LEARNING INCENTIVES, NOT UTILITIES: CERTIFIED SELECTION AND STEERING IN EXTENSIVE-FORM GAMES

Anonymous authors
Paper under double-blind review

ABSTRACT

We study how a principal can steer no-regret learning agents in a repeated extensive-form game without knowing their utility functions. Left alone, such agents can settle on outcomes that are poor for the principal. The principal knows the game tree and its own payoff, observes play, and can privately recommend actions and pay the agents at its own cost. Given a class of candidate rules for doing so, it must decide which are robust: under the true utilities, every agent loses by deviating from their recommendations. A large enough bonus for following recommendations is robust without any learning, but it is paid in every round. Whether a cheaper rule is also robust depends on utilities the principal cannot see. We show that the principal never needs the utility functions themselves. Learning only the agents' incentives, it can check every candidate at once, and choosing the highest-value certified candidate is safe. Steering to it then earns its value, net of payments, up to the agents' regret and the rounds spent learning. This is possible because in extensive form, incentives depend on the utilities only up to one constant per information set. Every relevant deviation gain is linear in what remains, the incentive-relevant part of the utilities. As in recent work on normal-form games, this part can be learned by a zero-sum payment game against one agent at a time. The proof is a modular reduction whose steps each have a concrete construction, and we report diagnostics on small games. For bonus-based rules, all steps can be carried out together under extra assumptions on observation and learning; steering a cheaper certified rule is left open. In short, learning buys value rather than feasibility: it lets the principal trust cheaper rules.

1 INTRODUCTION

A platform sits between a seller and a buyer who interact in a few moves. At each decision point it can privately recommend an action and pay a small bonus for following it, but it sees only the completed transaction. Both sides adapt over repeated interactions, using learning rules the platform does not control. Formally, the seller and the buyer are the agents of a repeated perfect-recall extensive-form game (Kuhn, 1953). The equilibrium notion that matches this setting is extensive-form correlated equilibrium (EFCE), in which a mediator privately recommends an action at each information set as it is reached (von Stengel & Forges, 2008). In equilibrium, no agent gains by ignoring a recommendation it has received and then playing freely. We call such a move a *trigger deviation*; a buyer who skips a recommended inspection makes one. We assume that the agents have no regret against trigger deviations. Learners of this kind approach the set of these equilibria (Farina et al., 2022), which can contain many outcomes, some of them poor for the platform. The platform, our principal, can act as this mediator and also pay, but it knows only the tree, the chance process, and its own payoff. Which recommendation-and-payment rules can it trust, and how can it steer play to the best of them, when trust depends on utilities it cannot see?

Two papers come closest to this question. Zhang et al. (2024) steer no-regret learners with recommendations and payments when the utilities are known; we keep their instruments and agent model, but our principal does not know the utilities. Zhang et al. (2026) learn a normal-form game up to strategic equivalence by playing a zero-sum payment game against each agent, and then steer to the best outcome of the learned game. We adapt their payment game to extensive form but, instead of

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.

the whole game, learn only what is needed to check a given set of rules and choose among them. Section 7 develops the comparison.

To make the question precise, call each rule the principal might adopt a *candidate*: it privately recommends actions and pays the agents. A candidate is *robust* with a given margin if, under the true utilities, every agent loses at least that margin by any trigger deviation from a recommendation it receives. The principal values a candidate by its expected payoff net of the payments, so paying more for compliance buys margin but lowers the value. The cost does not fade: when only the realized path of play is observed, no steering method with bounded payments can make the average payment vanish in general, even with known utilities (Zhang et al., 2024).

For the platform, the trade-off is concrete. One candidate needs no learning: a *compliance bonus* large enough to outweigh any gain from deviating is robust if the utilities have a known bound. But it is paid in every round, and whether the small bonus of the opening example would also suffice depends on utilities the platform cannot see. Given a class of candidates, such as several bonus levels, the principal must certify which are robust, select the best certified one as its *target*, and steer play to it. How much better than the large bonus can it safely do?

The principal cannot learn the utilities themselves, but it does not need to: certification needs only the agents' incentives, and extensive form gives them a structure the principal can exploit. To see this structure, hold everything except one agent's choices fixed; we call this a *frozen environment*. That agent's payoff is then linear in the probabilities it assigns to sequences of its own actions, with one coefficient per sequence. The vector of these probabilities is its *realization plan*, the sequence-form representation of its strategy (Koller et al., 1996).

