theorem for valid confidence bounds does not make a plug-in Gaussian bound valid.

Theorem 4.1 (Bounded passive ambiguity)**.** *Fix horizon $H \geq 1$ and per-step rewards in $[0, W]$, $W > 0$. There are two finite games $\mathcal{G}_0, \mathcal{G}_1$ and a deterministic candidate π with identical passive trajectory laws for every sample size, but*

$$G_s^{\mathcal{G}_0}(\pi) = 0, \qquad G_s^{\mathcal{G}_1}(\pi) = HW.$$

For any possibly randomized passive upper bound U satisfying $\Pr_{\mathcal{G}}\{G_s^{\mathcal{G}}(\pi) \leq U\} \geq 1-\delta$ uniformly over this game class,

$$\Pr_{\mathcal{G}_0}\{U < HW\} = \Pr_{\mathcal{G}_1}\{U < HW\} \leq \delta.$$

Thus on this ambiguous candidate, passive data cannot improve on the worst-case bound with probability greater than δ, even in the zero-gap game.

The complete proof is in Appendix B.1. *Construction.* At the first step the candidate chooses an action yielding zero forever. An unchosen action yields zero forever in $\mathcal{G}_0$ and W at each of H steps in $\mathcal{G}_1$. Other agents may have one action. The passive transcript is the same, so a bound below HW fails in $\mathcal{G}_1$ and has the same probability in $\mathcal{G}_0$. The constant HW remains a valid bound; certification at tolerance at least HW is possible. This is a worst-case missing-coverage result, not an impossibility for every policy or for auditors with model access.

Theorem 4.2 (Randomized decision–abstention tradeoff)**.** *For the pair in Theorem 4.1, fix $0 \leq \epsilon < HW$. Any passive rule taking the three mutually exclusive decisions Certify, Refute, and Abstain has the same decision probabilities $p_{\mathrm{C}}, p_{\mathrm{R}}, p_{\mathrm{A}}$ in both games. If its false-certification probability is at most δ_{C} and false-refutation probability at most δ_{R} in every game, then*

$$p_{\mathrm{C}} \leq \delta_{\mathrm{C}}, \qquad p_{\mathrm{R}} \leq \delta_{\mathrm{R}}, \qquad p_{\mathrm{A}} \geq \max\{0, 1 - \delta_{\mathrm{C}} - \delta_{\mathrm{R}}\}.$$

With zero error budgets it must abstain almost surely on this pair.

The complete proof is in Appendix B.2. The tradeoff follows because certification is false in $\mathcal{G}_1$ and refutation is false in $\mathcal{G}_0$. For $\delta_{\mathrm{C}} + \delta_{\mathrm{R}} \leq 1$, randomizing with exactly these decision probabilities attains equality on the pair. This tightness is pair-specific, not a construction of a uniformly valid rule on a larger class. It does not say every randomized, non-abstaining decision is invalid; it bounds how often such decisions are possible under both error constraints.

Proposition 4.3 (Active refutation)**.** *For a fixed admissible deviator b_i, let $g_i = J_i(b_i, \pi_{-i}) - J_i(\pi)$. If a lower bound L_i satisfies $\Pr\{L_i > g_i\} \leq \delta_i$, then*

$$\Pr\{L_i > \epsilon,\ G_s(\pi) \leq \epsilon\} \leq \delta_i. \tag{5}$$

For a fixed finite family of probes, valid bounds at levels $\delta_{i,k}$ give the same false-refutation bound with $\sum_{i,k} \delta_{i,k}$ when any probe crosses.

The complete proof is in Appendix B.3. Indeed, $g_i \leq G_s$, so a false crossing implies failure of the lower bound. Conditional validity given fitting data suffices for independently evaluated fitted probes. No independence across probes is required for the union bound. Common-random-number pairing preserves the gain estimand if both rollout marginals are correct; it reduces variance only when the induced covariance is positive (Proposition A.5).

Theorem 4.4 (Unconditional claim-sound selection)**.** *For a fixed candidate family, suppose $\Pr\{c_q(C_m) = 1, \Delta_q(\pi^{(m)}) > \epsilon_q\} \leq \delta_m$ for every m. Let S denote the output of Algorithm 1, with $S = \perp$ for abstention. Then*

$$\Pr\{S \neq \perp,\ \Delta_q(S) > \epsilon_q\} \leq \sum_{m=1}^{M} \delta_m. \tag{6}$$

Here the event means that a candidate is returned and its claim is false; Δ_q is not evaluated at $S = \perp$.

The complete proof is in Appendix B.4. A false selected return requires at least one false certification, so a union bound proves (6). Selection by return within the certified set would obey the same guarantee. Neither this theorem nor per-run probe correction gives conditional correctness among returned candidates or a global guarantee for retrospectively selected audit families. Detailed proofs and the assumptions for additional backends appear in the appendix.

5 Suite-Specific Claim Contracts

Exact soft leader–follower map. The constructed game has one initial state and two actions per player: the leader chooses $a \in \{0,1\}$, and the follower observes a before choosing $b \in \{0,1\}$. There is one terminal payoff, no discounting, and payoff matrices

$$A = \begin{pmatrix} 4 & 2 \\ 1 & 10 \end{pmatrix}, \qquad B = \begin{pmatrix} 2 & 0 \\ 2 & -2 \end{pmatrix}.$$

A policy is $x = (p, q_0, q_1) \in [0,1]^3$, where $p = \Pr(a = 1)$ and $q_a = \Pr(b = 1 \mid a)$. Set $\tau = 0.5$ and $s_\tau(u, v) = e^{v/\tau}/(e^{u/\tau} + e^{v/\tau})$. The declared map is

$$\begin{aligned} q_a^* &= s_\tau(B_{a0}, B_{a1}), \quad v_a^* = (1 - q_a^*)A_{a0} + q_a^* A_{a1}, \\ F(x) &= (s_\tau(v_0^*, v_1^*), q_0^*, q_1^*), \qquad \Delta_e(x) = \|x - F(x)\|_\infty . \end{aligned} \tag{7}$$

Crucially, F is *constant in* x. Each soft choice uses its decision maker's payoff and a Shannon-entropy bonus with temperature τ; the leader's v_a^* contains no follower-entropy bonus. Reported return is the leader's raw expected payoff, without entropy. The surrogate F_u replaces B by A in the follower response; $R_u = \|x - F_u(x)\|_\infty$. The exact-map claim is $\Delta_e \le 0.08$, in policy-probability units, with zero statistical error. The maximum covers both follower contingencies, including off-path actions. This is conformity to a specified soft map, not an unregularized Nash or value-residual certificate.

Learned MPE finite-horizon gap. We use the *Multi-Agent Particle Environment* (MPE) (11) through PettingZoo's AEC interface (18). For Simple Spread (three agents) and Simple Reference (two), let $d(\pi)_i(o)$ be the loaded actor's deterministic argmax action, with the actor's tie rule. The evaluation wrapper supplies zero recurrent state at each call, so it does not evaluate a recurrent checkpoint's full hidden-state dynamics. Actions are discrete, and the local ratio parameter is fixed at (0.5). Let μ be the task implementation's native reset-state distribution and

$$J_i^{25}(\eta) = \mathbb{E}_{\mu,\eta}\left[\sum_{t=0}^{24} r_{i,t}\right], \qquad \Delta_m(\pi) = \max_i \sup_{b_i} \left[J_i^{25}(b_i, d(\pi)_{-i}) - J_i^{25}(d(\pi))\right].$$

The finite-horizon claim is $\Delta_m \le 0.05$, in agent-specific undiscounted episode-return units. There is no entropy bonus, discount, or division by 25 in this audit objective. The deviation space contains all policies using the agent's own observation/action history, including the evaluated memoryless neural argmax probes; no privileged state is supplied to those probes. Opponents remain frozen. A counterexample here concerns this transformed finite-horizon policy under μ, not the original stochastic checkpoint's discounted objective, and not equilibrium at every state. Native reset laws depend on the PettingZoo/MPE version; runtime identity is incompletely recorded for the exploratory measurements. The intended expectation over resets and the uncertainty of its finite-seed estimate must therefore be distinguished.

