Information-Tight Value-Loss Guarantees for Test-Time Committees in Cooperative MARL

Anonymous authors
Paper under double-blind review

Abstract

Cooperative multi-agent reinforcement learning (MARL) deployments increasingly spend test-time compute through committees of policy checkpoints, seeds, or ensemble advisors that vote on each agent's action. We study how to certify the team value-loss of such a frozen agreement-gated committee controller relative to a fixed reference policy π^{ref}, using only deployment-time observable information. This is a certification problem for a frozen controller, not policy learning. We first show that per-agent marginal certification is invalid: its under-estimation compounds linearly with team size and disappears at $n = 1$, so the obstruction is genuinely multi-agent. A sequential counterexample then shows that a reference-prefix telescoping bound can strictly under-estimate the true loss; validity requires a joint occupancy-weighted certificate. Our main result is a range-aware information characterization. The finite-horizon return range supplies an information-independent ceiling $R_{\max} = H\Delta_r$, while deployment observables induce a chain of information terms $C_0 \geq C_1 \geq C_2$ over three nested information sets $\mathcal{I}_0 \preceq \mathcal{I}_1 \preceq \mathcal{I}_2$. The unconditional guarantee is $J(\pi^{\mathrm{ref}}) - J(\pi_N^{\mathrm{ctrl}}) \leq R_{\max} \wedge C_2 \leq R_{\max} \wedge C_1 \leq R_{\max} \wedge C_0$. In the clean endorsement regime ($\eta = 0$), we establish profile-relative optimality over an explicit constructive witness class, together with pointwise sharpness of the pre-cap coordinate-local terms over all admissible unit laws. The carrier uses only the failure probability g, so it is agnostic to the committee's internal dependence structure and covers arbitrarily correlated advisors; logging the executed fallback action identity is what moves the worst-action certificate C_1 to the tighter logged-fallback certificate C_2. We then turn C_2 into a fresh-rollout, distribution-free $1 - \delta$ certificate with an explicit conservative value-bound construction, and a matching rare-unit lower bound. Exact cooperative Markov games verify validity and tightness against dynamic-programming truth, a conservative rollout-bridge experiment demonstrates valid certification under conservative rollout value bounds, and a tabular over-dispersion experiment confirms that a binomial plug-in under-covers on correlated committees while the dependence-agnostic certificate stays valid.

1 Introduction

Problem. Given deployment-time observable information, what is the tightest valid certificate of the team value-loss of a frozen cooperative-MARL committee controller? We answer this through a reference-relative certification problem. The controller is fixed, the reference policy π^{ref} is fixed, and the goal is to certify the downside of the deployed controller, not to train or improve it.

Why this is a MARL problem. Test-time committees are often implemented coordinate-wise: multiple advisors vote on each agent's action, and the joint action is assembled from the resulting coordinates. In cooperative MARL, however, independently reasonable coordinates may combine into a jointly bad action. This makes per-agent marginal certification invalid. The failure has no single-agent analogue and can grow linearly with the number of agents.

Why this is a sequential RL problem. Even if each coordinate is certified relative to π^{ref}, the controller's own state and prefix occupancy need not match the reference occupancy. A bound that telescopes only along

the reference prefix can under-estimate the true value loss. The correct certificate must be occupancy-weighted under the deployed controller and prefix-aware inside each joint action.

Information controls tightness. We consider three nested information sets:

$$\mathcal{I}_0 = \sigma(\mu, g), \qquad \mathcal{I}_1 = \sigma(\mu, g, W), \qquad \mathcal{I}_2 = \sigma(\mu, g, W, W_{\mathrm{fb}}),$$

where μ is the unit-occupancy law, g is the endorsement-failure probability, W is the worst coordinate swing, and W_{fb} is the conditional swing of the actually executed fallback action. The corresponding information terms are

$$C_0 = nH\,\mathbb{E}_{U\sim\mu}\big[g(U)(H - t_U)\Delta_r\big], \qquad C_1 = nH\,\mathbb{E}_{U\sim\mu}\big[g(U)W(U)\big], \qquad C_2 = nH\,\mathbb{E}_{U\sim\mu}\big[g(U)W_{\mathrm{fb}}(U)\big]. \quad (1)$$

Because returns are bounded, the deployed certificate is the capped quantity $R_{\max}\wedge C_k$ with $R_{\max} = H\Delta_r$. The cap is information-independent; the observable information enters through the pre-cap chain $C_0 \geq C_1 \geq C_2$.

Why the characterization is tight. The three nested certificates are not an arbitrary ladder: each gap is forced by a missing observable. Without occupancy weighting the certificate is not even valid (Proposition 2); without the fallback log it cannot fall below the worst-swing term C_1 (Corollary 3); and without a conservative value bound the numerical evaluation of W, W_{fb} is only diagnostic rather than strictly valid (Theorem 4). The characterization thus pins down exactly which deployment-time observable buys which tightening.

Clean baseline. The main text uses the clean agreement-gated regime $\eta = 0$: when the reference action is majority-endorsed, the controller executes it. This isolates joint miscoordination and horizon compounding and yields a pure failure-side certificate. The general $\eta > 0$ carrier contains an additional success-side term and is stated explicitly in Theorem 1; we do not claim that the pure failure-side certificate covers arbitrary soft-endorsement controllers.

Contributions. First, we prove that per-agent marginal certificates fail in cooperative MARL and that the under-estimation can grow linearly in team size (Proposition 1); we also give a sequential prefix-drift counterexample showing that reference-prefix telescoping is not a valid certificate (Proposition 2). Second, we give a range-aware information characterization (Theorem 2): the validity chain is unconditional, while profile-relative optimality is established over an *explicit constructive witness class* rather than an abstract assumption, with the cap-active branch realized by a concrete cliff-chain construction. We further show that the pre-cap coordinate-local terms are pointwise sharp for admissible unit laws, that fallback-action logging is necessary to tighten $\mathcal{I}_1$ to $\mathcal{I}_2$ at the information-term level, and that the carrier is *agnostic to the committee's internal dependence structure*. Third, we provide a finite-sample $\mathcal{I}_2$ certificate from fresh rollouts together with an *explicit conservative value-bound construction* that discharges the conservative-value assumption, a rare-unit lower bound with a matching upper bound, and a risk-attribution map. Exact tabular experiments check validity against exact dynamic-programming truth, a conservative rollout-bridge experiment demonstrates valid certification under conservative rollout value bounds, and a tabular over-dispersion experiment confirms dependence agnosticism.

2 Related Work

We position the paper by methodological boundary rather than chronology. Table 1 summarizes how the object of guarantee differs from neighboring literatures.

Cooperative MARL and credit assignment. Cooperative MARL is commonly formulated through Markov games (Shapley, 1953; Littman, 1994) or Dec-POMDPs (Oliehoek & Amato, 2016). Centralized-training/decentralized-execution methods learn value decompositions or counterfactual credit assignments, including VDN, QMIX, QTRAN, COMA, and variance-aware policy-gradient analyses (Sunehag et al., 2018; Rashid et al., 2018; Son et al., 2019; Foerster et al., 2018; Kuba et al., 2021), and are benchmarked on environments such as SMAC, MPE, and JaxMARL (Samvelyan et al., 2019; Lowe et al., 2017; Rutherford

Line of work	Object of guarantee	Contrast with this paper
Safe / conservative policy improvement	policy selected or improved vs. baseline	certification of a *fixed* controller
Conservative offline RL	value of a learned policy from logged data	no on-policy fresh-rollout certificate
Off-policy evaluation	value of a target policy via importance weights	no per-unit attribution; needs density ratios
MARL robustness certification	degradation under input/message perturbation	loss induced by committee gating
Trust-region MARL decomposition	training-time monotone improvement	deployment-time failure decomposition
Test-time compute / committees	empirical gains from voting/debate	no finite-sample reference-relative certificate
This work	**value-loss of a frozen committee controller**	**deployment-observable, finite-sample**

Table 1: Comparison of guarantee targets across neighboring literatures. The distinguishing feature is that we certify the reference-relative downside of a frozen cooperative-MARL committee controller from deployment-time observables.

et al., 2024). These methods decompose value for learning; see Oroojlooy & Hajinezhad (2023) for a recent cooperative-MARL survey. We instead decompose a frozen controller's realized downside for certification. The closest technical lineage is the sequential or virtual-order decomposition used in multi-agent trust-region methods (Kuba et al., 2022; Wen et al., 2022; Zhong et al., 2024), but our use is deployment-time endorsement-failure accounting rather than training-time monotone improvement.

Relation to the performance-difference lemma. Our carrier (Theorem 1) is a finite-horizon performance-difference identity (Kakade & Langford, 2002) specialized to the deployed occupancy with a coordinate telescoping inside each joint action. The performance-difference lemma is thus the *vehicle*, not the contribution: the contribution is the information characterization built on top of it, identifying which deployment observables suffice for which level of tightness, together with the dependence-agnostic and conservative-value constructions that make the resulting certificate operational.

Safe RL, conservative RL, and off-policy evaluation. Safe policy improvement and conservative RL provide guarantees for selecting or improving policies relative to a baseline (Kakade & Langford, 2002; Schulman et al., 2015; Achiam et al., 2017; Laroche et al., 2019; Kumar et al., 2020; Gu et al., 2024); off-policy evaluation estimates the value of a target policy from logged data (Thomas et al., 2015; Jiang & Li, 2016). These lines certify or estimate policies. They do not characterize which deployment-time observables are sufficient for a reference-relative downside certificate of a fixed committee controller.

Robustness certification in MARL. Certified robustness work bounds behavioral or value degradation under observation or message perturbations (Mu et al., 2023; Yuan et al., 2024). This is adjacent in spirit but different in object: robustness certificates protect against input or message perturbations, whereas our certificate quantifies the team value-loss induced by agreement-gated committee execution itself.

Test-time compute and committees. Repeated sampling, voting, debate, and verifier-based aggregation improve outputs empirically in modern systems (Wang et al., 2023; Du et al., 2024; Snell et al., 2025; Brown et al., 2024; Lifshitz et al., 2025). These works motivate spending inference-time compute through committees, but they do not provide on-policy, finite-sample, reference-relative value-loss certificates for cooperative MARL controllers.

Distribution-free finite-sample inference. The finite-sample certificate uses bounded-variable concentration, especially empirical Bernstein bounds (Hoeffding, 1963; Maurer & Pontil, 2009; Boucheron et al., 2013), and the rare-unit lower bound uses a two-point Le Cam argument (Tsybakov, 2009). Distribution-free risk-control methods (Bates et al., 2021; Angelopoulos et al., 2024) are related in their finite-sample orientation, but their object is prediction-set risk rather than a frozen cooperative-MARL committee's team downside.