For example, suppose the buyer can decline (L) or inspect (R) and, after inspecting, buy (a) or not (b). Raise the coefficient of R by a constant and lower those of Ra and Rb by the same constant. No strategy's total changes, since a buyer who inspects gains the constant there and gives it back at the next step. The difference between the coefficients of Ra and Rb, which decides whether to buy, is also untouched. In general, such a change at an information set, up on the sequence leading into it and down on each sequence continuing from it, is an *information-set potential*. It shifts every strategy's payoff by the same amount and so changes no comparison between strategies.

The part of the coefficients that potentials cannot change is exactly what incentives depend on, so no incentive comparison can reveal a utility table. In each frozen environment, we call the coefficients modulo these potentials the agent's *quotient representative*. It is also all the principal needs: every deviation gain that robustness constrains is a known payment term plus a linear function of one frozen environment's representative. The representatives are unknown, so certification uses estimates.

Contributions. The paper answers five questions.

(1) How can the representatives be learned, and how well? In a frozen environment the principal learns one agent's representative by playing a zero-sum payment game against it. The estimate's error is at most a constant times both sides' combined average regret in that game (Theorem 3.2).

(2) Is that enough to steer play to one target? Yes, if three steps can be carried out; we call them *interfaces*, since any implementation with a stated guarantee can fill them. The steps are learning the representatives while the other agents are held to the target, reading every deviation gain off the estimates, and steering a target whose gains are certified. Suppose the target is truly robust and the estimated gains are within half its margin of the truth. Then, with high probability, the principal steers to it and earns its value, up to the agents' regret and the rounds spent learning (Theorem 4.5).

(3) Which candidate should the principal choose? The certified one with the highest value. One learning phase, run on the environments of all the candidates, estimates every candidate's gains at once, and a candidate is certified when its estimated margin clears a threshold. The selected candidate is safe: its true margin falls short of the threshold by at most the estimation error. It is also never worse than any candidate whose true margin exceeds the threshold by at least that error (Theorem 5.1).

(4) What does learning buy? Value, not feasibility: when compliance is verifiable, a large enough compliance bonus needs no learning. If the class contains such a bonus, the choice is never worse than it, and strictly better if a cheaper payment rule for the same recommendations is certified.

(5) Can the three steps be carried out, and where does the paper stop? An explicit algebraic construction can always read gains off the estimates. A monitored construction carries out the other two for compliance-bonus targets, under bounded utilities and three extra assumptions: full feedback while learning, restarts from chosen points in the game, and a stronger form of no-regret (Section 6). Its steering step needs a margin on every path, which cheaper certified candidates have only in expectation; steering them is left open. Small-game diagnostics test each component separately.

Section 2 defines the model precisely; the remaining sections follow the order of these questions.

2 Model and Benchmark

Let $\Gamma = (N, H, Z, \mathcal{I}, \mathcal{A}, \chi, u_0, u_1, \dots, u_n)$ be a finite perfect-recall extensive-form game. The principal knows the tree $(N, H, Z, \mathcal{I}, \mathcal{A})$, the chance process χ, and its own payoff u_0, but not the agents' utilities $u_1, \dots, u_n$. Where stated, utilities are bounded by a known $U_{\max}$, $|u_i(z)| \le U_{\max}$. The game is repeated for T rounds: in each round the principal chooses private recommendations and nonnegative payments, the agents choose realization plans x_i^t, a terminal history z^t is sampled and observed, and the principal receives $u_0(z^t) - \sum_i q_i^t$, where q_i^t is paid to player i.

Definition 2.1 (Payments and benchmark)**.** Recommendations and payments. *A recommendation distribution μ samples a private recommendation profile r. A payment rule $q = (q_i)_i$ assigns each player i a payment $q_i(r, z) \ge 0$ determined by the recommendation profile and the observed terminal history; we call such rules terminal-history-dependent. Nonnegativity is a limited-liability convention.*

Benchmark. *Under obedience, (μ, q) induces a distribution over recommendation-terminal pairs (r, z). The principal's benchmark value is*

$$V_0(\mu, q) := \mathbb{E}_{\mu,\mathrm{ob}}\left[u_0(z) - \sum_i q_i(r, z)\right],$$

so its own payments are charged as costs.