6 Evidence and Its Limits

We examine an exact contract illustration, a fixed-protocol neural audit, and exploratory observations. No experiment below demonstrates superiority over a best-response evaluator using the same bounds and an honest abstention rule.

Exact ranking reversal. Table 1 contains six hand-constructed candidates from the single payoff example in Section 5, not six independent games. Reward-only selection chooses a return-10 candidate with residual approximately 1; attaching the exact map claim to that choice is a false promotion. The contract selects the oracle with return 3.956 and residual 0; the noisy candidate also

Table 1: Exact map-conformity audit, $\epsilon = 0.08$. Verdicts follow the unrounded residual. The star marks reward-only selection; the check marks CAPA selection.

Candidate	Return	R_s	R_u	Verdict
Semantic oracle ✓	3.956	0.000	1.000	Certified
Noisy semantic	3.829	0.060	0.970	Certified
Surrogate oracle	10.000	1.000	0.000	Refuted
Reward-greedy joint ⋆	10.000	1.000	0.018	Refuted
Simultaneous uniform	4.250	0.500	0.500	Refuted
Wrong-order proxy	1.000	0.997	1.000	Refuted

Table 2: Complete seeded Simple Spread audit. Each row aggregates 15 probes with 128 paired evaluation episodes each. LCBs use Gaussian quantiles with sample SD and nominal level $0.05/15$; a crossing is approximate refutation evidence, not a finite-sample guarantee. No candidate has an unrestricted upper-bound certificate.

Candidate	Return/agent	Best LCB	Positive probes	Audit outcome
IPPO	-27.16	-0.70	0/15	abstain
MAPPO	-24.34	0.22	1/15	approximate refutation
HAPPO	-15.22	-0.59	0/15	abstain

passes but has a larger residual. All four other candidates are *Refuted* for this exact claim. Their surrogate/performance tags are explanatory only. Zero false promotions follows from using the same exact residual to define truth and certification; it is not an independent calibration result.

Seeded deviation audit. We freeze one Simple Spread checkpoint from each of IPPO, MAPPO, and official HARL HAPPO and fit five neural deviations for each of the three agents. The protocol fixes all candidates, probe settings, and evaluation seeds before running; all 45 probes are reported without best-seed filtering. Fixed-probe confirmation is evaluated separately. Each probe receives 200 training episodes and 128 paired evaluation episodes at horizon 25. Training uses REINFORCE with discount 0.99 and a two-hidden-layer, 64-unit Tanh network; evaluation uses the undiscounted deterministic policies defined in Section 5. Python, NumPy, Torch, and environment RNG streams are fixed. The collector associates each sampled action with the cumulative reward on the agent's next AEC turn, including the terminal reward. Regression tests cover every agent position, both episode endings, real MPE parameter updates, and repeated CPU runs.

In Table 2, MAPPO has one crossing: agent 1, probe 3, has mean gain 1.231, sample SD 4.235, and LCB 0.216. Using a nominal allocation of $0.05/45$ across all three candidates gives 0.086, still above the tolerance 0.05; the Gaussian approximation remains necessary for interpreting either number. IPPO has no crossing in this audit, unlike the exploratory observations under a different collection protocol. HAPPO also has no crossing; its best LCB is -0.587, despite its higher observed return. No unrestricted upper bounds are available: the candidate selector therefore abstains even when a candidate has not been refuted. This is one audit of three fixed checkpoints, not an estimate of algorithm-level exploitability or training-seed robustness. A separate evaluation reconstructs the named MAPPO response with an identical weight hash and tests 512 independent paired resets. Its mean gain is 1.305 (95% Student-t interval $[0.910, 1.699]$); the one-sided bound at nominal level $0.05/45$ is 0.687. This supports the same fixed-policy violation on independent resets, not an unrestricted certificate or additional candidate-training replication; the interval remains distribution-dependent. Two fixed policy-perturbation controls give restoration-gain LCBs of 10.14 and 7.06 using the original HAPPO agent as the deviator (Table 3). These check the evaluation path with a supplied response. In a separate three-seed search control on the fixed-action HAPPO perturbation, the supplied response gains 6.880, whereas all three 200-episode neural searches have negative gains. Across fixed budgets of 50, 200, and 800 episodes, the three-seed mean gains are -2.778, -3.759, and -2.190, with no detected learned response at any budget. Re-evaluating the nine fixed responses on three disjoint 256-reset sets leaves all 27 checkpoint–set gains negative, while the restoration gain is 7.274–7.403. Training seeds are the replication unit conditional on the fixed evaluation sets;

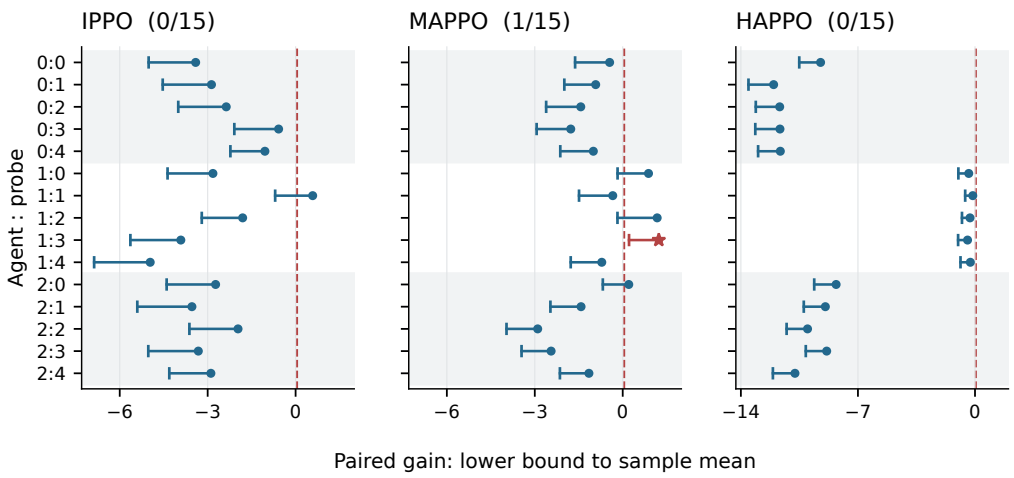

Figure 2: **Seeded deviation evidence.** Each marker is one of the 45 fixed neural probes; the segment runs from the Bonferroni Gaussian lower confidence bound to the paired sample mean over 128 evaluation episodes. The dashed line is the declared 0.05 gap tolerance. Only the MAPPO agent-1/probe-3 row crosses it. These are approximate lower-bound diagnostics, not unrestricted upper-bound certificates.

budgets reuse the same seeds and sets. These controls show that the tested search can miss a large admissible improvement, not that larger budgets are generally ineffective or that the unrestricted gap is small (Appendix C.1).

The full probe geometry in Figure 2 is consistent with the table-level conclusion: the absence of crossings for IPPO and HAPPO is abstention, whereas the one MAPPO crossing is approximate refutation evidence.

Reward-collection and randomization controls. The exploratory collection protocol supplied zero training rewards to non-final agents and left Torch randomness unseeded. Those measurements combine random and partially trained deviations; they cannot establish reproducible trained-search effectiveness. Their values and limitations are reported in Appendix C.2, including a reused HAPPO row that does not constitute held-out evidence. They are not pooled with Table 2.

Calibration and sensitivity diagnostics. Repeated Reference records contain crossings in one of three named audit streams for IPPO and two for MAPPO, and none for HAPPO. Named streams do not reproduce the unseeded torch randomness. Candidate-training-seed summaries and threshold sweeps reuse the same exploratory collection protocol and do not isolate probe-search variability. Synthetic boundary checks (Table 8) give five-probe false-refutation rates of 6.44% $[5.79, 7.15]$ for Gaussian plug-in bounds and 5.04% $[4.47, 5.68]$ for matched Student-t bounds over 5000 trials, with pointwise 95% Wilson Monte Carlo intervals. The latter guarantee requires iid Gaussian gains; under skewed noise the corresponding rates are 9.66% and 8.14%, so neither rule is distribution-free. Ensemble power comparisons use more samples than single-probe evaluation. A budget plot uses deterministic slack $0.32/\sqrt{B}$ added to an exact residual, not sampled confidence intervals. Full values, sample-budget qualifications, and reuse relationships are retained in Appendix C.