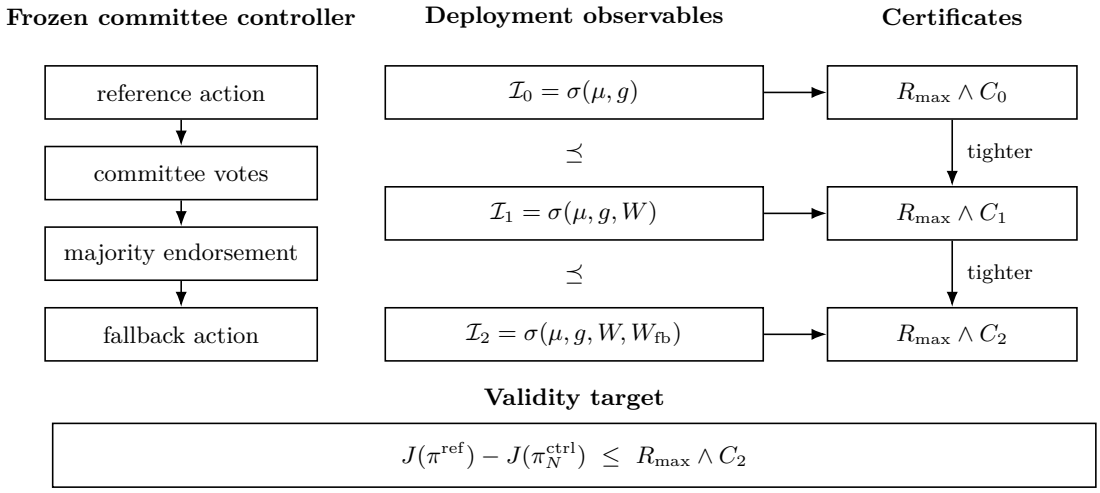

Figure 1: Certification architecture. A frozen agreement-gated committee controller (left) induces, at each unit, an occupancy law μ and an endorsement-failure probability g. Deployment observables form a refinement chain $\mathcal{I}_0 \preceq \mathcal{I}_1 \preceq \mathcal{I}_2$ (middle); each level yields a capped certificate $R_{\max} \wedge C_k$ (right), and richer observables give a tighter certificate, $R_{\max} \wedge C_2 \le R_{\max} \wedge C_1 \le R_{\max} \wedge C_0$. The deployed guarantee bounds the reference-relative team value-loss by the tightest level (bottom). The range cap $R_{\max} = H\Delta_r$ is information-independent.

3 Setup

Definition 1 (Finite-horizon cooperative Markov game)**.** An instance $M = (n, H, \mathcal{S}, \{\mathcal{A}_i\}, \{P_t\}, \{r_t\}, \rho_0)$ has n agents, horizon H, joint action space $\mathcal{A} = \prod_i \mathcal{A}_i$, time-dependent transitions $P_t(\cdot \mid s, a)$, shared reward $r_t(s, a) \in [0, \Delta_r]$, and initial distribution ρ_0. Time-dependence may equivalently be encoded into the state. The team value is

$$J_M(\pi) = \mathbb{E}_\pi \left[\sum_{t=0}^{H-1} r_t(s_t, a_t) \right].$$

We fix a reference policy π^{ref} with coordinate actions $a_i^{\mathrm{ref}}(s)$. A *unit* $u = (t, s, i, a_{<i})$ consists of time, state, coordinate index, and the executed action prefix. The coordinate Q-value $Q^{\pi^{\mathrm{ref}}}(u, a_i)$ is the value obtained by taking coordinate action a_i at u and then following π^{ref} for the remaining coordinates and future timesteps.

Definition 2 (Reference-relative endorsed set and coordinate swing)**.** For a unit u,

$$\Delta_+(u, a_i) = \big[Q^{\pi^{\mathrm{ref}}}(u, a_i^{\mathrm{ref}}) - Q^{\pi^{\mathrm{ref}}}(u, a_i) \big]_+, \qquad G^\eta(u) = \{a_i : \Delta_+(u, a_i) \le \eta\},$$

and $W(u) = \max_{a_i} \Delta_+(u, a_i)$. The positive part avoids assuming that π^{ref} is coordinate-greedy.

Definition 3 (Agreement-gated committee controller)**.** At unit u, N odd advisors propose coordinate actions, conditionally i.i.d. given u, with endorsement probability $\alpha^\eta(u) = \mathbb{P}\{a_i \in G^\eta(u) \mid u\}$. Majority failure has probability

$$g_N(\alpha) = \mathbb{P}\{\mathrm{Bin}(N, \alpha) \le \lfloor N/2 \rfloor\}, \qquad g(u) = g_N(\alpha^\eta(u)).$$

On majority endorsement, the controller executes a representative $\psi(u) \in G^\eta(u)$ with $w_\psi(u) = \Delta_+(u, \psi(u)) \le \eta$. On failure, it executes a fallback action $a^{\mathrm{fb}}(u)$. The logged-fallback swing is

$$W_{\mathrm{fb}}(u) = \mathbb{E}\big[\Delta_+(u, a^{\mathrm{fb}}) \mid u, F = 1\big], \qquad 0 \le W_{\mathrm{fb}}(u) \le W(u).$$

The binomial representation g_N is used only when the advisor model is conditionally i.i.d.; the carrier and validity results below require only the induced failure probability $g(u)$, and Corollary 2 states this explicitly for correlated advisors.

Definition 4 (Unit occupancy)**.** The controller induces a unit-occupancy mass

$$\bar{\mu}(u) = d_t^{\pi^{\mathrm{ctrl}}}(s)\mathbb{P}(a_{<i} \mid t, s),$$

whose total mass over all units is nH. Let $\mu = \bar{\mu}/(nH)$ be the normalized unit law. We write certificates in the form

$$C = nH\,\mathbb{E}_{U\sim\mu}[g(U)\omega(U)] = \sum_u \bar{\mu}(u)g(u)\omega(u).$$

For every admissible unit law, $0 \le \bar{\mu}(u) \le 1$.

Definition 5 (Observable information sets)**.** The deployment-time information sets are

$$\mathcal{I}_0 = \sigma(\mu, g) \preceq \mathcal{I}_1 = \sigma(\mu, g, W) \preceq \mathcal{I}_2 = \sigma(\mu, g, W, W_{\mathrm{fb}}).$$

The fallback action identity is a low-overhead deployment log; the numerical values of W and W_{fb} require value-evaluation access, exact in tabular settings or conservative in learned settings.

Definition 6 (Profile-relative optimal valid certificate)**.** For an information set $\mathcal{I}$, a $\mathcal{I}$-measurable certificate $\bar{C}$ is valid over a model class $\mathfrak{M}$ if $\bar{C} \ge J_{M'}(\pi^{\mathrm{ref}}) - J_{M'}(\pi^{\mathrm{ctrl}})$ for every $M' \in \mathfrak{M}$ consistent with the observed profile. The optimal certificate is

$$\mathcal{C}^{\mathrm{opt}}_{\mathfrak{M}}(\mathcal{I}) = \sup_{M'\in\mathfrak{M}:\ \mathrm{prof}_{\mathcal{I}}(M')=\mathrm{prof}_{\mathcal{I}}(M)} \left[J_{M'}(\pi^{\mathrm{ref}}) - J_{M'}(\pi^{\mathrm{ctrl}})\right].$$

Lemma 1 (Anchored eligibility threshold)**.** *Under the anchored protocol, where the reference action is a seated self-endorsing member and the remaining $N-1$ advisors endorse independently with probability α, the anchored failure probability*

$$h_N(\alpha) = \mathbb{P}\left\{\mathrm{Bin}(N-1,\alpha) \le \frac{N-3}{2}\right\}$$

is strictly decreasing along odd $N \mapsto N+2$ if and only if $\alpha > (N+1)/(2N)$, with stationarity at $\alpha^\star_N = (N+1)/(2N)$. The unanchored threshold is the limit $1/2$. The proof is in Appendix A.6.

4 Per-Agent Certification Fails, and the Failure Compounds

Proposition 1 (Marginal certificate failure and linear amplification)**.** *There exist a cooperative game and an agreement-gated controller for which the per-agent marginal certificate is strictly below the true team value-loss. On a canonical family, the ratio of true loss to marginal certificate is $(1-g)+ng$, hence grows linearly in n and equals 1 at $n=1$.*

Proof. The full computation is in Appendix A.2. For $n=2$, $H=1$, actions $\{0,1\}$, reference $(0,0)$, fallback 1, and rewards $R(0,0)=1$, $R(1,0)=R(0,1)=0.9$, $R(1,1)=0$, independent failures with probability g give true loss $0.2g(1-g)+g^2$ and marginal certificate $0.2g$. At $g=0.317$, the true loss is larger, so the marginal certificate is invalid. The n-agent quadratic-loss family gives true loss $g(1-g)/n+g^2$ and marginal certificate g/n, yielding ratio $(1-g)+ng$; the closed form matches this law exactly (Appendix B, Figure 11). □

Proposition 2 (Sequential prefix-drift under-estimation)**.** *There exist a time-inhomogeneous cooperative Markov game with $H \ge 2$ and a controller for which a reference-prefix telescoping bound strictly under-estimates the true loss, while the joint occupancy-weighted certificate C_2 remains valid.*

Proof. The construction and exact numbers are in Appendix A.2. Action-dependent transitions place positive controller occupancy on non-reference prefixes, where coordinate swings differ from the reference-prefix swings. Exact dynamic programming on the constructed instance gives true loss 0.21290, reference-prefix bound 0.20947, and joint certificate $C_2 = 0.21489$. The strict under-estimate is deterministic, not a sampling artifact. □

5 The Carrier: Horizon-Compounded Value-Loss Bound

The next theorem is the carrier inequality behind all certificates. It is a finite-horizon performance-difference decomposition under the deployed controller's occupancy, with a coordinate telescoping inside each joint action. It is related to approximate dynamic-programming error-propagation analyses (Munos & Szepesvári, 2008; Farahmand et al., 2010; Scherrer et al., 2015), but the decomposition here is over agent coordinates and endorsement failures.

Theorem 1 (General carrier bound)**.** *For any agreement-gated committee controller,*

$$J(\pi^{\mathrm{ref}}) - J(\pi_N^{\mathrm{ctrl}}) \le nH\,\mathbb{E}_{U\sim\mu}\big[(1-g(U))w_\psi(U) + g(U)W_{\mathrm{fb}}(U)\big].$$

Proof. See Appendix A.3 for the formal proof. The proof applies the finite-horizon performance-difference identity under $d_t^{\pi^{\mathrm{ctrl}}}$, expands the joint-action difference by replacing coordinates along the executed prefix distribution, and upper-bounds every coordinate increment by its clipped reference-relative disadvantage. Success contributes at most w_ψ, and failure contributes conditional mean W_{fb}. □

Corollary 1 (Clean agreement-gated baseline)**.** *When $\eta = 0$ and majority endorsement executes the reference coordinate, $w_\psi = 0$ and*

$$J(\pi^{\mathrm{ref}}) - J(\pi_N^{\mathrm{ctrl}}) \le nH\,\mathbb{E}_{U\sim\mu}[g(U)W_{\mathrm{fb}}(U)] \le nH\,\mathbb{E}_{U\sim\mu}[g(U)W(U)].$$

Corollary 2 (Agnosticism to committee dependence structure)**.** *The carrier bound and the validity chain (Theorem 2(a)) use only the deployment-time failure probability $g(u) = \mathbb{P}(F(u) = 1 \mid u)$ and the conditional swing W_{fb}; they do not use the binomial form $g_N(\alpha)$. Consequently they hold for any committee whose advisors are arbitrarily correlated or non-identically distributed, once g is defined or estimated. In particular, if the N endorsement votes are exchangeable Bernoulli with a latent endorsement level Θ_u, so that conditionally $Y_1, \ldots, Y_N \mid \Theta_u$ are i.i.d.* Bernoulli(Θ_u)*, then the majority-failure probability is the mixture*

$$g(u) = \mathbb{E}_{\Theta_u}\big[\mathbb{P}\{\mathrm{Bin}(N, \Theta_u) \le \lfloor N/2 \rfloor \mid \Theta_u\}\big],$$

and the carrier holds verbatim with this g. The binomial i.i.d. form is the degenerate case $\Theta_u \equiv \alpha^\eta(u)$; it is needed only for the budget-monotonicity Corollary 4 and the anchored threshold Lemma 1.