A candidate of Section 1 is such a pair (μ, q), and its value is $V_0(\mu, q)$. Whether it can be trusted depends on how the agents can deviate from its recommendations.

Definition 2.2 (Trigger deviations, gains, and robustness)**.** *Let $\mathcal{D}_i^{\mathrm{IC}}$ be player i's proper trigger deviations. A private recommendation context is one of player i's information sets together with the private recommendations it has received so far. Each $d \in \mathcal{D}_i^{\mathrm{IC}}$ obeys every recommendation before one such context, its* trigger context $\mathsf{C}_{i,d}$*, chooses a nonrecommended action there, and then follows a continuation rule with no further recommendations. If $\Pr_\mu(\mathsf{C}_{i,d}) > 0$, define the context-conditioned values*

$$V_i^u(\mu, q; a \mid \mathsf{C}_{i,d}) := \mathbb{E}_{\mu,a}\left[u_i(z) + q_i(r, z) \mid \mathsf{C}_{i,d}\right], \qquad a \in \{\mathrm{obey}, d\},$$

where $\mathbb{E}_{\mu,d}$ applies d for player i at that context and keeps the other players obedient. The trigger-deviation gain is

$$\Delta_{i,d}^u(\mu, q) := V_i^u(\mu, q; d \mid \mathsf{C}_{i,d}) - V_i^u(\mu, q; \mathrm{obey} \mid \mathsf{C}_{i,d}).$$

Zero-probability trigger contexts are assigned gain 0 and impose no incentive constraint. Incentive constraints quantify over these deviations, so we also call them incentive deviations and define the active incentive-deviation set $\mathcal{D}_i^+(\mu) := \{d \in \mathcal{D}_i^{\mathrm{IC}} : \Pr_\mu(\mathsf{C}_{i,d}) > 0\}$ and the active trigger-context set $\mathcal{C}_i^+(\mu) := \{\mathsf{C}_{i,d} : d \in \mathcal{D}_i^+(\mu)\}$. The pair (μ, q) is payment-implementable for u if payments are nonnegative and $\Delta_{i,d}^u(\mu, q) \le 0$ for all i and $d \in \mathcal{D}_i^+(\mu)$, and ρ-robust if $\Delta_{i,d}^u(\mu, q) \le -\rho$ for all such (i, d).

For instance, if a candidate recommends inspection and pays a bonus for full compliance, a buyer told to inspect who buys without inspecting gains its payoff difference over obeying, minus the forfeited bonus; this deviation loses exactly when the bonus exceeds that difference. Checking only active trigger deviations loses nothing: the gain of any causal deviation plan, one that reacts only to what the agent has seen, is a probability-weighted sum of their gains, so no such plan gains if none of them does (Lemma A.2). It remains to state what we assume about the agents, whose regret is also measured against obedience comparators used only in the playback construction of Section 6.

Definition 2.3 (Contextual causal regret)**.** *A* playback context *is a move-tagged information-set/recommended-action pair; the obedience comparator at a playback context c replaces player i's entire round plan by obedience. Let $\mathcal{D}_i^{\mathrm{ob}}$ contain one such comparator per playback context and set*

$\mathcal{D}_i^{\mathrm{reg}} := \mathcal{D}_i^{\mathrm{IC}} \cup \mathcal{D}_i^{\mathrm{ob}}$. *Player i has contextual causal regret at most* $\mathrm{Reg}_i(T)$ *if, for every relevant trigger or playback context and every matching* $d \in \mathcal{D}_i^{\mathrm{reg}}$*, the cumulative utility-plus-payment gain from applying* d *on rounds with that context is at most* $\mathrm{Reg}_i(T)$*. We assume* $\mathrm{Reg}_i(T) = o(T)$.

In words, on the rounds where an agent receives a given recommendation, its payoff, payments included, is at least that of any trigger deviation from it, up to a sublinear total.