7 Limitations

The contract certifies only its declared mathematical object. The bounded ambiguity theorem concerns missing off-policy coverage; stronger observation models can avoid it. Finite-cover certification requires a verified cover and Lipschitz constant, not a finite neural ensemble alone. Exploratory learned measurements have incomplete provenance and reward-collection and randomization defects. The seeded audit uses complete reward collection and fixed random streams, but remains limited to one fixed checkpoint per learner and approximate intervals. Its finite search provides no

unrestricted upper bound, no learned equilibrium certification, and no demonstration of stronger search than existing best-response evaluators. The supported conclusion is narrower: return ranking and evidence for a declared claim should be reported separately, with unconditional error events and explicit abstention when the required evidence is absent.

AI USE STATEMENT

Generative AI tools assisted with mathematical revisions, implementation, and the design, execution, and analysis of probe experiments and synthetic diagnostics. The authors take responsibility for the final content.

REFERENCES

[1] Dimitri P Bertsekas. *Dynamic Programming and Optimal Control, Vol. II*. Athena Scientific, 4th edition, 2012.

[2] Michael Bowling and Manuela Veloso. Rational and convergent learning in stochastic games. In *Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence (IJCAI)*, volume 17, pp. 1021–1026, 2001.

[3] Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

[4] Matthieu Geist, Bruno Scherrer, and Olivier Pietquin. A theory of regularized markov decision processes. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2160–2169. PMLR, 2019.

[5] Wenbo Guo, Xian Wu, Sui Huang, and Xinyu Xing. Adversarial policy learning in two-player competitive games. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 3910–3919. PMLR, 2021.

[6] Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963.

[7] Marc Lanctot, Vinicius Zambaldi, Audrunas Gruslys, Angeliki Lazaridou, Karl Tuyls, Julien Pérolat, David Silver, and Thore Graepel. A unified game-theoretic approach to multiagent reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 30, pp. 4190–4203, 2017.

[8] Marc Lanctot, Edward Lockhart, Jean-Baptiste Lespiau, Vinicius Zambaldi, Satyaki Upadhyay, Julien Pérolat, Sriram Srinivasan, Finbarr Timbers, Karl Tuyls, Shayegan Omidshafiei, Daniel Hennes, Dustin Morrill, Paul Muller, Timo Ewalds, Ryan Faulkner, János Kramár, Bart De Vylder, Brennan Saeta, James Bradbury, David Ding, Sebastian Borgeaud, Matthew Lai, Julian Schrittwieser, Thomas Anthony, Edward Hughes, Ivo Danihelka, and Jonah Ryan-Davis. OpenSpiel: A framework for reinforcement learning in games. *arXiv preprint arXiv:1908.09453*, 2019.

[9] Martin Lauer and Martin Riedmiller. An algorithm for distributed reinforcement learning in cooperative multi-agent systems. In *Proceedings of the Seventeenth International Conference on Machine Learning (ICML)*, pp. 535–542. Morgan Kaufmann, 2000.

[10] Zun Li, Feiran Jia, Aditya Mate, Shahin Jabbari, Mithun Chakraborty, Milind Tambe, and Yevgeniy Vorobeychik. Solving structured hierarchical games using differential backward induction. In *Proceedings of the Thirty-Eighth Conference on Uncertainty in Artificial Intelligence*, volume 180 of *Proceedings of Machine Learning Research*, pp. 1107–1117. PMLR, 2022.

[11] Ryan Lowe, Yi Wu, Aviv Tamar, Jean Harb, Pieter Abbeel, and Igor Mordatch. Multi-agent actor-critic for mixed cooperative-competitive environments. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pp. 6382–6393, Red Hook, NY, USA, 2017. Curran Associates Inc.

[12] Georgios Papoudakis, Filippos Christianos, Arrasy Rahman, and Stefano V Albrecht. Dealing with non-stationarity in multi-agent deep reinforcement learning. *arXiv preprint arXiv:1906.04737*, 2019.

[13] Martin L Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, 2014.

[14] Tabish Rashid, Mikayel Samvelyan, Christian Schroeder de Witt, Gregory Farquhar, Jakob Foerster, and Shimon Whiteson. QMIX: Monotonic value function factorisation for deep multi-agent reinforcement learning. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, pp. 4295–4304. PMLR, 2018.

[15] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

[16] Reinhard Selten. Reexamination of the perfectness concept for equilibrium points in extensive games. *International Journal of Game Theory*, 4(1):25–55, 1975.

[17] Peter Sunehag, Guy Lever, Audrunas Gruslys, Wojciech M Czarnecki, Vinicius Zambaldi, Max Jaderberg, Marc Lanctot, Nicolas Sonnerat, Joel Z Leibo, Karl Tuyls, and Thore Graepel. Value-decomposition networks for cooperative multi-agent learning. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems (AAMAS)*, pp. 2085–2087, 2018.

[18] J Terry, Benjamin Black, Nathaniel Grammel, Mario Jayakumar, Ananth Hari, Ryan Sullivan, Luis S Santos, Clemens Dieffendahl, Caroline Horsch, Rodrigo Perez-Vicente, Niall Williams, Yashas Lokesh, and Praveen Ravi. PettingZoo: Gym for multi-agent reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 34, pp. 15032–15043. Curran Associates, Inc., 2021. URL `https://proceedings.neurips.cc/paper_files/paper/2021/hash/803f7c4c3ff61b71be53a0c803bfb57f-Abstract.html`.

[19] Finbarr Timbers, Nolan Bard, Edward Lockhart, Marc Lanctot, Martin Schmid, Neil Burch, Julian Schrittwieser, Thomas Hubert, and Michael Bowling. Approximate exploitability: Learning a best response. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence*, pp. 3487–3493, 2022. doi: 10.24963/ijcai.2022/484.

[20] Rui Yan, Xiaoming Duan, Zongying Shi, Yisheng Zhong, Jason R. Marden, and Francesco Bullo. Policy evaluation and seeking for multi-agent reinforcement learning via best response. *arXiv preprint arXiv:2006.09585*, 2020.

[21] Chao Yu, Akash Velu, Eugene Vinitsky, Yu Gao, Yu Wang, Alexandre Bayen, and Yi Wu. The surprising effectiveness of PPO in cooperative multi-agent games. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pp. 24611–24624, 2022.

[22] Kaiqing Zhang, Zhuoran Yang, and Tamer Başar. Multi-agent reinforcement learning: A selective overview of theories and algorithms. In Kyriakos G. Vamvoudakis, Yan Wan, Frank L. Lewis, and Derya Cansever (eds.), *Handbook of Reinforcement Learning and Control*, pp. 321–384. Springer International Publishing, Cham, 2021.

[23] Yifan Zhong, Jakub Grudzien Kuba, Xidong Feng, Siyi Hu, Jiaming Ji, and Yaodong Yang. Heterogeneous-agent reinforcement learning. *Journal of Machine Learning Research*, 25(32): 1–67, 2024.

A Additional Backend Contracts

These results specify sufficient evidence, not capabilities established for the learned experiments. The proofs are included in Appendix B.

Assumption A.1 (Complete invariant value space)**.** For fixed π, the semantic operator is a self-map on a nonempty complete normed value space $\mathcal{V}$, and for all $V, V' \in \mathcal{V}$,

$$\|\mathcal{T}_s^\pi V - \mathcal{T}_s^\pi V'\|_\infty \le \rho \|V - V'\|_\infty, \qquad 0 \le \rho < 1.$$

Thus it has a unique fixed point V_s^π.

Proposition A.2 (Residual-to-value error)**.** *Under Assumption A.1, $R_s(V, \pi) \le \epsilon$ implies*

$$\|V - V_s^\pi\|_\infty \le \epsilon/(1-\rho).$$

The proof is in Appendix B.5. This standard contraction argument (1; 13) controls evaluation of the fixed policy only. It supplies no bound on G_s without an additional relation between value evaluation and policy optimality. For a value-error claim at tolerance η, the residual threshold must be $(1-\rho)\eta$, not automatically η. Neither empirical suite applies this proposition as an equilibrium certificate.