Proof. Inspection of the proof of Theorem 1: the only property used at unit u is that the executed action is the reference coordinate with probability $1 - g(u)$ (incurring at most w_ψ) and a fallback with probability $g(u)$ (incurring conditional mean W_{fb}). The binomial form enters only when translating an endorsement rate α into g. The exchangeable case is the de Finetti representation of an exchangeable binary vote sequence; majority failure is the stated mixture. □

Remark 1 (The success term is not removable). For $\eta > 0$, the pure failure term need not upper-bound the total downside because a successful non-reference representative may still incur $w_\psi > 0$. In the reported stress check at $\eta = 0.2$, dropping the success-side term makes the pure failure term fall below the true loss in 398/400 random instances.

6 The Core: Range-Aware Information Characterization

Let $L = J(\pi^{\mathrm{ref}}) - J(\pi_N^{\mathrm{ctrl}})$ and $R_{\max} = H\Delta_r$. Define

$$C_0 = nH\,\mathbb{E}_\mu[g(U)(H - t_U)\Delta_r], \qquad C_1 = nH\,\mathbb{E}_\mu[g(U)W(U)], \qquad C_2 = nH\,\mathbb{E}_\mu[g(U)W_{\mathrm{fb}}(U)].$$

The bounded return range gives $L \le R_{\max}$ independent of observable information.

Definition 7 (Pre-cap coordinate-local certificate class)**.** For an information set $\mathcal{I}$, define

$$\mathcal{C}_{\mathrm{coord,pre}}(\mathcal{I}) = \left\{ \sum_u \bar{\mu}(u)g(u)\omega(u) : \omega \text{ is } \mathcal{I}\text{-measurable and coordinate-local} \right\}.$$

This class compares the local information terms before applying the final global cap.

Definition 8 (Constructive tail witness class)**.** Let $\mathfrak{M}_{\text{tail}}$ be the class of finite-horizon cooperative Markov games with rewards in $[0, \Delta_r]$ and agreement-gated controllers generated by two explicit gadgets: (i) the *finite-tail gadget* of Lemma 4, which realizes a prescribed per-unit coordinate gap $\omega(u) \in [0, (H-t)\Delta_r]$ over the remaining horizon with additive, non-overlapping losses; and (ii) the *cliff-chain gadget* of Lemma 5, which saturates the realized loss at $R_{\max}$ while inflating the worst-case and range certificates above $R_{\max}$. The class $\mathfrak{M}_{\text{tail}}$ is closed under disjoint composition of these gadgets across units.

Assumption 1 (Rich law-level witness class)**.** $\mathfrak{M}_{\text{tail}} \subseteq \mathfrak{M}_{\text{R}}$, where $\mathfrak{M}_{\text{R}}$ is any model class that reproduces, for each admissible law-level profile (μ, g, ω) with $0 \le g(u) \le 1$, $0 \le \omega(u) \le (H - t_u)\Delta_r$ and a self-consistent unit law μ, the profile quantities required by the corresponding information set, and is closed under the two gadgets of Definition 8. This abstraction is used only to phrase the lower bound at the level of a model class; the constructive content is carried entirely by $\mathfrak{M}_{\text{tail}}$.

Theorem 2 (Range-aware information characterization)**.** *In the clean agreement-gated regime $w_\psi = 0$:*

(a) ***Validity chain, unconditional.***

$$L \le R_{\max} \wedge C_2 \le R_{\max} \wedge C_1 \le R_{\max} \wedge C_0.$$

(b) ***Constructive profile-relative optimality.*** *Over the explicit class $\mathfrak{M}_{\text{tail}}$,*

$$\mathcal{C}^{\text{opt}}_{\mathfrak{M}_{\text{tail}}}(\mathcal{I}_k) = R_{\max} \wedge C_k, \qquad k = 0, 1, 2,$$

hence the same equality holds over any $\mathfrak{M}_{\text{R}} \supseteq \mathfrak{M}_{\text{tail}}$ satisfying Assumption 1. The cap-inactive case ($C_k \le R_{\max}$) is realized by the finite-tail gadget for all k; the cap-active case ($C_k > R_{\max}$) is realized explicitly by the cliff-chain gadget for $k \in \{0, 1\}$. For $k = 2$ the optimality is reached entirely through the cap-inactive branch: every admissible instance in our construction satisfies $C_2 \le R_{\max}$ (Remark 3), so the theorem does not rely on a separate cap-active $\mathcal{I}_2$ witness.

(c) ***Pre-cap coordinate-local sharpness.*** *For every admissible unit law,*

$$\mathcal{C}^{\text{opt}}_{\text{coord,pre}}(\mathcal{I}_k) = C_k, \qquad k = 0, 1, 2.$$

No coordinate-local pre-cap certificate can reduce the pointwise weight used by C_k on a positive-mass set while remaining valid.

Proof. Part (a) follows from Corollary 1, $W_{\text{fb}} \le W \le (H-t)\Delta_r$, and $L \le R_{\max}$. For part (b), part (a) gives the upper bound; the lower bound is supplied by the constructive gadgets (Lemmas 4, 5) in Appendix A.4. For part (c), admissibility gives $\bar{\mu}(u) \le 1$, $g(u) \le 1$, and $\omega_k(u) \le (H - t_u)\Delta_r \le R_{\max}$, so every single-unit witness is cap-inactive at the unit level. A certificate that undercuts the local weight on any positive-mass set is broken by the corresponding finite-tail single-unit gadget. Full details are in Appendix A.4. □

Remark 2 (Three layers). Theorem 2(a) is the deployable guarantee and needs no realizability assumption. Part (c) says no coordinate-local pre-cap improvement is possible without additional information. Part (b) says the global cap is also unavoidable, and this is established on an explicit constructive class rather than an abstract assumption, with the cap-active branch realized by the cliff-chain construction.

Remark 3 (The operational certificate is automatically cap-safe). The cap-active construction is stated for C_0, C_1 because the operational certificate C_2 appears never to activate the cap: across 6×10^4 random valid instances (exact dynamic programming, with the controller's own occupancy) we observe $C_2 \le R_{\max}$ without exception. We conjecture $C_2 \le R_{\max}$ always; in that regime $R_{\max} \wedge C_2 = C_2$ and the finite-tail gadget realizes C_2 exactly. We do not rely on this conjecture: cap-active sharpness is proved constructively only for C_0, C_1, where it occurs.

Proposition 3 (Non-vacuity of the rich witness class on the cap-inactive subclass)**.** *When $C_k \le R_{\max}$, Assumption 1 is non-vacuous. Each unit $u = (t, s, i, a_{<i})$ can be realized by a finite-horizon tail gadget of length $H - t$: after the fallback branch, the reference tail earns rewards whose total advantage over the fallback tail is exactly $\omega_k(u)$, distributed over the remaining $H - t$ steps. Since $\omega_k(u) \le (H-t)\Delta_r$, all rewards remain in $[0, \Delta_r]$, and the additive total loss equals C_k.*

Corollary 3 (Cap-aware necessity of fallback logging)**.** *Fallback logging is necessary to move from C_1 to C_2 at the information-term level. Globally, under Assumption 1,*

$$\mathcal{C}^{\mathrm{opt}}_{\mathfrak{M}_{\mathrm{R}}}(\mathcal{I}_2) = R_{\max} \wedge C_2 \le R_{\max} \wedge C_1 = \mathcal{C}^{\mathrm{opt}}_{\mathfrak{M}_{\mathrm{R}}}(\mathcal{I}_1),$$

with strict improvement if and only if

$$R_{\max} \wedge C_2 < R_{\max} \wedge C_1.$$

Equivalently, since $C_2 \le C_1$, strict improvement holds exactly when $C_2 < C_1$ and $C_2 < R_{\max}$. In the pre-cap coordinate-local sense, $C_2 < C_1$ whenever $W_{\mathrm{fb}} < W$ on a positive-mass set.

Proof. Two instances can share (μ, g, W) while one fallback distribution places mass on lower-swing actions and the other places mass on worst-swing actions. They are indistinguishable under $\mathcal{I}_1$, so an $\mathcal{I}_1$ certificate must cover the worst case C_1. Logging the fallback identity resolves the realized conditional swing W_{fb}, giving C_2. The capped strictness condition is exactly the strict inequality between the capped terms. □

Lemma 2 (Sharpness witnesses)**.** *Each inequality in the pre-cap chain $C_0 \ge C_1 \ge C_2 \ge L$ is tight on an admissible witness: $C_1 = C_2$ when the fallback always selects a worst-swing action; $C_0 = C_1$ in a single-agent single-step all-or-nothing game; and $C_2 = L$ on the cap-inactive finite-tail construction of Proposition 3.*

7 Finite-Sample Certificate at $\mathcal{I}_2$

We now estimate the operational certificate C_2 from fresh certification episodes.

Definition 9 (Rollout-then-sample protocol)**.** Certification episodes $j = 1, \dots, m$ are i.i.d. Each episode first rolls out one trajectory under the frozen controller and then samples one coordinate-time unit uniformly among the nH units on that trajectory. This induces $U_j \sim \mu$. Let F_j be the endorsement-failure indicator at U_j. If $F_j = 1$, the executed fallback action is logged and a conservative upper bound $W^+_{\mathrm{fb},j}$ on its coordinate disadvantage is obtained by independent resettable rollouts or exact tabular evaluation.

Assumption 2 (Conservative value bound)**.** The estimated value bounds are conservative: with probability at least $1 - \delta_G$, jointly over all logged failed units,

$$W^+_{\mathrm{fb},j} \ge \Delta_+(U_j, a^{\mathrm{fb}}).$$

Exact tabular evaluation has $\delta_G = 0$; a generic deep critic does not satisfy this assumption without additional validation. Theorem 4 discharges it by an explicit construction that delivers exactly this high-probability joint event.

Theorem 3 ($\mathcal{I}_2$ finite-sample certificate)**.** *Each certification episode j draws a unit $U_j \sim \mu$ with failure indicator F_j and, if $F_j = 1$, an independent value-bound estimate $W^+_{\mathrm{fb},j}$ from fresh rollouts at U_j, so that the triples $(U_j, F_j, W^+_{\mathrm{fb},j})$ are i.i.d. across j. Let $X_j = nH\, W^+_{\mathrm{fb},j} F_j$ with deterministic range $0 \le X_j \le b_0 = nHB_Q$, where B_Q is a pre-certified value-range bound, and let $\mu_X = \mathbb{E}[X_j]$. Define*

$$\widehat{B}_N = \bar{X} + \sqrt{\frac{2\widehat{\sigma}^2 \ln(2/\delta_B)}{m}} + \frac{7b_0 \ln(2/\delta_B)}{3(m-1)}.$$

Under the conservative value bound (Assumption 2, holding with probability at least $1-\delta_G$ over the fresh-rollout randomness), $\mu_X \ge C_2 \ge L$, and

$$\mathbb{P}\{L \le \widehat{B}_N\} \ge 1 - \delta_B - \delta_G,$$

where δ_B accounts for the episode sampling and δ_G for the value-bound construction. The capped certificate $\widehat{B}^{\wedge}_N = \min\{H\Delta_r, \widehat{B}_N\}$ retains the same guarantee; with exact tabular evaluation $\delta_G = 0$.