3 What Incentives Reveal

The gains of Definition 2.2 are the objects to estimate, and they depend on the utilities only through a quotient. A frozen environment is the fixed distribution of play that one player faces when the other players, chance, recommendations, and payments are all held to prescribed rules from that point on; the sandboxed elicitation interface of Section 4 approximates such environments by finite phases. In a frozen environment the payoff coefficient is identified only modulo information-set potentials, and a centered payment game elicits one quotient representative; that representative is what enters active gains.

Fix player i and freeze the rest of the system into an environment η. Let Σ_i be player i's sequences, its own actions along a path, and let $X_i = \{x \in \mathbb{R}_+^{|\Sigma_i|} : E_i x = e_i\}$ be its realization plans. Each row of E_i is the flow constraint at one information set: the probability of the sequence leading into it equals the total probability of the sequences continuing from it. The root row fixes $x_\emptyset = 1$ for the empty sequence. Where a continuation may depend on private recommendations, sequences and information sets carry them, as in Definition 2.2 (Appendix A). Player i's expected utility is linear, $\langle c_i(\eta), x\rangle$, with a coefficient determined by the environment. Call c and c' equivalent if $c - c' \in \mathrm{Im}(E_i^\top)$.

Proposition 3.1 (Infoset-potential invariance)**.** *If $c' = c + E_i^\top \lambda$, then every feasible realization plan $x \in X_i$ satisfies*

$$\langle c', x\rangle = \langle c, x\rangle + \langle \lambda, e_i\rangle .$$

Consequently $\langle c', x - y\rangle = \langle c, x - y\rangle$ for all $x, y \in X_i$, so best responses, regret comparisons, and incentive constraints are unchanged by adding infoset-potential directions.

In the two-move example of Section 1, the constant added at the information set after R is such a direction, and at that information set the identifiable quantity is the coefficient difference along $Ra - Rb$. We fix a complement H_i to $\mathrm{Im}(E_i^\top)$ and write $h_i(\eta) \in H_i$ for the component of $c_i(\eta)$ in H_i along $\mathrm{Im}(E_i^\top)$, the quotient representative of Section 1 in this gauge; the gauge is arbitrary, and all guarantees are stated in terms of this representative.

To learn $h_i(\eta)$, the principal looks for a payment that leaves player i indifferent among all its realization plans; any such payment offsets the representative exactly. In the two-move example, it makes the buyer's plans L, Ra, and Rb equally good, offsetting the two independent payoff differences. The principal pays $\langle p, x_i\rangle$ for a vector p in a bounded slice $\mathcal{P}_i$: the complement H_i shifted by a fixed $a_i \in \mathrm{Im}(E_i^\top)$ and cut to nonnegative, bounded coordinates (Appendix C).

Let $x_i^\circ \in \mathrm{ri}(X_i)$ be a fixed reference point and define the centered payment game

$$\min_{p \in \mathcal{P}_i} \max_{x_i \in X_i} \langle c_i(\eta) + p, x_i - x_i^\circ\rangle .$$

Its value is nonnegative and, if a_i and the coordinate bound are large enough (Appendix C), vanishes exactly at $p_i^\star(\eta) = a_i - h_i(\eta)$ (Lemma C.4). The construction adapts the zero-sum payment game of Zhang et al. (2026) from mixed strategies in normal form to realization plans.

Theorem 3.2 (Frozen-environment elicitation)**.** *Fix a frozen environment η with $p_i^\star(\eta) \in \mathcal{P}_i$. At round t, the principal chooses $p_t \in \mathcal{P}_i$ and the player chooses $x_t \in X_i$. Suppose the principal has regret $R_i^P(T)$ for the observable losses $\ell_t(p) = \langle p, x_t - x_i^\circ\rangle$ and the player has regret $R_i^A(T)$ for perturbed utilities $\langle c_i(\eta) + p_t, x_i\rangle$. With slice coercivity constant $\gamma_i > 0$ (Appendix C), the average payment $\bar{p} = T^{-1}\sum_t p_t$ satisfies*

$$\|\bar{p} - p_i^\star(\eta)\| \le \frac{R_i^P(T) + R_i^A(T)}{\gamma_i T}.$$

Equivalently, the learned representative $\widehat{h}_i(\eta) = a_i - \bar{p}$ has the same error bound relative to $h_i(\eta)$.