Proposition A.3 (Surrogate transfer)**.** *For operators defined at the same V, π,*

$$R_s(V, \pi) \le R_u(V, \pi) + B_\Delta(V, \pi), \quad B_\Delta(V, \pi) = \|\mathcal{T}_u^\pi V - \mathcal{T}_s^\pi V\|_\infty.$$

Under Assumption A.1, bounds $R_u \le \epsilon_u$ and $B_\Delta \le \beta$ imply $\|V - V_s^\pi\|_\infty \le (\epsilon_u + \beta)/(1-\rho)$.

The proof is in Appendix B.6. The mismatch bound must itself be valid on the claimed event. A small surrogate residual does not estimate β. If residual and mismatch bounds are statistical, their joint failure budget must be included in (3). The same triangle inequality bounds distances to two policy maps, but those distances have a different estimand from value error.

Proposition A.4 (Response upper oracle)**.** *Fix a candidate and agent i. Suppose a backend supplies a response $\hat{b}_i$, an error certificate $\alpha_i \ge 0$, and bounds U_i^b, L_i^π such that, jointly with probability at least $1 - \delta_i$,*

$$\sup_{b_i} J_i(b_i, \pi_{-i}) \le J_i(\hat{b}_i, \pi_{-i}) + \alpha_i, \quad J_i(\hat{b}_i, \pi_{-i}) \le U_i^b, \quad J_i(\pi) \ge L_i^\pi.$$

Then $U_i = U_i^b + \alpha_i - L_i^\pi$ satisfies $\Pr\{G_i(\pi) > U_i\} \le \delta_i$. Certifying only when every $U_i \le \epsilon$ has false-certification probability at most $\sum_i \delta_i$.

The proof is in Appendix B.7. The joint event includes optimization error and both value bounds. Finding a high-return response does not establish α_i; no learned row supplies this oracle.

Proposition A.5 (Paired rollout gains)**.** *Fix a candidate and deviator independently of their evaluation samples, or condition on independent fitting data. Let the coupled returns $R_i^b(\omega), R_i^\pi(\omega)$ have the correct marginals for $J_i(b_i, \pi_{-i})$ and $J_i(\pi)$. For iid coupled seeds $\omega_1, \ldots, \omega_n$,*

$$D_i = R_i^b - R_i^\pi, \qquad \mathbb{E}D_i = J_i(b_i, \pi_{-i}) - J_i(\pi).$$

A valid lower bound L_i for $\mathbb{E}D_i$ yields (5)*. For finite second moments,*

$$\mathrm{Var}(D_i) = \mathrm{Var}(R_i^b) + \mathrm{Var}(R_i^\pi) - 2\,\mathrm{Cov}(R_i^b, R_i^\pi).$$

The proof is in Appendix B.8. Identical seed labels do not alone prove correct coupling marginals or iid sampling. Positive covariance is sufficient for a variance reduction relative to independent rollouts, but is not automatic. For a bounded-return interval, a valid range and concentration inequality must be supplied; Gaussian plug-in bounds are not replaced by such an inequality merely by calling the evaluation paired.

Proposition A.6 (Finite-cover deviation audit)**.** *For each i, fix $\mathcal{B}_i = \{b_{i,1}, \ldots, b_{i,K_i}\}$ independently of evaluation data, and define*

$$G_i^{\mathcal{B}} = \max_{b \in \mathcal{B}_i} J_i(b, \pi_{-i}) - J_i(\pi).$$

Suppose every probe value and the candidate value for every agent has a simultaneous confidence interval, with total failure probability at most δ. Maximizing the differences of probe lower and

candidate upper endpoints gives a lower bound on $G_i^{\mathcal{B}}$; maximizing probe upper minus candidate lower endpoints gives an upper bound. If $\mathcal{B}_i$ is a ζ_i-net of the full deviation space in metric d_i, and $J_i(\cdot, \pi_{-i})$ is L_i-Lipschitz, then

$$G_i \le G_i^{\mathcal{B}} + L_i \zeta_i.$$

Specifically, assume each evaluated return lies in an interval of width W_i, has n_i iid samples, and choose $0 < \delta_i < 1$ with $\sum_i \delta_i \le \delta < 1$. Writing $\widehat{J}_i^b = \widehat{J}_i(b, \pi_{-i})$ and $\widehat{J}_i^\pi = \widehat{J}_i(\pi)$, Hoeffding's inequality gives

$$r_i = W_i \sqrt{\frac{\log(2(K_i+1)/\delta_i)}{2n_i}}, \qquad U_i = \max_{b \in \mathcal{B}_i} \widehat{J}_i^b - \widehat{J}_i^\pi + 2r_i + L_i \zeta_i.$$

Then $\Pr\{\exists i : G_i > U_i\} \le \delta$, and certifying when all $U_i \le \epsilon$ obeys (3)*.*

The proof is in Appendix B.9. Independence across different evaluated returns is unnecessary, but iid samples for each fixed return and a valid joint budget are necessary for this stated construction. An independently fitted random cover is allowed if the metric, covering and Lipschitz requirements hold conditional on the fit. Without cover transfer, a class upper bound is not an unrestricted upper bound. The available finite-cover check uses supplied means, sample sizes, cover radii, and Lipschitz constants. It checks arithmetic, not whether neural deviations form a cover or whether those constants are valid in MPE.

Corollary A.7 (Probe-ensemble false-refutation control)**.** *Let $N = \sum_i K_i > 0$, and let every fixed or independently fitted admissible probe have a valid gain lower bound $L_{i,k}$ with failure probability at most δ/N. Then*

$$\Pr\{\max_{i,k} L_{i,k} > \epsilon,\ G_s(\pi) \le \epsilon\} \le \delta.$$

If no bound crosses, there is no unrestricted certification without an upper-bound backend.

The proof is in Appendix B.10. Bonferroni combines valid marginal error bounds; it does not calibrate invalid ones. This corollary does not establish nominal finite-sample coverage for the archived Gaussian-approximate learned rows.

B Proofs

B.1 Bounded passive ambiguity

Proof of Theorem 4.1. At the first step the audited agent has two actions a_0, a_1; any other agents have one action. The remaining $H-1$ steps have no effective choices. The candidate always takes a_0, giving reward zero at every step in both games. In $\mathcal{G}_0$, a_1 also gives zero at every step. In $\mathcal{G}_1$, it gives reward W immediately and moves to a state paying W for every remaining step. The first action is the only consequential deviation, so the two gaps are 0 and HW. All transitions, observations, and rewards on the candidate path agree. Thus any number of passive trajectories, together with identically distributed internal randomness, has the same law in both games. Let this common law be P. Coverage in $\mathcal{G}_1$ requires $P(U \ge HW) \ge 1 - \delta$. Consequently $P(U < HW) \le \delta$ in $\mathcal{G}_0$ as well. Every H-step return lies in $[0, HW]$, so HW is always a valid gap upper bound. The result prohibits an informative improvement on the ambiguous transcript, not all upper bounds. □

Unbounded-class extension. If the game class permits arbitrary positive off-policy reward with the same zero passive transcript, a uniformly valid extended-real upper bound must satisfy $P(U \ge c) \ge 1 - \delta$ for every $c > 0$. Continuity of probability as integer $c \to \infty$ gives $P(U = +\infty) \ge 1 - \delta$. Hence for $\delta < 1$ no almost-surely finite upper bound is uniformly valid on that class. This extension needs unbounded rewards and is not the statement used for the bounded theorem.