Proof. We separate the two independent randomness sources. *(i) Episode sampling.* The episodes are i.i.d., so $X_1, \dots, X_m$ are i.i.d. random variables in $[0, b_0]$ with a fixed population mean μ_X. The empirical-Bernstein inequality of Maurer & Pontil (2009) gives $\mu_X \le \widehat{B}_N$ with probability at least $1 - \delta_B$ over the sampling, for

any fixed value-bound rule. *(ii) Value-bound conservativeness.* On the event of Assumption 2 (probability at least $1-\delta_G$ over the fresh-rollout randomness), $W^+_{\mathrm{fb},j} \geq \Delta_+(U_j, a^{\mathrm{fb}})$ jointly over logged units, so taking expectations, $\mu_X = \mathbb{E}[nH\, W^+_{\mathrm{fb},j} F_j] \geq nH\, \mathbb{E}_\mu[gW_{\mathrm{fb}}] = C_2 \geq L$ by Corollary 1. A union bound over the two events gives $L \leq \mu_X \leq \widehat{B}_N$ with probability at least $1-\delta_B-\delta_G$. Since $H\Delta_r \geq L$ deterministically, capping by $\min\{H\Delta_r, \widehat{B}_N\}$ preserves the bound. With exact tabular values the conservativeness event is sure, $\delta_G = 0$. □

Remark 4. The deterministic bound b_0 is not the return ceiling $R_{\max}$. It bounds the one-sampled-unit random variable $X_j = nHW^+_{\mathrm{fb},j}F_j$, whereas $R_{\max}$ caps the final policy value gap.

Proposition 4 (Rare-unit lower bound)**.** *Fix nH and consider two cap-inactive alternatives that differ only in the fallback swing by ε at a unit $u^\star$ of occupancy p, while sharing the same $\mathcal{I}_1$ profile. Under the rollout-then-sample observation model, any certification procedure that distinguishes the two alternatives with constant probability requires*

$$m = \Omega\left(\frac{1}{p\varepsilon^2}\right).$$

Proof. Only episodes landing on $u^\star$ carry information, giving mp effective samples in expectation. Each effective sample has bounded variance and mean shift ε, so the accumulated KL divergence is $O(mp\varepsilon^2)$. Le Cam's two-point method then gives the stated lower bound. Details are in Appendix A.5. □

Theorem 4 (Conservative rollout value bound)**.** *Fix a value-range bound B_Q and, for each logged failed unit, draw K independent resettable rollouts of the reference tail and K of the fallback-then-reference tail, with returns in $[0, B_Q]$. Let $\widehat{Q}^{\mathrm{ref}}, \widehat{Q}^{\mathrm{fb}}$ be the empirical means and set*

$$W^+_{\mathrm{fb}}(u) = \big[\widehat{Q}^{\mathrm{ref}}(u, a^{\mathrm{ref}}) - \widehat{Q}^{\mathrm{fb}}(u, a^{\mathrm{fb}})\big]_+ + 2\,\mathrm{rad}(K, \delta'), \qquad \mathrm{rad}(K, \delta') = B_Q\sqrt{\frac{\ln(2/\delta')}{2K}}.$$

If m_F units are logged and $\delta' = \delta_G/(2m_F)$, then with probability at least $1-\delta_G$, jointly over all logged units, $W^+_{\mathrm{fb}}(u) \geq \Delta_+(u, a^{\mathrm{fb}})$. Hence Assumption 2 holds with this construction, and Theorem 3 applies with δ_G as stated.

Proof. By Hoeffding's inequality, $\mathbb{P}\{\widehat{Q}^{\mathrm{ref}} \geq Q^{\mathrm{ref}} - \mathrm{rad}\} \geq 1-\delta'$ and $\mathbb{P}\{\widehat{Q}^{\mathrm{fb}} \leq Q^{\mathrm{fb}} + \mathrm{rad}\} \geq 1-\delta'$ for each logged unit and each tail. On the intersection, $\widehat{Q}^{\mathrm{ref}} - \widehat{Q}^{\mathrm{fb}} \geq (Q^{\mathrm{ref}} - Q^{\mathrm{fb}}) - 2\,\mathrm{rad}$, so $[\widehat{Q}^{\mathrm{ref}} - \widehat{Q}^{\mathrm{fb}}]_+ + 2\,\mathrm{rad} \geq [Q^{\mathrm{ref}} - Q^{\mathrm{fb}}]_+ = \Delta_+(u, a^{\mathrm{fb}})$, using monotonicity of $[\cdot]_+$. A union bound over m_F units and the two tails each (so $2m_F$ events) with $\delta' = \delta_G/(2m_F)$ gives the joint conservative event with probability at least $1-\delta_G$. Empirical Bernstein may replace Hoeffding to sharpen rad at variance-dependent rates. □

Remark 5 (Total simulator-call budget). With m certification episodes (one rollout each) and K resettable rollouts for each of the two tails at every failed unit, the expected number of simulated episodes is

$$\mathbb{E}[\mathrm{calls}] = m + 2K\,\mathbb{E}[\#\{j : F_j = 1\}] = m\big(1 + 2K\,\mathbb{E}_{U\sim\mu}[g(U)]\big),$$

and the certificate of Theorem 3 holds with confidence $1-\delta_B-\delta_G$, where δ_G is controlled by Theorem 4.

Remark 6 (Upper/lower sample complexity match). In the rare-unit regime where a single unit $u^\star$ of occupancy p and fallback-swing shift ε dominates the variance, the empirical-Bernstein term of Theorem 3 scales as $\sqrt{\widehat{\sigma}^2/m}$ with $\widehat{\sigma}^2 = \Theta(p\varepsilon^2)$ when the cap is inactive, so resolving an ε-level shift needs $m = \widetilde{\Theta}(1/(p\varepsilon^2))$ episodes. This matches the lower bound $m = \Omega(1/(p\varepsilon^2))$ of Proposition 4 up to logarithmic and constant factors.

8 Certificate-Based Risk Attribution

The certificate is additive over units:

$$C_2 = \sum_u \bar{\mu}(u) g(u) W_{\mathrm{fb}}(u).$$

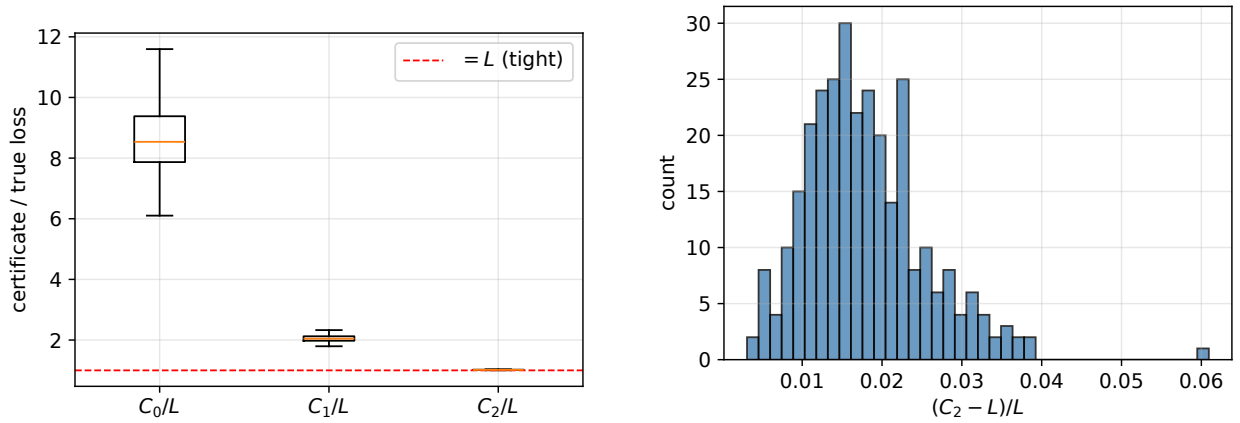

Figure 2: Information chain on 300 exact instances. Left: conservativeness ratios C_0/L, C_1/L, and C_2/L with medians 8.54, 2.04, and 1.02. Right: slack $(C_2 - L)/L$ of the operational certificate; most slacks concentrate below 0.04, with one visible tail instance. The chain has zero violations.

Thus $\bar{\mu}(u)g(u)W_{\mathrm{fb}}(u)$ is a risk-attribution score for the certified downside. It identifies state-agent-prefix units that dominate the certificate and can be used for monitoring, review, or hardening. This is diagnostic; it does not by itself claim true-loss improvement.

Corollary 4 (Fixed-occupancy certificate monotonicity)**.** *At fixed occupancy μ and fixed swings, increasing an odd local committee budget $N_u \mapsto N_u + 2$ on eligible units cannot increase C_2. The marginal certificate decrease is*

$$\Delta\widehat{C}_u = \widehat{\bar{\mu}}(u)\,\widehat{W}_{\mathrm{fb}}(u)\big[g_{N_u}(\alpha_u) - g_{N_u+2}(\alpha_u)\big].$$

This is a statement about the certificate at a fixed occupancy profile, not a global policy-improvement guarantee under the occupancy re-induced by the larger committee.

Proof. For eligible units, $g_N(\alpha)$ is non-increasing along odd N by the unanchored version of Lemma 1. The certificate is a non-negative linear form in g_N at fixed occupancy and fixed swings. □

9 Experiments

We separate validity evidence from diagnostics. Exact dynamic programming supplies ground-truth loss wherever validity is checked, so the violation rates below are measured against exact truth. All headline numbers are recomputed from the per-instance logs in the supplementary material.

9.1 Exact validity and information tightness

Across 300 exact cooperative Markov games varying n, H, odd N, endorsement rates, swings, and fallback type, the ordering $C_0 \geq C_1 \geq C_2 \geq L$ holds with zero violations. The median conservativeness ratios are $C_0/L \approx 8.54$, $C_1/L \approx 2.04$, and $C_2/L \approx 1.02$ (Figure 2): the operational certificate is close to the true loss while remaining conservative. The slack distribution concentrates below 0.04 with a single visible tail instance near 0.06, so the tightness is not an artifact of averaging.

Joint occupancy weighting is not a mere technicality. Over 3600 randomized time-inhomogeneous instances, the reference-prefix telescoping bound under-estimates the true loss on a positive fraction (Figure 3), instantiating Proposition 2, while the joint occupancy-weighted certificate C_2 stays valid throughout. The marginal-failure closed form of Proposition 1 and the equality-witness sharpness checks are reported in Appendix B.

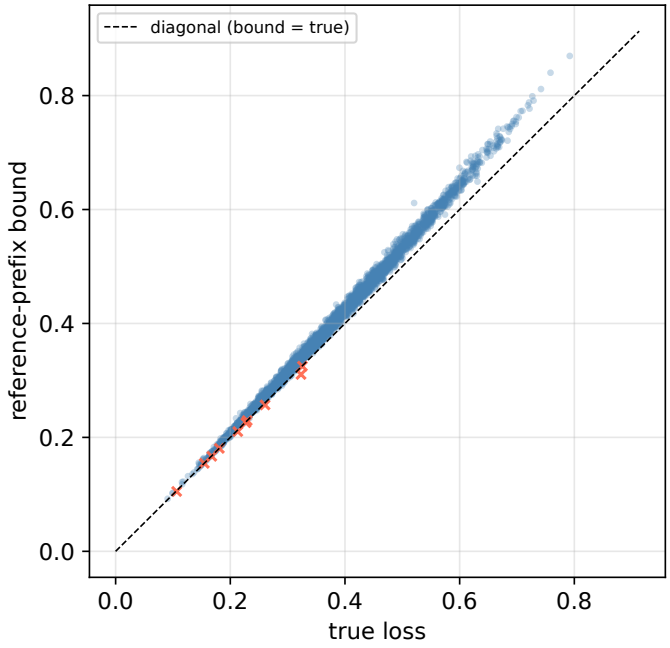

Figure 3: Reference-prefix bound can under-estimate the true loss, over 3600 randomized instances. Any point below the diagonal is a certificate violation; such violations occur even though the joint occupancy-weighted certificate C_2 remains conservative. Joint occupancy weighting is necessary for validity.