Both regrets are interfaces of this game. The principal's losses need the player's plan x_t, which terminal histories alone do not reveal, so observing it is part of the interface (Appendix D.1, Lemma C.9).

The monitored construction supplies it by full feedback while learning, and learning from on-path observations alone is left open. The player's regret R_i^A is measured against fixed continuation plans, which the contextual regret of Definition 2.3 does not imply; the monitored construction assumes it blockwise (Assumption D.11).

The other requirements can be met (Appendix C). The coercivity constant is positive for every complement and interior reference point; with the reference point at the average of a set $\mathcal{S}_i^{\mathrm{test}}$ of pure test plans and the norm chosen there, $\gamma_i \geq 1/|\mathcal{S}_i^{\mathrm{test}}|$ (Lemmas C.5 and C.6). For representatives bounded by M_i, a $\mathcal{P}_i$ with coordinate bound polynomial in the explicit sequence-form size and in M_i contains every $p_i^\star(\eta)$ (Lemmas C.2 and C.3). Elicitation also tolerates rounds where the environment is not kept frozen: if they are an α_η fraction, the error bound grows by $\kappa_\eta \alpha_\eta$, with $\kappa_\eta \leq 4U_{\max}/\gamma_i$ when $|u_i| \leq U_{\max}$ (Lemma C.7). Section 4 uses this guarantee as the sandboxed elicitation interface.

4 One Target: From the Quotient to Steering

The repeated game is not frozen, so the learning of Section 3 must run inside sandbox phases, one per player and trigger context. Quotient estimates nevertheless suffice, because every active trigger gain is linear in the representative of a single frozen environment.

Proposition 4.1 (Gain linearization)**.** *Fix a target (μ, q). For each active trigger context $c \in \mathcal{C}_i^+(\mu)$, let $X_{i,c} = \{x \geq 0 : E_{i,c}x = e_{i,c}\}$ be the continuation realization-plan polytope rooted at c and let $\eta_{i,c}^{\mu,q}$ be the frozen continuation environment induced by the target at c. For every $d \in \mathcal{D}_i^+(\mu)$ with $c = \mathsf{C}_{i,d}$,*

$$\Delta_{i,d}^u(\mu, q) = \beta_{i,d}(\mu, q) + \left\langle h_{i,c}^u(\eta_{i,c}^{\mu,q}), y_{i,d}^{\mu,q} \right\rangle,$$

with $\beta_{i,d}(\mu, q)$ a known payment term, linear in q, and $y_{i,d}^{\mu,q} \in \ker(E_{i,c})$ fixed by the tree, μ, and d.

In the buyer example of Section 2, $\beta_{i,d}$ is minus the forfeited bonus and the inner product the payoff difference over obeying; the proof is in Appendix E (Lemmas E.1 and E.2). Learning in each trigger environment $\eta_{i,c}^{\mu,q}$ therefore bounds the error of an estimated gain: $|\langle \widehat{h}_{i,c} - h_{i,c}^u, y_{i,d}^{\mu,q} \rangle| \leq \|y_{i,d}^{\mu,q}\|_* \, \|\widehat{h}_{i,c} - h_{i,c}^u\|$.

Setting. The principal's strategy has a learning phase of T_{learn} rounds and a steering phase of $T_{\mathrm{steer}} := T - T_{\mathrm{learn}}$ rounds. In the learning phase, *sandboxed elicitation* takes a finite list $\mathcal{E}$ of frozen environments and returns estimates $\widehat{h}_i(\eta)$ within $\varepsilon_{\mathrm{el}} + \varepsilon_{\mathrm{sb}}$ of the true representatives, where $\varepsilon_{\mathrm{el}}$ is the elicitation error in an exactly frozen environment and $\varepsilon_{\mathrm{sb}}$ the additional error from approximating it by finite phases. *Gain recovery* turns these into estimates of every active gain with error $L_{\mathrm{cov}}(\varepsilon_{\mathrm{el}} + \varepsilon_{\mathrm{sb}})$. In the steering phase, *known-utility steering* takes a target whose active gains have *certified margin* $\rho_{\mathrm{cert}} := \rho - L_{\mathrm{cov}}(\varepsilon_{\mathrm{el}} + \varepsilon_{\mathrm{sb}})$ and plays it. Each interface declares its rounds, failure probability, and payment bounds, B_{sb} and $B_{\mathrm{steer}}(\rho_{\mathrm{cert}})$ on one-round total payments. The steering bound uses the aggregate regret quantity $\mathrm{Reg}(T_{\mathrm{steer}})$, and B dominates both $|V_0(\mu^\star, q^\star)|$ and the principal's one-round payoff-plus-payment magnitude during learning. Theorem 3.2 bounds $\varepsilon_{\mathrm{el}}$, and Section 6 supplies the rest for compliance-bonus targets under further premises.