B.2 Randomized decision–abstention tradeoff

Proof of Theorem 4.2. The common transcript law, including an independent auditor random seed, implies identical probabilities for all three decisions. Since $0 \le \epsilon < HW$, certification is a false decision in $\mathcal{G}_1$, so $p_{\mathrm{C}} \le \delta_{\mathrm{C}}$. Refutation is a false decision in $\mathcal{G}_0$, so $p_{\mathrm{R}} \le \delta_{\mathrm{R}}$. Mutual exclusivity

and exhaustiveness give $p_{\mathrm{A}} = 1 - p_{\mathrm{C}} - p_{\mathrm{R}} \geq 1 - \delta_{\mathrm{C}} - \delta_{\mathrm{R}}$, as well as $p_{\mathrm{A}} \geq 0$. At zero budgets both non-abstaining probabilities vanish. For budgets summing to at most one, a rule that ignores the transcript and chooses Certify, Refute, or Abstain with probabilities $\delta_{\mathrm{C}}, \delta_{\mathrm{R}}, 1 - \delta_{\mathrm{C}} - \delta_{\mathrm{R}}$ attains the bound on this pair. No deterministic-decision premise is used. □

B.3 Active refutation

Proof of Proposition 4.3. Admissibility gives $g_i \leq G_i \leq G_s$. On $\{G_s \leq \epsilon, L_i > \epsilon\}$ we therefore have $L_i > g_i$. This event inclusion proves (5). The union over all crossing probes proves the finite-family statement. If bounds hold conditional on independent fitting data, averaging their error bounds over that data gives the unconditional result. □

B.4 Unconditional claim-sound selection

Proof of Theorem 4.4. Let $E_m = \{c_q(C_m) = 1, \Delta_q(\pi^{(m)}) > \epsilon_q\}$. Any false selected return belongs to $\bigcup_m E_m$, because the algorithm returns a candidate only from the certified set. Therefore its probability is at most $\sum_m \Pr(E_m) \leq \sum_m \delta_m$. When no candidate is admissible, $S = \perp$, which is outside the false-return event. The proof never conditions on $S \neq \perp$, and requires no independence between the candidate audits. For randomly fitted candidates it applies conditional on fitting data only if each stated error bound remains valid under that conditioning. □

B.5 Residual-to-value error

Proof of Proposition A.2. The contraction self-map on the complete space has a unique fixed point. Using it and Assumption A.1,

$$\|V - V_s^\pi\|_\infty \leq \|V - \mathcal{T}_s^\pi V\|_\infty + \|\mathcal{T}_s^\pi V - \mathcal{T}_s^\pi V_s^\pi\|_\infty \leq \epsilon + \rho \|V - V_s^\pi\|_\infty .$$

Rearrange using $1 - \rho > 0$. □

B.6 Surrogate transfer

Proof of Proposition A.3. Insert $\mathcal{T}_u^\pi V$ between V and $\mathcal{T}_s^\pi V$ and apply the triangle inequality. Substitute the two supplied bounds and apply Proposition A.2. For statistical inputs this reasoning holds on their joint validity event, so a union bound on input failures gives the claimed output error budget. □

B.7 Response upper oracle

Proof of Proposition A.4. On the stated joint event,

$$G_i = \sup_{b_i} J_i(b_i, \pi_{-i}) - J_i(\pi) \leq J_i(\hat{b}_i, \pi_{-i}) + \alpha_i - J_i(\pi) \leq U_i^b + \alpha_i - L_i^\pi .$$

If every displayed upper bound is at most ϵ while $\max_i G_i > \epsilon$, some agent's joint event must fail. A union bound over agents gives the unconditional false-certification bound. □

B.8 Paired rollout gains

Proof of Proposition A.5. Correct rollout marginals and linearity give $\mathbb{E}(R_i^b - R_i^\pi) = J_i(b_i, \pi_{-i}) - J_i(\pi)$. Apply Proposition 4.3 to any valid lower bound for this mean. Expanding the variance of the difference yields the stated covariance identity. Independent evaluation has zero covariance, so the paired variance is smaller precisely when the covariance is positive, with equality at zero. □

B.9 Finite-cover deviation audit

Proof of Proposition A.6. On the simultaneous event, each difference of a probe lower endpoint and the candidate upper endpoint is below its true gain; the analogous difference of upper and

Table 3: New restoration controls with a supplied reference policy, not learned best responses. Gains use the repaired evaluator and Gaussian-approximate bounds. Both controls retain all 128 paired returns in the artifact.

Agent-0 perturbation	Mean restoration gain	LCB	Crossing
action permutation	11.42	10.14	1
constant	8.09	7.06	1

lower endpoints is above it. Taking maxima preserves both inequalities for $G_i^{\mathcal{B}}$. For any full-space deviation b, choose a net point b' with $d_i(b, b') \le \zeta_i$. Lipschitzness gives

$$J_i(b, \pi_{-i}) \le J_i(b', \pi_{-i}) + L_i \zeta_i.$$

Take the supremum over b and subtract $J_i(\pi)$. For the concrete intervals, Hoeffding (6) gives $P(|\widehat{J} - J| > r_i) \le \delta_i/(K_i+1)$ for each of the K_i probe means and the candidate mean. The union includes $\sum_i (K_i + 1)$ means, not only the probes, and has failure probability at most $\sum_i \delta_i \le \delta$. The probe upper endpoint contributes r_i and the candidate lower endpoint another r_i, giving $2r_i + L_i\zeta_i$. This proves both coverage and the false-certification event bound. □

B.10 Probe-ensemble false-refutation control

Proof of Corollary A.7. Apply Proposition 4.3 to each admissible probe with failure budget δ/N. A false declaration from any crossing is contained in the union of the N lower-bound failures. Its probability is at most δ. No inference about unsearched deviations follows from the absence of a crossing. □

C Additional Empirical Diagnostics

C.1 Seeded audit protocol and restoration controls

The new audit uses public IPPO/MAPPO final actor weights and the fixed official HARL HAPPO actor weights included in the supplement. Probe seeds are 91700, 91800, and 91900 by candidate, with offsets $10000i + 1000k$ for agent i and probe k; phase-specific SHA256-derived seeds initialize the RNGs. Evaluation resets use 917000–917127 for all candidates and probes, with paired candidate/deviator rollouts. The training learning rate is 0.003, discount 0.99, and REINFORCE returns are centered and scaled within an episode as specified in the code. Each probe uses 200 training episodes; these are audit-search episodes, not additional candidate learner steps. The artifact checks all 5760 paired gain records against the 45 probe summaries and preserves all non-crossings.

For separate controls we change only HAPPO agent 0: either map discrete actions by $[1, 2, 3, 4, 0]$, or fix its action to 0. The unchanged original agent-0 policy is a known reference deviation, with the other two agents frozen. Both controls were fixed before their evaluation and are reported; no response is trained for these controls. They use 128 paired resets, seeds 918000–918127, and nominal $\alpha = 0.05/2$ per test in this separate two-test family. There is no joint confidence claim combining these controls with the 45 neural probes or the historical searches.

Fixed-probe re-evaluation and pairing diagnostics. The MAPPO confirmation retains the original candidate and agent-1/probe-3 response, reconstructed under the original numerical runtime and verified by its weight hash. The 512 evaluation resets are disjoint from the initial audit and an earlier cross-runtime diagnostic. That diagnostic reproduced all 128 historical return pairs despite differing response weights, and gave gain 1.110 on a separate 512-reset bank; it is not used as proof of identical policies. Neither bank is pooled into the initial table. For all 45 retained initial probes, paired returns have positive sample covariance: paired gain variance is 0.089–0.566 times the sum of the two marginal return variances. This checks the realized benefit of pairing, not its universal validity. Replacing the original global Gaussian bound by a Student-t bound gives 0.062 for the named crossing; it falls below the 0.05 tolerance in 40 of 128 leave-one-observation-out variants. These overlapping diagnostics motivate the separate confirmation rather than additional independent evidence or distribution-free coverage.

Table 4: Response-search budget diagnostic. Learned gain is mean ± SD across three training seeds; the one-sided lower bound uses those three seed means and is conditional on the fixed 128-reset evaluation set. Restoration is a supplied positive control, not a learned response.

Budget	Learned gain	Seed LCB	Detected	Restoration [LCB]
50	-2.778 ± 0.481	-3.588	$0/3$	$6.880\ [6.111]$
200	-3.759 ± 0.589	-4.752	$0/3$	$6.880\ [6.111]$
800	-2.190 ± 0.590	-3.186	$0/3$	$6.880\ [6.111]$

Table 5: Robustness across disjoint evaluation sets. Learned columns average the same three training-seed checkpoints within budget and set. Restoration bounds use the 256 unique paired resets in each set; repeated control values stored beside different fits do not increase the sample size.