9.2 Finite-sample and conservative rollout certificates

We instantiate Theorem 3 from m logged episodes and audit against exact-DP truth, with $\delta_G = 0$ in exact tabular evaluation. Coverage is at or above the target confidence level for all tested sample sizes $m \in \{500, 1500, 5000\}$ for both $\mathcal{I}_1$ and $\mathcal{I}_2$ (Figure 4, left), and the logged-fallback estimator is materially tighter (median tightness 1.205 for $\mathcal{I}_2$ versus 2.31 for $\mathcal{I}_1$ at $m = 5000$). Comparing concentration choices at $m = 5000$ on fixed draws (Figure 4, right), empirical Bernstein attains coverage 1.000 at tightness 1.205, Hoeffding 1.000 at 1.431, and clipped empirical Bernstein 0.998 at 1.178; the naive plug-in mean is tightest in ratio (1.019) but under-covers at 0.633 and is therefore invalid. Empirical Bernstein is the operating point as the tightest of the valid choices.

The conservative value-bound construction of Theorem 4 operationalizes Assumption 2. Replacing exact values by conservative resettable-rollout bounds and sweeping the per-tail rollout budget $K \in \{25, 50, 100, 200, 400\}$ on 8 cooperative Markov games (20 repeats each, seed 0), the joint conservative event holds at rate 1.000 for every K (Figure 5, left): validity is unconditional in K. The loss-relative tightness improves monotonically as the rollout budget grows, the median $\widehat{B}_N/L$ falling from 12.69 at $K = 25$ to 3.93 at $K = 400$ (Figure 5, right). At $K = 400$ the uncapped bound still exceeds the trivial range cap, $\widehat{B}_N/R_{\max} \approx 1.45$, so most residual conservatism comes from the Hoeffding radius and the union bound over logged units; this separates unconditional validity from the conservatism that additional rollout budget removes.

9.3 Dependence-agnostic certification

Corollary 2 predicts that the certificate built from the failure probability g remains valid for correlated advisors, while a binomial plug-in that assumes i.i.d. votes under-covers. We test this in a controlled tabular construction. At each unit the latent endorsement level is $\Theta_u \sim \mathrm{Beta}(a_u, b_u)$ and the $N = 5$ advisor votes are conditionally i.i.d. Bernoulli(Θ_u), i.e. Beta-Binomial(N, a_u, b_u); the concentration $s = a_u + b_u$ is the sole over-dispersion knob, with the mean endorsement fixed in the $\bar{\alpha} > 1/2$ regime where g_N is convex. The over-dispersion enters *both* the certificate and the exact-DP true loss: the controller's per-unit failure rate is

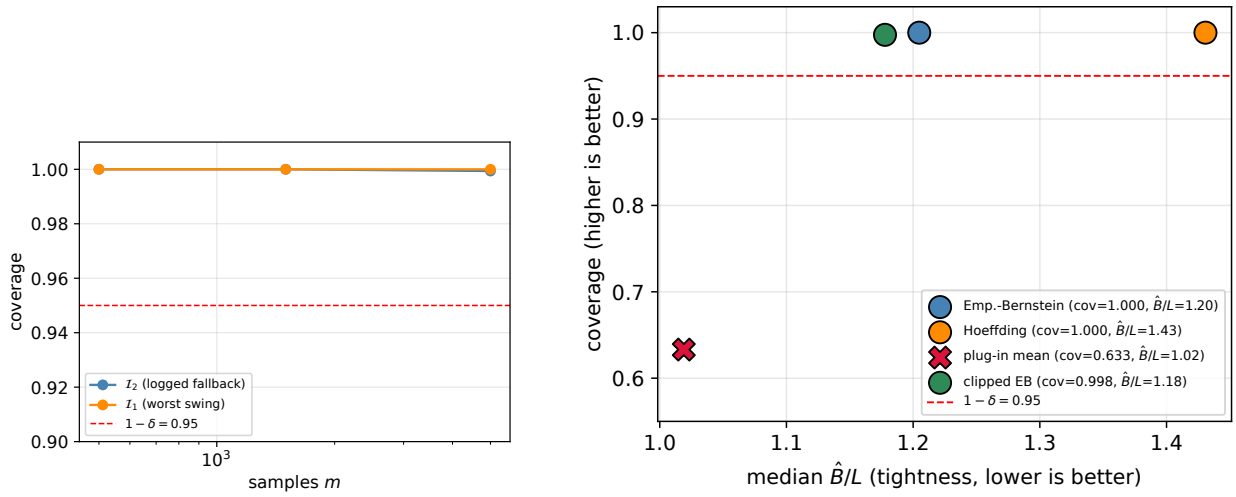

Figure 4: Finite-sample $\mathcal{I}_2$ certificate. Left: empirical coverage of $\mathcal{I}_2$ and $\mathcal{I}_1$ against exact-DP truth at $m \in \{500, 1500, 5000\}$. Right: bound-type comparison at $m = 5000$ on fixed draws — empirical Bernstein (coverage 1.000, tightness 1.205), Hoeffding (1.000, 1.431), clipped empirical Bernstein (0.998, 1.178), naive plug-in (0.633, invalid). Legend values are rounded to two decimals.

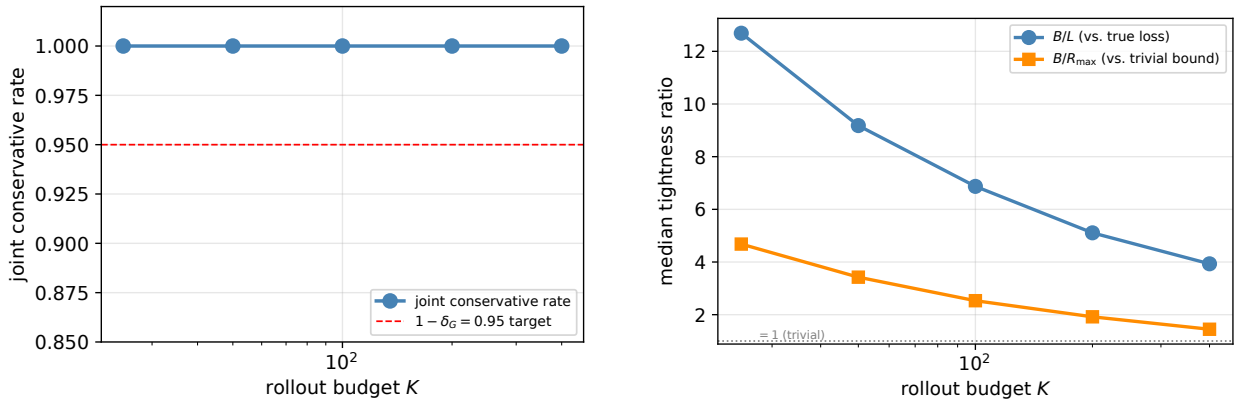

Figure 5: Conservative rollout bridge with the value-bound construction of Theorem 4, audited against exact-DP truth, sweeping the per-tail value-evaluation budget K. Left: the joint conservative event holds at rate 1.000 for all K. Right: median loss-relative tightness $\widehat{B}_N/L$ falls from 12.69 ($K = 25$) to 3.93 ($K = 400$); validity is independent of K while additional rollout budget reduces conservatism.

the de Finetti mixture $g(u) = \mathbb{E}_{\Theta_u}[g_N(\Theta_u)]$, so the true loss itself rises as s shrinks (0.371 → 0.544 across the sweep, Figure 6 right).

The mixture-g certificate, which uses the true mixture g (a controlled exact-DP audit), maintains coverage 1.000 with zero violations across all over-dispersion levels (Figure 6, left). The binomial plug-in certificate, which substitutes $g_N(\hat{\alpha}_u)$ at the mean endorsement, under-covers increasingly as concentration shrinks, its under-cover rate rising monotonically from 0.00 at $s = 500$ through $0.18, 0.42, 0.80, 0.90, 0.98$ to 1.00 at $s \leq 30$. The direction is determined by Jensen's inequality on the convex g_N, which forces $\mathbb{E}_\Theta[g_N(\Theta)] > g_N(\mathbb{E}[\Theta])$: the plug-in systematically under-estimates the failure rate and hence the loss. This supports the dependence-agnostic certificate as the appropriate validity target under correlated committee votes, and shows that validity does not rely on the binomial form.

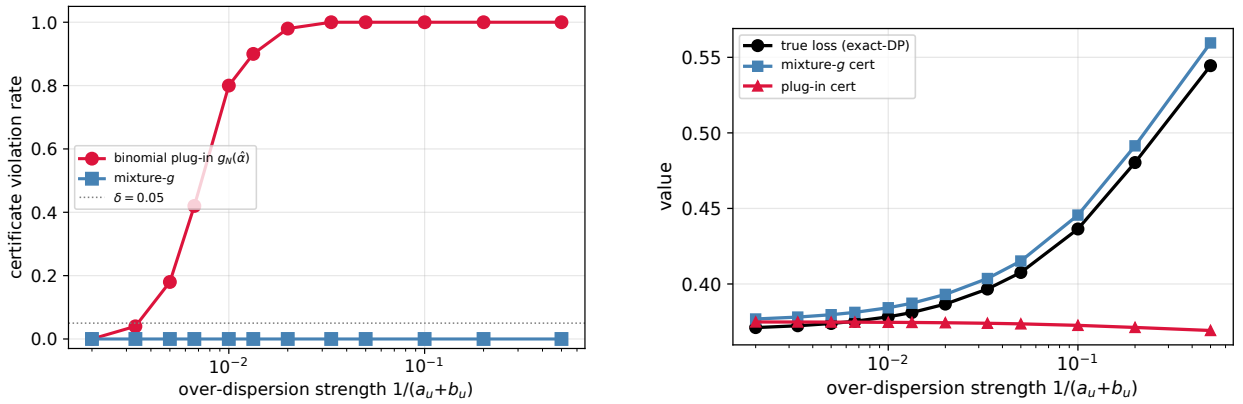

Figure 6: Correlated committees via Beta-Binomial over-dispersion (concentration s; smaller s = more over-dispersed). Left: the mixture-g certificate stays valid (coverage 1.000) while the binomial plug-in's under-cover rate rises monotonically to 1.0. Right: the exact-DP true loss itself rises with over-dispersion ($0.371 \to 0.544$), since the controller's failure rate is the de Finetti mixture.

9.4 Diagnostics and additional checks

Appendix B reports additional checks that are not core to the validity claims: a rare-unit sample-complexity scaling consistent with Proposition 4; a comparison against adjacent off-policy and robust-RL baselines; a computational observation that rollout-based certification avoids exact enumeration under a fixed sampling budget; and the $\eta = 0$ boundary and component-falsification ablations. As a learned-committee diagnostic, on MPE `simple_spread` with trained IPPO and MAPPO committees (Yu et al., 2022; Rutherford et al., 2024) the debiased endorsement rate is stable across committee budgets ($\widehat{p} \approx 0.26$), supporting the conditional-i.i.d. advisor approximation in that environment; this is diagnostic only, and strict validity rests on exact tabular evaluation or on Assumption 2 discharged by Theorem 4.