Assumption 4.2 (Sandboxed elicitation interface)**.** *For any finite certificate-visible list $\mathcal{E}$ of requested frozen environments, elicitation tolerance $\varepsilon_{\mathrm{el}}$, sandboxing discrepancy $\varepsilon_{\mathrm{sb}}$, and failure budget δ_{sb}, there is a sandboxed elicitation protocol using bounded nonnegative payments based on recommendations and observed terminal histories, together with any additional monitoring declared by the protocol, that outputs estimates $\widehat{h}_i(\eta)$ satisfying, simultaneously for every $(i, \eta) \in \mathcal{E}$,*

$$\|\widehat{h}_i(\eta) - h_i^u(\eta)\| \leq \varepsilon_{\mathrm{el}} + \varepsilon_{\mathrm{sb}}$$

with probability at least $1 - \delta_{\mathrm{sb}}$. The protocol declares its total number of rounds, its allocation across $\mathcal{E}$, and a one-round total sandboxing-payment bound B_{sb}.

Assumption 4.3 (Local coverage: gain recovery from elicited representatives)**.** *For a fixed target (μ, q), a finite family of sandboxed continuation environments determines every active incentive-deviation gain $\Delta_{i,d}^u(\mu, q)$ with $d \in \mathcal{D}_i^+(\mu)$ from the elicited local representatives, with Lipschitz error $L_{\mathrm{cov}}(\varepsilon_{\mathrm{el}} + \varepsilon_{\mathrm{sb}})$.*

Appendix E formalizes this interface as a scalar-query certificate: each query asks for one number, such as the inner product in Proposition 4.1. Querying that inner product for every active gain gives

$L_{\mathrm{cov}} = \max_{i,d} \|y^{\mu,q}_{i,d}\|_*$ (Remark E.6). The algebraic reference-action baseline bounds this dual norm by $2|\Gamma_{\mathrm{rec}}|L^{\max}_{\mathrm{ref}}$, with $L^{\max}_{\mathrm{ref}}$ the largest dual norm of its basis vectors (Theorem E.12). Here $|\Gamma_{\mathrm{rec}}|$ is the size of the explicit recommendation-augmented description (Appendix A), which can be exponentially larger than $|\Gamma|$. Representatives that answer the queries alike may still disagree on the target's robustness, which is what coverage excludes (Proposition E.15).

Assumption 4.4 (Known-utility steering interface)**.** *For $T_{\mathrm{steer}} \geq 1$, when a known-utility steering procedure is given estimated gains $\widehat{\Delta}_{i,d}(\mu, q) \leq -\rho_{\mathrm{cert}}$ for every i and $d \in \mathcal{D}_i^+(\mu)$, with $\rho_{\mathrm{cert}} > 0$, the procedure implements the fixed target (μ, q) for T_{steer} rounds and guarantees*

$$\bar{V}_{\mathrm{steer}} \geq V_0(\mu, q) - O\left(\frac{\mathrm{Reg}(T_{\mathrm{steer}})}{\rho_{\mathrm{cert}} T_{\mathrm{steer}}}\right)$$

in the true repeated game against the actual agents, where $\bar{V}_{\mathrm{steer}}$ is the average realized value; applying a known-utility guarantee to the surrogate game defined by the estimated gains does not satisfy this interface. The procedure declares its failure probability, a nonnegative payment domain $\mathcal{Q}_{\mathrm{steer}}(\rho_{\mathrm{cert}})$, and a one-round total payment bound $B_{\mathrm{steer}}(\rho_{\mathrm{cert}})$, all at the same margin ρ_{cert}.