Set	Restoration [LCB]	Budget 50	Budget 200	Budget 800	Detected
1	$7.403\ [6.849]$	-2.381	-3.150	-1.858	$0/9$
2	$7.274\ [6.752]$	-2.529	-3.157	-1.924	$0/9$
3	$7.303\ [6.734]$	-2.606	-3.319	-1.906	$0/9$

Known-violation search control. We retain the previously specified HAPPO agent-0 fixed-action-0 perturbation and compare three independently initialized response searches, their untrained initializations, and the original agent as a supplied response. All searches receive 200 training episodes and share 128 fresh paired evaluation resets, disjoint from the earlier restoration test. The supplied response has mean gain 6.880 and one-sided Student-t LCB 5.881, conservatively using nominal level $0.05/3$. The three trained gains are $-4.157, -3.083, -4.038$, versus initialized gains $-1.848, -4.100, -2.231$. Training improves one initialization but none of the trained probes finds the known violation. These approximate intervals and three search seeds characterize this fixed-budget search, not independent candidate generalization or its best-response optimum.

A preceding diagnostic fixed MAPPO agent 1 to action 0. Restoring its original policy gave gain -1.452, so that perturbation was not a validated positive control. Its trained gains were $-1.292, -3.092, -2.060$; one untrained initialization gained 0.884 but lost that advantage after training. We retain this failed control rather than infer a known profitable restoration from the perturbation alone. Both diagnostics reinforce the distinction between unsuccessful search and evidence of a small unrestricted gap.

Budget and evaluation-set robustness. We retain every fit in a fixed 3×3 grid of training budgets and seeds. All fits use the same candidate, response architecture, optimizer settings, and paired evaluation set. Table 4 reports the complete budget-level summary. Increasing the budget to 800 episodes makes the fitted responses less harmful than at 200 episodes on average, but every mean gain remains negative and the profitable restoration remains well separated.

The nine fixed responses are also evaluated on three mutually disjoint sets of 256 resets that do not overlap fitting or preceding evaluation seeds. Candidate, learned-response, and restoration returns are paired within each reset. Table 5 shows that the restoration is profitable in every set, whereas all 27 learned checkpoint–set means are negative. The 27 detection indicators are descriptive crossed diagnostics, not independent tests or multiplicity-adjusted confirmations.

Because budgets share training seeds and evaluation sets, budget effects require paired contrasts rather than differences between marginal lower bounds. Table 6 first subtracts budgets within training seed and evaluation set and then averages the same three sets within each seed. The intervals use only the three resulting training-seed differences and are conditional on the selected checkpoints and evaluation sets. Although the 800-episode fits are less harmful for all three seeds, none has positive gain; the small number of seeds does not establish a general monotone budget effect.

C.2 Historical evidence status and budgets

The historical learned measurements in this subsection are exploratory, from mixed random/partially trained search with unseeded torch initialization and training action sampling. No historical row is

Table 6: Paired budget contrasts after averaging the same evaluation sets within each training seed. Intervals are two-sided 95% Student-t intervals over three matched training-seed differences.

Contrast	Seed 1	Seed 2	Seed 3	Mean [95% CI]
$200 - 50$	-1.243	-0.739	-0.129	$-0.704\ [-2.089, 0.682]$
$800 - 200$	1.653	1.551	0.734	$1.313\ [0.062, 2.564]$
$800 - 50$	0.411	0.812	0.606	$0.609\ [0.111, 1.108]$

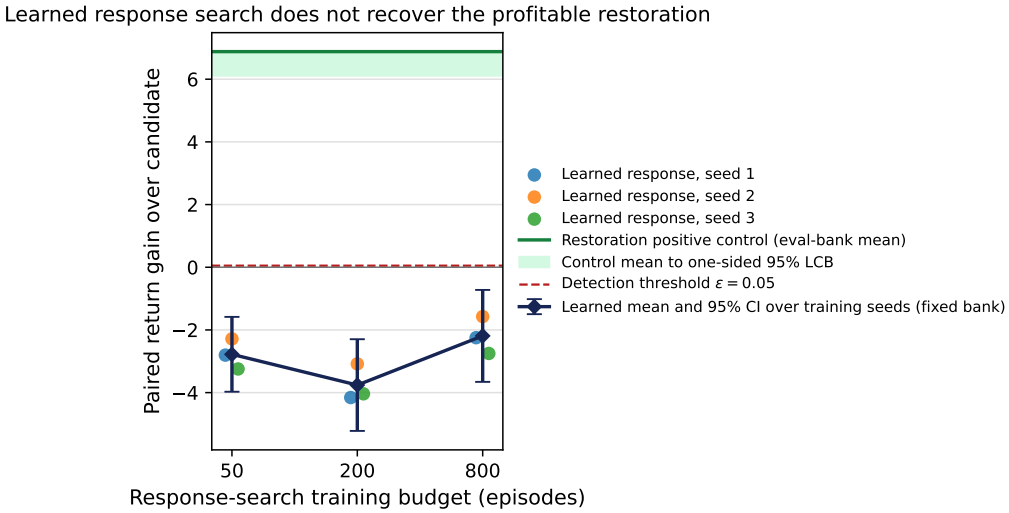

Figure 3: Learned response gains under the fixed budget grid. Points are training seeds; the dark summary uses training-seed uncertainty conditional on the fixed evaluation set. The restoration line is a supplied positive control.

asserted to have been regenerated by repaired code or reproducible from its named audit seed. The exploratory collector supplied zero training rewards to non-final agents. In particular, the exploratory Simple Spread IPPO crossing and the positive fixed-candidate Reference audit-stream crossings concern agent 0. A profitable random policy is a legitimate deviation, but neither the strength of trained search nor the meaning of non-refutation can be inferred from these records.

The exploratory Simple Spread table has one seed-101 candidate per generator: 12.8M environment-step IPPO and MAPPO checkpoints and a fixed official HARL HAPPO checkpoint. Spread uses three agents, five probes per agent, and 128 paired evaluation episodes, with 256 in focused confirmations. Reference uses two agents, three probes per agent, 120 nominal training episodes, and 96 paired evaluation episodes. Training-episode counts describe attempted legacy probe fitting, including zero-signal cases, rather than successful learning. Training used a discounted REINFORCE objective with nominal discount 0.99; the audited objective is nevertheless the undiscounted 25-cycle return in Section 5.

With α_0 denoting the nominal level assigned to one test, the historical gain bound is

$$L = \overline{D} - z_{1-\alpha_0}\, s_D/\sqrt{n}.$$

Individual full-agent runs allocated nominal family level 0.05 across their agent–probe tests. Focused runs used their own allocations; preferred-source summaries combined full and focused records after exploration. There is no documented prospective confirmation-selection rule giving the combined table global coverage. Disjoint evaluation seed ranges support separation of some evaluation streams, not independence of checkpoint selection or exact historical executable identity. Future confirmatory evidence would need candidates, transformations, probe fitting, family membership, and allocation fixed before independent evaluation; none is inferred retroactively here.

For the official HAPPO main row, the return -14.9948157 and LCB -0.4533632 are exactly the evidence reused by the former held-out-source entry. No additional held-out checkpoint evaluation is

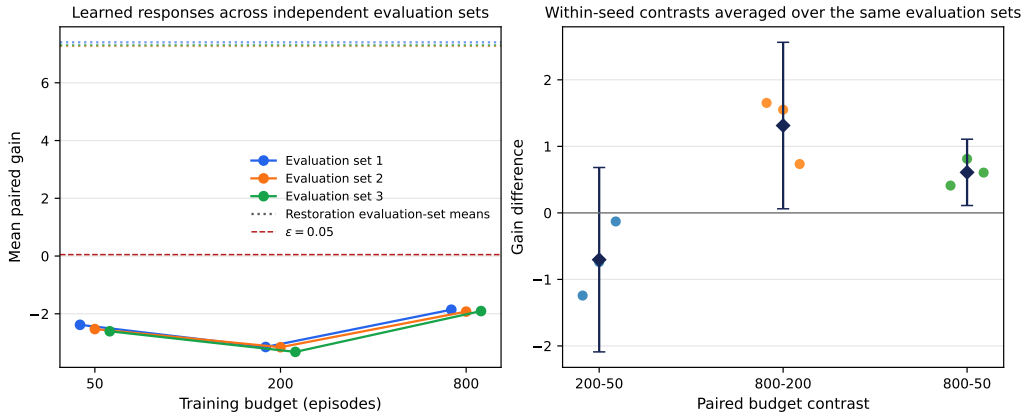

Figure 4: Evaluation-set robustness and paired budget contrasts. The left panel averages the same three training seeds within each setting; the right panel shows the three seed-level paired differences and their conditional intervals.