10 Discussion and Limitations

The certificate is profile-relative and worst-case over the information available. The validity chain is unconditional, and the optimality claim is established over an explicit constructive witness class rather than an abstract assumption. The carrier is agnostic to the committee's dependence structure: it requires only the failure probability g, and the over-dispersion experiment confirms that validity does not rely on the binomial form. The main text uses $\eta = 0$ to isolate endorsement-failure downside; for $\eta > 0$, the success-side term $(1-g)w_\psi$ is necessary. Logging the fallback identity is low-overhead at deployment, but numerical evaluation of W and W_{fb} requires exact values or the conservative confidence bound of Theorem 4. Learned-critic results are therefore diagnostic unless a uniform conservative value bound is independently established. One route to such a bound in learned settings is a value estimator with an explicit confidence penalty, in the spirit of conservative offline value estimation (Kumar et al., 2020); whether a penalty calibrated to a target δ_G delivers the uniform conservative guarantee of Assumption 2 is left to future work.

Reproducibility Statement

All experiments are CPU-only and seeded (`seed` $= 0$). We provide anonymized code, configuration files, and per-instance logs in the supplementary material; all headline numbers in the paper are recomputed from these logs. Validity experiments are audited against exact dynamic-programming truth; the conservative rollout-bridge uses the value-bound construction of Theorem 4; the correlated-committee experiment uses exact-DP truth under the de Finetti mixture failure process. Per-experiment configurations (instance counts, m, δ, value source) are listed in Appendix B.

Broader Impact Statement

This work develops deployment-time guarantees for agreement-gated committee controllers in cooperative multi-agent systems. As test-time committees of policies or ensemble advisors are increasingly used to spend additional compute at deployment, the ability to certify their value-loss relative to a reference policy from observable quantities is directly relevant to the safe operation of such systems. The certificate is conservative by construction and is intended to support, not replace, domain-specific safety review: the certificate bounds the reference-relative downside, but the bound is only as meaningful as the modelling assumptions behind it. Two cautions are worth stating explicitly. First, the strict-validity guarantees hold under exact value evaluation or under the conservative value-bound construction of Theorem 4; a generic learned critic does not satisfy the conservative-value assumption without separate validation, and treating learned-critic outputs as certified could give a false sense of safety. Second, the certificate bounds value-loss against a chosen reference policy and does not by itself adjudicate whether that reference is itself appropriate for the deployment. We therefore recommend that practitioners verify the conservative-value condition and the choice of reference before drawing operational conclusions from any certificate produced by this method.

References

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

Anastasios N. Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. Conformal risk control. In *International Conference on Learning Representations (ICLR)*, 2024.

Stephen Bates, Anastasios N. Angelopoulos, Lihua Lei, Jitendra Malik, and Michael I. Jordan. Distribution-free, risk-controlling prediction sets. *Journal of the ACM*, 68(6):1–34, 2021.

Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, 2013.

Bradley Brown, Jordan Juravsky, Ryan Ehrlich, Ronald Clark, Quoc V. Le, Christopher Ré, and Azalia Mirhoseini. Large language monkeys: Scaling inference compute with repeated sampling. *arXiv preprint arXiv:2407.21787*, 2024.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. Improving factuality and reasoning in language models through multiagent debate. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, volume 235 of *PMLR*, pp. 11733–11763, 2024.

Amir-massoud Farahmand, Rémi Munos, and Csaba Szepesvári. Error propagation for approximate policy and value iteration. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2010.

Jakob N. Foerster, Gregory Farquhar, Triantafyllos Afouras, Nantas Nardelli, and Shimon Whiteson. Counterfactual multi-agent policy gradients. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence*, 2018.

Shangding Gu, Long Yang, Yali Du, Guang Chen, Florian Walter, Jun Wang, and Alois Knoll. A review of safe reinforcement learning: Methods, theory and applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. doi: 10.1109/TPAMI.2024.3457538.

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

Nan Jiang and Lihong Li. Doubly robust off-policy value evaluation for reinforcement learning. In *Proceedings of the 33rd International Conference on Machine Learning (ICML)*, pp. 652–661, 2016.

Sham Kakade and John Langford. Approximately optimal approximate reinforcement learning. In *Proceedings of the 19th International Conference on Machine Learning (ICML)*, pp. 267–274, 2002.

Jakub Grudzien Kuba, Muning Wen, Linghui Meng, Haifeng Zhang, David Mguni, Jun Wang, and Yaodong Yang. Settling the variance of multi-agent policy gradients. *Advances in Neural Information Processing Systems (NeurIPS)*, 34:13458–13470, 2021.

Jakub Grudzien Kuba, Ruiqing Chen, Muning Wen, Ying Wen, Fanglei Sun, Jun Wang, and Yaodong Yang. Trust region policy optimisation in multi-agent reinforcement learning. In *International Conference on Learning Representations (ICLR)*, 2022.

Aviral Kumar, Aurick Zhou, George Tucker, and Sergey Levine. Conservative Q-learning for offline reinforcement learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Romain Laroche, Paul Trichelair, and Rémi Tachet des Combes. Safe policy improvement with baseline bootstrapping. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, pp. 3652–3661, 2019.

Shalev Lifshitz, Sheila A. McIlraith, and Yilun Du. Multi-agent verification: Scaling test-time compute with multiple verifiers. *arXiv preprint arXiv:2502.20379*, 2025.

Michael L. Littman. Markov games as a framework for multi-agent reinforcement learning. In *Machine Learning Proceedings 1994*, pp. 157–163. Morgan Kaufmann, 1994.

Ryan Lowe, Yi Wu, Aviv Tamar, Jean Harb, Pieter Abbeel, and Igor Mordatch. Multi-agent actor-critic for mixed cooperative-competitive environments. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Andreas Maurer and Massimiliano Pontil. Empirical Bernstein bounds and sample-variance penalization. In *Proceedings of the 22nd Conference on Learning Theory (COLT)*, 2009.

Ronghui Mu, Leandro Soriano Marcolino, Yanghao Zhang, Tianle Zhang, Xiaowei Huang, and Wenjie Ruan. Certified policy smoothing for cooperative multi-agent reinforcement learning. In *Proceedings of the 37th AAAI Conference on Artificial Intelligence*, 2023.

Rémi Munos and Csaba Szepesvári. Finite-time bounds for fitted value iteration. *Journal of Machine Learning Research*, 9:815–857, 2008.

Frans A. Oliehoek and Christopher Amato. *A Concise Introduction to Decentralized POMDPs*. Springer, 2016.

Afshin Oroojlooy and Davood Hajinezhad. A review of cooperative multi-agent deep reinforcement learning. *Applied Intelligence*, 53(11):13677–13722, 2023.

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, 2018.

Alexander Rutherford, Benjamin Ellis, Matteo Gallici, Jonathan Cook, Andrei Lupu, Gardar Ingvarsson, Timon Willi, Akbir Khan, Christian Schroeder de Witt, Alexandra Souly, Saptarashmi Bandyopadhyay, Mikayel Samvelyan, Minqi Jiang, Robert Tjarko Lange, Shimon Whiteson, Bruno Lacerda, Nick Hawes, Tim Rocktaschel, Chris Lu, and Jakob Nicolaus Foerster. JaxMARL: Multi-agent RL environments and algorithms in JAX. In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, 2024.

Mikayel Samvelyan, Tabish Rashid, Christian Schroeder de Witt, Gregory Farquhar, Nantas Nardelli, Tim G. J. Rudner, Chia-Man Hung, Philip H. S. Torr, Jakob Foerster, and Shimon Whiteson. The StarCraft multi-agent challenge. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems (AAMAS)*, pp. 2186–2188, 2019.

Bruno Scherrer, Mohammad Ghavamzadeh, Victor Gabillon, Boris Lesner, and Matthieu Geist. Approximate modified policy iteration and its application to the game of Tetris. *Journal of Machine Learning Research*, 16(49):1629–1676, 2015.

John Schulman, Sergey Levine, Pieter Abbeel, Michael I. Jordan, and Philipp Moritz. Trust region policy optimization. In *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, pp. 1889–1897, 2015.

Lloyd S. Shapley. Stochastic games. *Proceedings of the National Academy of Sciences*, 39(10):1095–1100, 1953.

Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling LLM test-time compute optimally can be more effective than scaling parameters for reasoning. In *International Conference on Learning Representations (ICLR)*, 2025.

Kyunghwan Son, Daewoo Kim, Wan Ju Kang, David Earl Hostallero, and Yung Yi. QTRAN: Learning to factorize with transformation for cooperative multi-agent reinforcement learning. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, pp. 5887–5896, 2019.

Peter Sunehag, Guy Lever, Audrunas Gruslys, Wojciech Marian 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 based on team reward. In *Proceedings of the 17th International Conference on Autonomous Agents and MultiAgent Systems (AAMAS)*, pp. 2085–2087, 2018.

Philip S. Thomas, Georgios Theocharous, and Mohammad Ghavamzadeh. High-confidence off-policy evaluation. In *Proceedings of the 29th AAAI Conference on Artificial Intelligence*, 2015.

Alexandre B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer, 2009.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *International Conference on Learning Representations (ICLR)*, 2023.

Muning Wen, Jakub Grudzien Kuba, Runji Lin, Weinan Zhang, Ying Wen, Jun Wang, and Yaodong Yang. Multi-agent reinforcement learning is a sequence modeling problem. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

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

Lei Yuan, Tao Jiang, Lihe Li, Feng Chen, Zongzhang Zhang, and Yu Yang. Robust cooperative multi-agent reinforcement learning via multi-view message certification. *Science China Information Sciences*, 67:142102, 2024. doi: 10.1007/s11432-023-3853-y.

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 Proofs and Additional Details

A.1 Formal setup

Definition 10 (Coordinate value)**.** For $u = (t, s, i, a_{<i})$,

$$Q^{\pi^{\mathrm{ref}}}(u, a_i) = \mathbb{E}\big[r_t(s_t, a_t) + V_{t+1}^{\pi^{\mathrm{ref}}}(s_{t+1}) \mid s_t = s,\ a_{<i},\ a_i,\ a_{>i} \sim \pi^{\mathrm{ref}}\big].$$

The return ceiling follows from bounded rewards: $J_M(\pi^{\mathrm{ref}}) - J_M(\pi) \le H\Delta_r = R_{\max}$ for every policy π.

A.2 Per-agent failure and prefix drift

Proof of Proposition 1. The two-agent example gives true loss

$$2g(1-g)(0.1) + g^2 = 0.2g(1-g) + g^2,$$

while the marginal certificate gives $0.1g + 0.1g = 0.2g$. For $g = 0.317$, the true loss is approximately 0.1438 and the marginal certificate is approximately 0.0634. For the n-agent quadratic family, if $K \sim \mathrm{Bin}(n, g)$ is the number of deviating coordinates, then

$$\mathbb{E}\left[\left(\frac{K}{n}\right)^2\right] = \frac{ng(1-g) + (ng)^2}{n^2} = \frac{g(1-g)}{n} + g^2.$$

The marginal certificate is g/n, so the ratio of true loss to marginal certificate is $(1-g) + ng$. □

Proof of Proposition 2. The reference-prefix bound evaluates coordinate swings only at $a^{\mathrm{ref}}_{<i}$, while the deployed controller visits non-reference prefixes with positive probability. The joint certificate integrates over the deployed prefix law:

$$C_2 = \sum_{t,s,i,a_{<i}} d_t^{\pi^{\mathrm{ctrl}}}(s)\mathbb{P}(a_{<i} \mid t, s)g(t, s, i, a_{<i})W_{\mathrm{fb}}(t, s, i, a_{<i}).$$

In the constructed time-inhomogeneous instance, exact dynamic programming gives true loss 0.21290, reference-prefix bound 0.20947, and joint certificate 0.21489, proving strict invalidity of the reference-prefix certificate. □

A.3 Carrier bound

Proof of Theorem 1. For finite-horizon Markov games, the performance-difference identity gives

$$J(\pi^{\mathrm{ref}}) - J(\pi^{\mathrm{ctrl}}) = \sum_{t=0}^{H-1} \mathbb{E}_{s_t \sim d_t^{\pi^{\mathrm{ctrl}}}} \left[V_t^{\pi^{\mathrm{ref}}}(s_t) - \mathbb{E}_{a_t \sim \pi_t^{\mathrm{ctrl}}(\cdot|s_t)} Q_t^{\pi^{\mathrm{ref}}}(s_t, a_t) \right].$$

Since $V_t^{\pi^{\mathrm{ref}}}(s) = \mathbb{E}_{a \sim \pi_t^{\mathrm{ref}}(\cdot|s)} Q_t^{\pi^{\mathrm{ref}}}(s, a)$, decompose the joint-action difference by replacing the coordinates in a fixed virtual order. Conditioning on the deployed prefix $a_{<i}$, the ith increment is

$$Q^{\pi^{\mathrm{ref}}}(u, a_i^{\mathrm{ref}}) - \mathbb{E}_{a_i \sim \pi^{\mathrm{ctrl}}(\cdot|u)} Q^{\pi^{\mathrm{ref}}}(u, a_i) \le \mathbb{E}_{a_i \sim \pi^{\mathrm{ctrl}}(\cdot|u)} \Delta_+(u, a_i),$$

because

$$Q^{\pi^{\mathrm{ref}}}(u, a_i^{\mathrm{ref}}) - Q^{\pi^{\mathrm{ref}}}(u, a_i) \le \left[Q^{\pi^{\mathrm{ref}}}(u, a_i^{\mathrm{ref}}) - Q^{\pi^{\mathrm{ref}}}(u, a_i) \right]_+.$$

At unit u, the controller succeeds with probability $1 - g(u)$ and incurs at most $w_\psi(u)$; it fails with probability $g(u)$ and incurs conditional mean $W_{\mathrm{fb}}(u)$. Therefore the coordinate increment is at most $(1 - g(u))w_\psi(u) + g(u)W_{\mathrm{fb}}(u)$. Summing over t, states, coordinates, and prefixes yields

$$J(\pi^{\mathrm{ref}}) - J(\pi_N^{\mathrm{ctrl}}) \le \sum_u \bar{\mu}(u) \left[(1 - g(u))w_\psi(u) + g(u)W_{\mathrm{fb}}(u) \right] = nH\mathbb{E}_{U \sim \mu}[(1-g)w_\psi + gW_{\mathrm{fb}}].$$

□

A.4 Information characterization

Lemma 3 (Admissible units are locally cap-inactive)**.** *For every admissible unit law and every information level k,*

$$\bar{\mu}(u)g(u)\omega_k(u) \le R_{\max}.$$

Proof. By construction, $\bar{\mu}(u) = d_t^{\pi^{\mathrm{ctrl}}}(s)\mathbb{P}(a_{<i} \mid t, s) \le 1$, $g(u) \le 1$, and $\omega_k(u) \le (H - t_u)\Delta_r \le H\Delta_r = R_{\max}$. Multiplying the three inequalities proves the claim. □

Lemma 4 (Finite-tail unit gadget)**.** *For any unit $u = (t, s, i, a_{<i})$ and any $\omega(u) \in [0, (H - t)\Delta_r]$, there is a deterministic tail gadget with rewards in $[0, \Delta_r]$ whose reference-minus-fallback value gap equals $\omega(u)$.*

Proof. Let $\omega(u) = q\Delta_r + r$ with integer $q \in \{0, \dots, H - t\}$ and $r \in [0, \Delta_r)$, truncating q at $H - t$ with $r = 0$ when equality holds. Along the reference tail, assign reward Δ_r for q future steps and reward r for one additional step if $r > 0$, with zero elsewhere. Along the fallback tail, assign zero on those steps and match the reference rewards elsewhere. The total gap is $q\Delta_r + r = \omega(u)$ and every reward lies in $[0, \Delta_r]$. □

Lemma 5 (Cliff-chain cap-active gadget)**.** *For every $H \ge 2$ and $\Delta_r > 0$ there is a single-agent ($n = 1$) deterministic instance with rewards in $[0, \Delta_r]$ and an agreement-gated controller, realizing the information profile $\bar{\mu}(u_t) = 1$, $g(u_t) = 1$, $W(u_t) = (H-t)\Delta_r$, $W_{\mathrm{fb}}(u_t) = \Delta_r$ at the units $u_t = (t, s_t, 1, \varnothing)$, $t = 0, \dots, H-1$, whose true loss is exactly $R_{\max} = H\Delta_r$ while*

$$C_0 = C_1 = \tfrac{H(H+1)}{2}\,\Delta_r > R_{\max}.$$

Hence $R_{\max} \wedge C_k = R_{\max} = L$ for $k \in \{0, 1\}$: the cap-active optimum is attained constructively.

Proof. Take states $s_0 \to s_1 \to \cdots \to s_{H-1}$ on a chain plus an absorbing dead state D with reward 0. The reference policy takes, at each s_t, an action earning Δ_r and moving to s_{t+1}, so $V^{\pi^{\mathrm{ref}}}(s_t) = (H - t)\Delta_r$ and $V^{\pi^{\mathrm{ref}}}(D) = 0$. At each s_t the controller fails ($g = 1$, e.g. endorsement rate 0) and executes the fallback a^{fb} that earns 0 and moves to s_{t+1}; thus the controller earns 0 throughout and $L = H\Delta_r - 0 = R_{\max}$. The available actions at s_t include a cliff action earning 0 and moving to D. Then $Q^{\pi^{\mathrm{ref}}}(u_t, a^{\mathrm{ref}}) = (H - t)\Delta_r$, $Q^{\pi^{\mathrm{ref}}}(u_t, a^{\mathrm{fb}}) = (H - 1 - t)\Delta_r$, and $Q^{\pi^{\mathrm{ref}}}(u_t, \mathrm{cliff}) = 0$, so $W(u_t) = (H - t)\Delta_r$ and $W_{\mathrm{fb}}(u_t) = \Delta_r$. Occupancy is deterministic, $\bar{\mu}(u_t) = 1$. Summing, $C_0 = C_1 = \sum_{t=0}^{H-1}(H - t)\Delta_r = \frac{1}{2}H(H + 1)\Delta_r > H\Delta_r = R_{\max}$ for $H \ge 2$, while $C_2 = \sum_{t=0}^{H-1}\Delta_r = H\Delta_r = R_{\max}$. All rewards lie in $[0, \Delta_r]$. The profile is $\mathcal{I}_1$-consistent with the stated certificate values, and the true loss equals $R_{\max} = R_{\max} \wedge C_k$ for $k \in \{0, 1\}$. □

Proof of Proposition 3. For each unit, use Lemma 4 to realize the desired coordinate gap $\omega_k(u)$ over the remaining horizon. Assign the unit occupancy and failure probability according to the admissible profile. Since $C_k \le R_{\max}$, summing the unit-level losses does not exceed the feasible return range. The total loss is exactly

$$\sum_u \bar{\mu}(u)g(u)\omega_k(u) = C_k,$$

and the rewards remain bounded by the tail-gadget construction. □

Proof of Theorem 2. Part (a) follows from Corollary 1 and $W_{\mathrm{fb}} \le W \le (H - t)\Delta_r$, plus $L \le R_{\max}$. For part (b), part (a) gives $\mathcal{C}^{\mathrm{opt}}_{\mathfrak{M}_{\mathrm{tail}}}(\mathcal{I}_k) \le R_{\max} \wedge C_k$. The cap-inactive case $C_k \le R_{\max}$ is attained by the finite-tail gadget (Lemma 4, via Proposition 3) for all k. The cap-active case $C_k > R_{\max}$ is attained for $k \in \{0, 1\}$ by the cliff-chain gadget (Lemma 5), which realizes true loss $R_{\max}$ with $C_0 = C_1 > R_{\max}$, so $R_{\max} \wedge C_k = R_{\max} = L$. Hence equality holds over $\mathfrak{M}_{\mathrm{tail}}$, and over any $\mathfrak{M}_{\mathrm{R}} \supseteq \mathfrak{M}_{\mathrm{tail}}$. For part (c), Lemma 3 allows a single-unit finite-tail witness for every positive-mass unit. Any coordinate-local pre-cap certificate that reduces the pointwise weight below $\omega_k(u)$ on such a set is violated by that witness. □

Proof of Lemma 2. The equality $C_1 = C_2$ holds when the logged fallback always selects an action attaining W. The equality $C_0 = C_1$ holds in a single-step all-or-nothing game where $W = (H - t)\Delta_r = \Delta_r$. The equality $C_2 = L$ holds for the cap-inactive finite-tail witness of Proposition 3. □

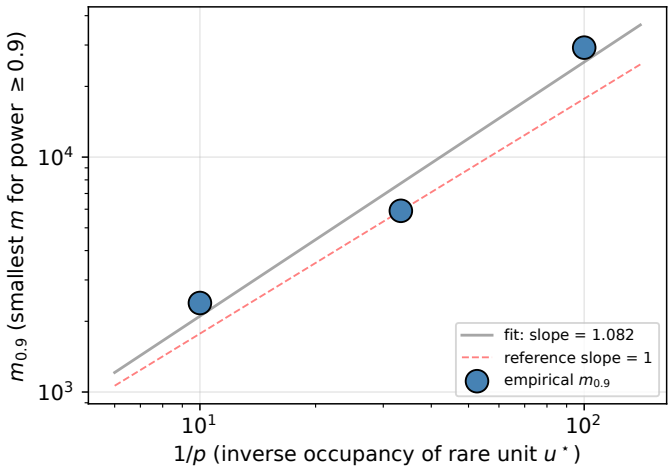

Figure 7: Rare-unit scaling. Episodes m to detect an ε-shift at a unit of occupancy p, against $1/p$ on log-log axes. The fitted slope is 1.082 ($R^2 = 0.966$), consistent with $m = \Theta(1/p)$.