Table 7: Historical exploratory Simple Spread measurements. These selected-source Gaussian-approximate crossings are not calibrated finite-sample verdicts. Search mixed random and partially trained probes with unseeded Torch initialization. The new audit is reported separately in Table 2.

Candidate	Return/agent	Best LCB	Legacy outcome
Official HAPPO ⋆	-14.99	-0.45	No crossing
MAPPO	-24.50	-0.58	Initial hit unconfirmed
IPPO	-27.19	1.09	Agent-0 crossing

represented. Action-permuted and fixed-action HAPPO controls retain reported LCBs 8.548714 and 6.500365, and per-agent returns -24.646 and -22.519. They are legacy exploratory perturbation controls, not independent source generalization or calibrated confirmation; companion raw deltas for these controls were not located.

C.3 Synthetic boundary calibration

The current estimator diagnostic uses 5000 trials per case/method, seed 123, tolerance 0.05, and nominal family level 0.05. It includes 77 case/method cells; no learned candidate or probe is regenerated by these simulations. Each fixed probe uses 40 gain observations, each requiring two returns. Three- and five-probe ensembles use 120 and 200 observations versus 40 for a single probe. In the moderate shared-Gaussian-noise case, current Gaussian-rule rejection rates are 12.4% $[11.5, 13.4]$, 41.8% $[40.4, 43.1]$, 67.7% $[66.4, 68.9]$, and 79.4% $[78.3, 80.5]$ for independent single, paired single, three-probe, and five-probe evaluation. This comparison does not establish equal-budget ensemble efficiency.

Unlike the earlier zero-gain nulls, every probe in the boundary controls has true gain $\epsilon = 0.05$ and discovery probability one. Table 8 transcribes the five-probe rows from the delivered calibration diagnostic. The Gaussian and Student-t methods use matched random streams. Intervals are pointwise 95% Wilson intervals for Monte Carlo rejection probabilities, not simultaneous intervals across cases or confidence intervals for learned gains. The Gaussian five-probe rate 6.44% excludes the nominal 5% level in its displayed interval. The matched t rate 5.04% is compatible with its exact Gaussian-model error probability.

For the Gaussian case, shared noise has SD 1 and each idiosyncratic return component SD 0.35; the noisy case doubles both scales. The symmetric heavy-tailed case replaces each independent idiosyncratic return component by scaled $t_3/\sqrt{3}$ noise. The skewed case uses Gaussian idiosyncratic noise for the original return and scaled $1 - \text{Exponential}(1)$ noise for the deviator. Their difference

Table 8: Synthetic five-probe boundary calibration, 5000 trials per cell, 40 paired gain samples per probe (200 total), nominal family level 5%. Entries are false-refutation percentages with pointwise 95% Wilson Monte Carlo intervals. Student-t validity requires iid Gaussian gains; non-Gaussian rows are stress tests. These are synthetic results, separate from all legacy learned-policy records.

Boundary noise	Gaussian plug-in	Matched Student-t
Gaussian	6.44 [5.79, 7.15]	5.04 [4.47, 5.68]
Noisy Gaussian	6.40 [5.75, 7.11]	5.16 [4.58, 5.81]
Symmetric t_3	5.58 [4.98, 6.25]	4.28 [3.75, 4.88]
Skewed	9.66 [8.87, 10.51]	8.14 [7.41, 8.93]

Table 9: Historical exploratory Reference audit streams. The source's stable-not-refuted labels denote no observed crossing only. Each cell counts crossings out of six agent–probe tests; no global family guarantee or trained-search robustness is established.

Generator	5151	5152	5153	Max LCB	Robust audit status
HAPPO fair-25000	0/6	0/6	0/6	-2.53	stable not-refuted
MAPPO	0/6	1/6	2/6	1.07	refuted in 2/3 seeds
IPPO	0/6	0/6	2/6	3.38	refuted in 1/3 seeds

remains skewed after the shared Gaussian component cancels. These assumptions describe the synthetic generator, not an established distributional model for MPE returns.

A normal quantile with estimated sample SD is not exact even under iid Gaussian gains. For five independent boundary probes, $n = 40$, and per-probe level 0.01, its analytical family error is

$$1 - \left[F_{t_{39}}\left(\Phi^{-1}(0.99)\right)\right]^5 \approx 0.06164;$$

the matched Student-t rule gives $1 - 0.99^5 \approx 0.04901$. The simulation intervals are consistent with these values. For skewed gains, the t method also over-rejects: its observed 8.14% rate is not evidence of distribution-free validity. A bounded-return guarantee requires a justified range and an appropriate concentration bound. No Student-t or bounded-return interval has been retroactively substituted into the legacy learned rows; their measured gains and approximate crossings remain unchanged.

C.4 Named streams, candidate seeds, and thresholds

Table 9 reuses fixed Reference candidates over three named audit streams. IPPO and MAPPO have crossings in one and two streams, respectively, with maximum LCBs 3.38 and 1.07. HAPPO has no crossing in 18 agent–probe rows and a maximum of -2.53. A no-crossing count is not a robustness certificate. Those 18 rows are probe tests, not 18 independently trained candidates. Unseeded torch randomness prevents interpreting named streams as reproducible full-audit seeds.

Table 10 retains three completed candidate-training seeds per Reference generator. The audit's named seed is 5151, but the torch stream was not fixed. The HAPPO seed-101 cell uses the fair-25000 checkpoint, a separate suite from the 12.8M-step Spread candidates and the official HAPPO row. Stored crossings occur in two candidate seeds for both IPPO and MAPPO and none for HAPPO. HAPPO's reported return is -20.99 ± 1.74; the spread is SD across three candidate-seed means, not SE or a confidence interval. Because probe randomness was uncontrolled, this comparison does not isolate candidate variability.

Table 11 changes the tolerance on the same 54 retained probe rows. HAPPO has no crossings from $\epsilon = 0$ to 1, while IPPO and MAPPO each retain at least one. This is threshold sensitivity on reused data, not independent replication.

C.5 Exact controls, selection, and deterministic slack

The oracle in Table 1 is $F(x)$, the surrogate oracle is $F_u(x)$, and the noisy candidate clips $F(x) + (0.06, -0.04, 0.03)$ coordinatewise to $[0, 1]$. The reward-greedy, uniform, and wrong-order proxy

Table 10: Historical exploratory Reference candidate-seed summary. Each cell is historical crossings out of six tests, with maximum approximate LCB in parentheses. Return is per-agent mean ± SD across three candidate-seed means. These Reference audits were not rerun.

Generator	Seed 101	Seed 102	Seed 103	Refuted seeds	Return
HAPPO	0/6 (-5.44)	0/6 (-2.31)	0/6 (-5.52)	0/3	-20.99 ± 1.74
MAPPO	0/6 (-1.09)	1/6 (0.64)	1/6 (0.54)	2/3	-37.63 ± 3.04
IPPO	0/6 (-0.03)	1/6 (2.48)	1/6 (4.32)	2/3	-40.88 ± 4.65

Table 11: Historical exploratory threshold sensitivity on reused Reference audit rows. Counts are approximate-bound crossings out of 18 stream–agent–probe tests per generator.