A.5 Finite-sample certificate and lower bound

Proof of Proposition 4. Let P_0 and P_1 differ only at unit $u^\star$, whose occupancy under the rollout-then-sample protocol is p. Conditional on landing on $u^\star$, the observation distribution has means separated by ε and bounded variance, so the one-hit KL is $O(\varepsilon^2)$. The total KL over m episodes is $O(mp\varepsilon^2)$. If $m \lesssim 1/(p\varepsilon^2)$, the total variation distance remains bounded away from one by Pinsker's inequality, and Le Cam's method implies constant error for any test. Thus constant-probability distinction requires $m = \Omega(1/(p\varepsilon^2))$. □

A.6 Supporting lemmas

Proof of Lemma 1. Write $M = N - 1$, $k = (N-3)/2$, and $q = 1 - \alpha$. Let $b_M(r) = \binom{M}{r}\alpha^r q^{M-r}$. A two-step binomial-CDF identity gives

$$h_{N+2}(\alpha) - h_N(\alpha) = q^2 b_M(k+1) - \alpha^2 b_M(k).$$

Since

$$\frac{b_M(k+1)}{b_M(k)} = \frac{M-k}{k+1}\frac{\alpha}{q},$$

the increment is negative iff

$$\frac{q}{\alpha}\frac{M-k}{k+1} < 1.$$

Substituting $M - k = (N+1)/2$ and $k + 1 = (N-1)/2$ yields $\alpha > (N+1)/(2N)$, with equality at the threshold. □

B Additional Experiments

Rare-unit sample-complexity scaling. As a sanity check consistent with the rare-unit lower bound (Proposition 4), we measure with pure exact-DP evaluation the episode count m needed to detect an ε-level fallback-swing shift at a unit of occupancy p. Over the feasible range $p \in \{10^{-1}, 3\times10^{-2}, 10^{-2}\}$ the required m is $\{2393, 5907, 29240\}$, a log-log slope of 1.082 against $1/p$ with $R^2 = 0.966$ (Figure 7), consistent with the $m = \Theta(1/p)$ scaling of Proposition 4 and Remark 6. For $p \le 3\times10^{-3}$ the theoretical sample size exceeds 10^6 and is computationally infeasible to verify directly; those points are archived rather than reported.

Method	Coverage	Median bound/loss	Comment
$\mathcal{I}_2$ certificate (this paper)	1.000	1.019	tightest valid
PDIS (Thomas et al., 2015)	1.000	1.607	valid, $\approx 58\%$ looser
Doubly-robust (Jiang & Li, 2016)	1.000	1.619	valid, $\approx 58\%$ looser
Robust simulation lemma (Kakade & Langford, 2002)	1.000	4.438	valid but uninformative

Table 2: Adjacent OPE / robust-RL baselines versus the $\mathcal{I}_2$ certificate, audited against exact truth (8 games, 2000 rollouts each). All are valid; the $\mathcal{I}_2$ certificate is the tightest, and the structural difference rather than the numerical gap is the point.

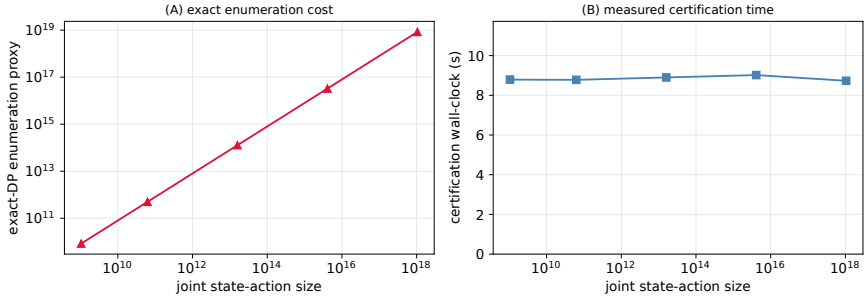

Figure 8: Computational observation on enumeration. The left panel reports an exact-DP enumeration proxy; the right panel reports the measured wall-clock of the rollout certificate under a fixed rollout budget. The figure supports infeasibility of exact enumeration at large joint state-action sizes, not strict validity on realistic neural MARL.

Adjacent baselines from OPE and robust RL. We calibrate the $\mathcal{I}_2$ certificate against adjacent off-policy and robust-RL alternatives (8 games, 2000 rollouts each), auditing all against exact truth (Table 2). The $\mathcal{I}_2$ certificate attains coverage 1.000 at median bound/loss 1.019. Per-decision importance sampling (PDIS) and the doubly-robust estimator are also valid but substantially looser, and the robust-simulation-lemma bound is valid but uninformative. The comparison is intended to calibrate the certificate against adjacent alternatives rather than to claim dominance over OPE methods in their native estimation setting: the relevant structural difference is that OPE baselines require importance weights and do not produce the per-unit attribution of Section 8.

Computational observation on enumeration. Fixing $n = 4$, $H = 8$, and $m = 300$ and varying the grid size, the rollout certificate uses a fixed number of sampled units while exact DP would require enumeration over the joint state-action space. Under a fixed sampling budget the measured certification wall-clock stays nearly flat (Figure 8); the exact-DP curve is an enumeration proxy, not a measured wall-clock in the same unit. This is a computational observation that rollout-based certification avoids exact enumeration under fixed sampling budgets, not a validity claim for realistic neural MARL.

$\eta = 0$ **boundary.** At $\eta = 0$, the value-loss bound is never violated across 1200 instances. For $\eta \geq 0.1$, the pure failure term $nH\mathbb{E}[gW_{\mathrm{fb}}]$ that drops the success-side contribution is violated in every instance (Figure 9).

Falsification of certificate components. Replacing the operational certificate C_2 by ablated variants and measuring against exact truth (Figure 10) shows that C_2 itself is never violated, the unweighted variant is violated in 99.7% of instances, the plug-in-α variant in 19.3%, and the no-logging variant C_1 remains valid but pays a median slack of 0.43 above C_2.

Rank consistency and budget monotonicity. The fixed-occupancy committee-sizing checks show certificate monotonicity in the budget in all tested instances, Spearman rank correlation 1.0 between certificate and true loss with no reversals (reversal fraction 0.000), and no true-loss increase in 99.3% of

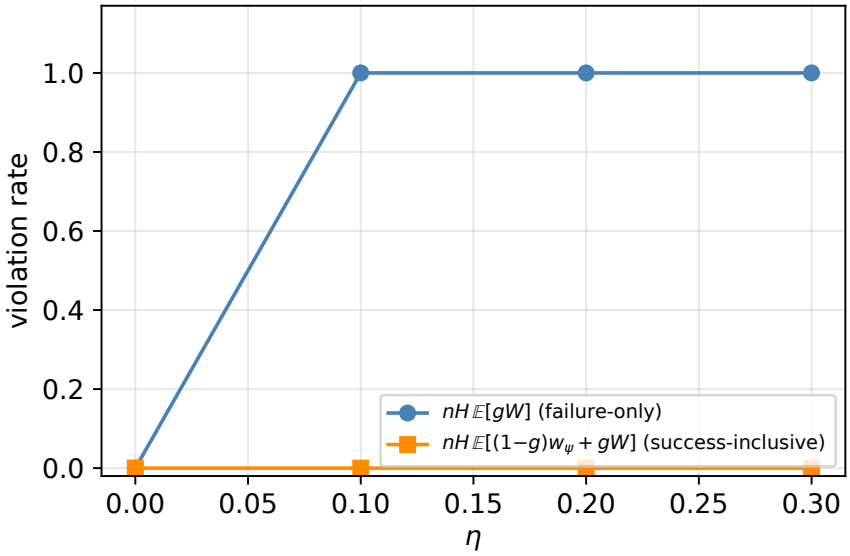

Figure 9: $\eta = 0$ boundary. Violation rate of the pure-failure term as a function of η: zero at $\eta = 0$ and one for $\eta \geq 0.1$, confirming that the success-side term is not removable for $\eta > 0$.

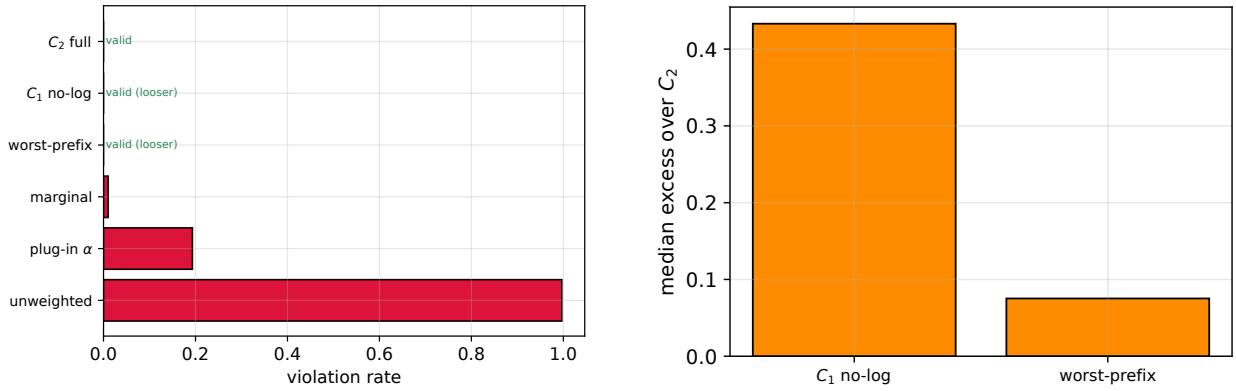

Figure 10: Falsification of certificate components. Left: the full logged-fallback certificate C_2 and conservative no-log variants remain valid, while removing occupancy weighting or calibrated failure probabilities breaks validity. Right: the no-log variants are valid but looser than C_2; they are not failed baselines.

tested fixed-occupancy budget increases. Because this is a diagnostic use of the certificate rather than a main contribution, we summarize it numerically here and leave the budget-drift plots out of the paper.

Sharpness witnesses. Closed-form witness values match exact-DP computations up to numerical precision; the maximum occupancy reconstruction error across 85 witnesses is 3.3×10^{-16}, confirming the equality cases of Lemma 2.

Per-experiment configurations. Table 3 lists the configuration of each experiment, including instance counts, sample sizes, confidence levels, and value source.

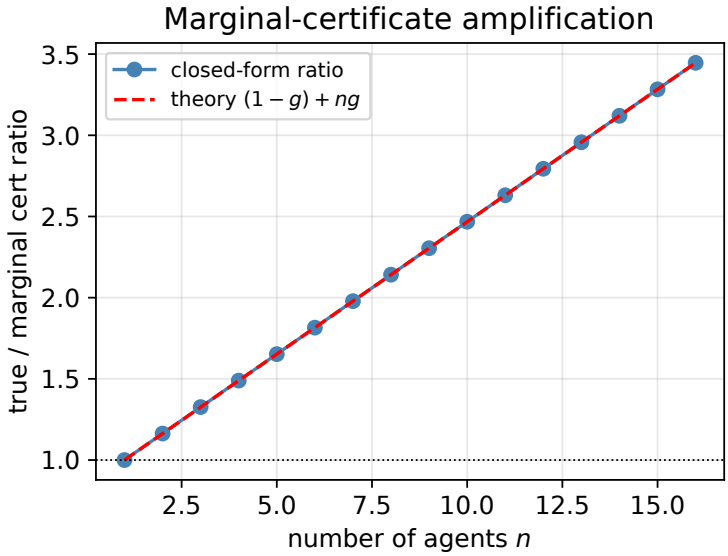

Figure 11: Marginal-certificate failure and amplification (closed-form check for Proposition 1). On the canonical n-agent family, the ratio of true team value-loss to the per-agent marginal certificate follows the predicted $(1-g)+ng$ law and equals 1 at $n=1$, so the per-agent certificate under-estimates the true loss for $n \geq 2$.

Experiment	Instances	m / repeats	Value source
Exact-tabular chain	300 games	—	exact DP (δ_G=0)
Finite-sample / bound comp.	8 games	$m \in \{500, 1500, 5000\}$, 200 rep	exact DP
K-sweep (Thm 4)	8 games	20 rep, $K \in \{25, \ldots, 400\}$	cons. rollout
Rare-unit scaling	$p \in \{10^{-1}, 3{\times}10^{-2}, 10^{-2}\}$	—	exact DP
Adjacent baselines	8 games	2000 rollouts	exact truth
Correlated committees	50 games $\times$ 12 levels	N=5, $s \in [2, 500]$	exact DP (mixture)

Table 3: Per-experiment configurations. All runs use seed 0 and are reproducible from the per-instance logs in the supplementary material.