Generator	Max LCB	$\epsilon = 0$	$\epsilon = 0.05$	$\epsilon = 0.1$	$\epsilon = 0.25$	$\epsilon = 0.5$	$\epsilon = 1$
HAPPO fair-25000	-2.53	0/18	0/18	0/18	0/18	0/18	0/18
MAPPO	1.07	3/18	3/18	2/18	2/18	2/18	1/18
IPPO	3.38	2/18	2/18	2/18	2/18	2/18	1/18

candidates are $(1, 0, 1)$, $(0.5, 0.5, 0.5)$, and $(1, 1, 0)$. Names such as learner or oracle in historical records describe constructed policies, not additional training experiments. The surrogate and reward-greedy rows have supplementary SURROGATE-ONLY tags, the uniform row a PRACTICAL tag, and the wrong-order proxy an UNCERTIFIED tag; all four are top-level Refuted for the declared exact map claim. The noisy and exact oracle rows are Certified.

Figures 5 to 7 retain visualizations of those exact quantities. Historical diagnostic tags in plots do not supersede the verdicts in Table 1. The previously named transfer sweep is a relisting of exact cases and protocol choices, not a continuous mismatch-parameter experiment. The cross-task learned summary likewise reuses Spread and Reference audits; its denominators of three agent-level rows and 18 probe rows are not comparable measures of coverage.

Selector ablations are logical illustrations on fixed tables. Reward-only or surrogate-only selectors make false promotions *if* they attach the exact map claim to their high-return choices. Omitting a mismatch requirement can falsely certify the surrogate oracle. Omitting abstention can attach an unsupported claim to official HAPPO, but no unrestricted truth is known for that learned row. All learned certification bits are zero by construction, so zero unsupported promotions by the full selector is not an independent test of its evidence calibration. The claim–evidence distinction is therefore: false promotion is demonstrated in the exact example; only unsupported promotion is established by the learned selector example.

For Figure 8, the plotted upper quantity is

$$U_B(x) = \Delta_e(x) + 0.32/\sqrt{B}, \qquad B \in \{1, 2, 4, 8, 16, 32, 64\}.$$

It is deterministic slack above already known truth, not a sampled interval. The slack values are $0.320, 0.226, 0.160, 0.113, 0.080, 0.057, 0.040$. The oracle first passes $U_B \leq 0.08$ at $B = 16$; no noisy data or sampling coverage is involved. Zero false promotions is mechanically guaranteed by nonnegative slack. The plot labels this quantity as deterministic slack.

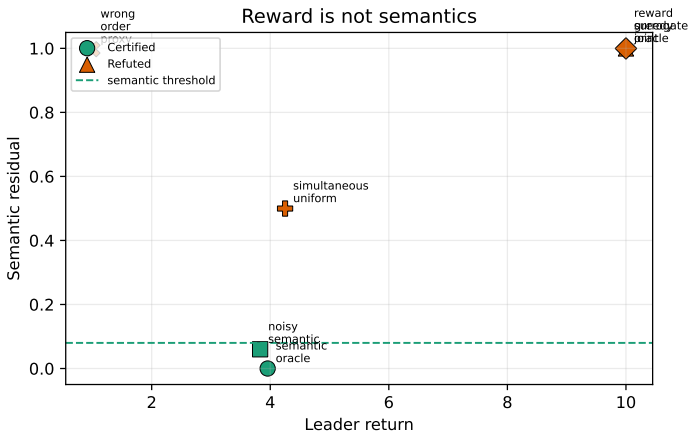

Figure 5: Retained exact return–map-residual separation. Four rows violate the map tolerance and are Refuted, regardless of any historical supplementary tag.

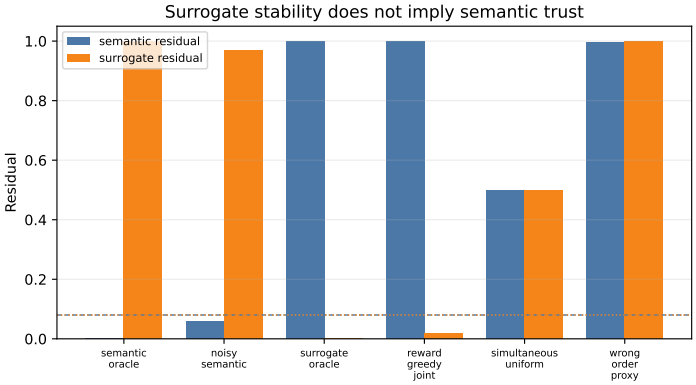

Figure 6: Retained exact surrogate diagnostic. Low surrogate residual alone does not imply low distance to the declared map.

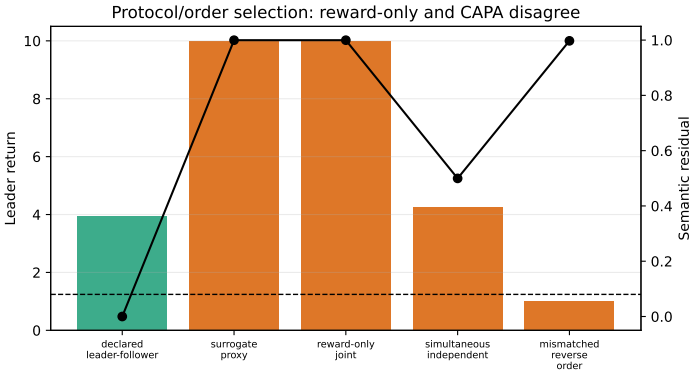

Figure 7: Retained constructed protocol-order comparison; these policies do not constitute independent learned replications.

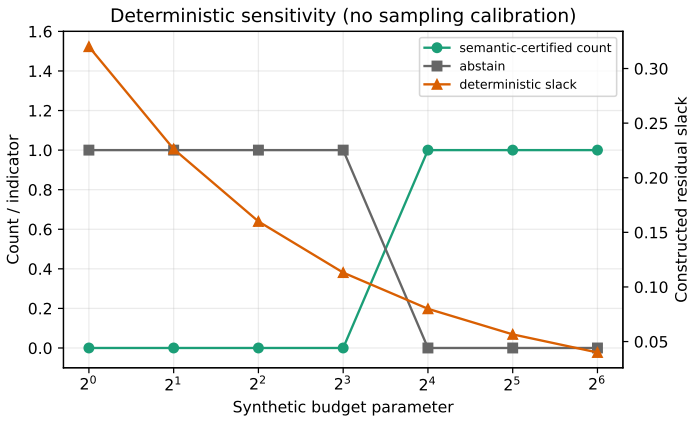

Figure 8: Deterministic slack sensitivity on the six exact candidates: $0.32/\sqrt{B}$ and the count of passing rows, not statistical intervals or empirical confidence coverage.

D PROVENANCE AND REPRODUCIBILITY

Evidence availability. Stored summary arithmetic is distinct from executable reproducibility. The review reconciled 179 summary rows across 26 selected-source, repeated-audit, and candidate-seed files. Retained companion deltas supported raw recomputation for 105 rows; the remaining 74 could not be checked that way from the located files. These checks support aggregation integrity only. The historical executed-code manifest, complete checkpoint/probe hashes, and runtime versions are not fully supplied. Current probe-code repairs do not establish how the old measurements were produced and do not regenerate them. The new audit is separate: every probe retains its ordered paired gains, checkpoint and probe hashes, phase seeds, configuration, and source hashes. Its results do not retroactively validate the historical measurements.

Reproduction scope. The supplement includes table and figure builders, their input statistics, retained paired returns, regression tests, and the fixed checkpoints used by the new audit. The README separates table reconstruction from neural-probe reruns and identifies the required Python packages and upstream HARL revision. Candidate learner training is outside this artifact's reproduction scope. Hash manifests identify delivered files; local paths in copied metadata are anonymized without changing numerical results. Regression tests establish reproducibility on the tested CPU stack, not across arbitrary hardware or library versions.

Reference scope. The approximate-exploitability citation refers to the IJCAI 2022 publication of *Approximate Exploitability: Learning a Best Response*. That work supplies the lower-bound precedent, not a new result of CAPA. The concentration and contraction arguments used here are standard, with explicit proofs and assumptions above; citations to candidate learners do not validate our audit implementation or its statistical intervals.