Theorem 4.5 (Fixed-target unknown-utility EFCE steering)**.** *Fix $\rho > 0$, $\delta \in [0, 1)$, and $T > T_{\mathrm{learn}} \geq 0$. Let $(\mu^\star, q^\star)$ be a ρ-robustly feasible payment-implementable EFCE target for the true game, and assume Assumptions 4.2–4.4 hold with total failure probability at most δ, with the quantities of the Setting. If $L_{\mathrm{cov}}(\varepsilon_{\mathrm{el}} + \varepsilon_{\mathrm{sb}}) \leq \rho/2$, then the principal has a two-phase learning-then-steering strategy such that*

$$\bar{V}_T \geq V_0(\mu^\star, q^\star) - O\left(\frac{\mathrm{Reg}(T_{\mathrm{steer}})}{\rho_{\mathrm{cert}} T_{\mathrm{steer}}}\right) - \frac{2BT_{\mathrm{learn}}}{T}$$

with probability at least $1 - \delta$, where $\bar{V}_T$ is the average realized value over all T rounds.

Proof idea. When all three interfaces succeed, elicitation and gain recovery give every active gain to within $L_{\mathrm{cov}}(\varepsilon_{\mathrm{el}} + \varepsilon_{\mathrm{sb}})$, so the true margin ρ shrinks to $\rho_{\mathrm{cert}} \geq \rho/2$ for the estimated gains (Corollary D.7), the step that certified selection in Section 5 runs in reverse. The ρ_{cert} in the denominator may therefore be read as ρ up to the constant in the $O(\cdot)$. Assumption 4.4 then applies at ρ_{cert}, and averaging its bound with the learning phase adds $2BT_{\mathrm{learn}}/T$; Appendix F has the full argument. With local coverage instantiated by the algebraic baseline, $L_{\mathrm{cov}} \leq \mathrm{poly}(|\Gamma_{\mathrm{rec}}|)$ if $L^{\max}_{\mathrm{ref}} \leq \mathrm{poly}(|\Gamma_{\mathrm{rec}}|)$, and the guarantee holds whenever $2|\Gamma_{\mathrm{rec}}|L^{\max}_{\mathrm{ref}}(\varepsilon_{\mathrm{el}} + \varepsilon_{\mathrm{sb}}) \leq \rho/2$ (Corollary F.1).

5 A Class of Targets: Certify, Select, and Why It Pays

Theorem 4.5 takes the robustness of one candidate as given. Robustness depends on the unknown utilities, so the principal cannot know a priori which candidates have it; it can, however, certify robustness from one set of gain estimates, for a whole known class at once.

Theorem 5.1 (Certified selection and steering over a known target class)**.** *Let $\mathcal{T}$ be a fixed nonempty finite class of candidate targets $s = (\mu, q)$ with $q \geq 0$, known to the principal. Write $V_0(s) := V_0(\mu, q)$, $\Delta^u_{i,d}(s) := \Delta^u_{i,d}(\mu, q)$, and let $A(s) := \{(i, d) : d \in \mathcal{D}_i^+(\mu)\}$, also written $A(\mu)$ since it depends on s only through μ. Suppose that, on an event $\mathcal{E}_{\mathrm{cov}}$ of probability at least $1 - \delta_{\mathrm{cov}}$, class-uniform estimates satisfy $|\widehat{\Delta}_{i,d}(s) - \Delta^u_{i,d}(s)| \leq \varepsilon_{\mathrm{cov}}$ for every $s \in \mathcal{T}$ and $(i, d) \in A(s)$. For $\rho_{\mathrm{sel}} > 0$, whenever the certified set is nonempty, choose*

$$\widehat{s} \in \operatorname*{argmax}_{s \in \mathcal{T}} V_0(s) \quad \text{subject to} \quad \widehat{\Delta}_{i,d}(s) \leq -\rho_{\mathrm{sel}} \quad \text{for every } (i, d) \in A(s),$$

using a fixed deterministic tie-breaking rule, so that $\widehat{s}$ is measurable with respect to the learning output. Then, on $\mathcal{E}_{\mathrm{cov}}$:

- *(i)* Safety. *$\Delta^u_{i,d}(\widehat{s}) \leq -\rho_{\mathrm{sel}} + \varepsilon_{\mathrm{cov}}$ for every $(i, d) \in A(\widehat{s})$; in particular $\widehat{s}$ is payment-implementable for the true utilities whenever $\rho_{\mathrm{sel}} > \varepsilon_{\mathrm{cov}}$.*
- *(ii)* Oracle comparison. *$V_0(\widehat{s}) \geq V_0(s^\star)$ for every $s^\star \in \mathcal{T}$ with $\Delta^u_{i,d}(s^\star) \leq -(\rho_{\mathrm{sel}} + \varepsilon_{\mathrm{cov}})$ for every $(i, d) \in A(s^\star)$.*
- *(iii)* Steering. *If $\rho_{\mathrm{sel}} \geq 3\varepsilon_{\mathrm{cov}}$, the certified set is nonempty throughout $\mathcal{E}_{\mathrm{cov}}$, and the known-utility steering interface of Assumption 4.4 is satisfied at $\widehat{s}$ with certified margin $\rho_{\mathrm{cert}} := \rho_{\mathrm{sel}} - 2\varepsilon_{\mathrm{cov}}$ and failure budget δ_{steer} (Proposition D.17), then with $T = T_{\mathrm{learn}} + T_{\mathrm{steer}}$ and probability at least $1 - \delta_{\mathrm{cov}} - \delta_{\mathrm{steer}}$,*

$$\bar{V}_T \geq V_0(\widehat{s}) - O\left(\frac{\mathrm{Reg}(T_{\mathrm{steer}})}{\rho_{\mathrm{cert}} T_{\mathrm{steer}}}\right) - \frac{2BT_{\mathrm{learn}}}{T},$$

where T_{learn} and $B \geq \max_{s \in \mathcal{T}} |V_0(s)|$ refer to the class-wide learning phase.

In the platform example, where the class is several bonus levels, the principal selects the cheapest level the estimates certify. The oracle comparison is margin-shifted: $\widehat{s}$ is ranked only against candidates whose margin exceeds the certification threshold by the estimation error. The class-uniform event is not an extra assumption: invoking Assumption 4.2 once on the union of the candidates' trigger-environment lists $\mathcal{E}_{\mathrm{trig}}(\mu)$ (Corollary D.4) gives it with $\varepsilon_{\mathrm{cov}} = \max_{s\in\mathcal{T}} L_{\mathrm{cov}}(s)(\varepsilon_{\mathrm{el}} + \varepsilon_{\mathrm{sb}})$ and $\delta_{\mathrm{cov}} = \delta_{\mathrm{sb}}$, and the algebraic baseline makes $L_{\mathrm{cov}}(s) \le 2|\Gamma_{\mathrm{rec}}|L_{\mathrm{ref}}^{\max}$ uniform over the class (Corollary D.18). Appendix D.6 has the proof.

A *compliance-bonus rule* pays each agent a fixed bonus M when the observed play shows that it obeyed at every reached recommendation (the event $\mathsf{E}_i(r,z) = 1$); compliance is *verifiable* when recommendations and terminal history reveal every active deviation, so that such a deviation forfeits the bonus with certainty (Definition F.2 calls this deterministic forfeit).

Proposition 5.2 (Elicitation buys value, not feasibility)**.** *Suppose $|u_i(z)| \le U_{\max}$, compliance is verifiable for the active deviations of a recommendation rule μ, and the class contains the compliance-bonus rule (μ, q^M) with $q_i^M(r,z) = M\mathbf{1}\{\mathsf{E}_i(r,z) = 1\}$ and $M \ge 2U_{\max} + \rho_{\mathrm{sel}} + \varepsilon_{\mathrm{cov}}$. Then on $\mathcal{E}_{\mathrm{cov}}$ the certified set of Theorem 5.1 is nonempty and $V_0(\widehat{s}) \ge V_0(\mu, q^M) = \mathbb{E}_{\mu,\mathrm{ob}}[u_0(z)] - nM$: the selected target is never worse than the bonus rule, which needs no elicitation to be safe, and strictly better whenever a cheaper payment rule for the same μ is certified.*

Proof. By Lemma F.4 the bonus rule's true active margin is at least $\rho_{\mathrm{sel}} + \varepsilon_{\mathrm{cov}}$, so on $\mathcal{E}_{\mathrm{cov}}$ its estimated margin is at least ρ_{sel} (Corollary D.7) and it is certified; part (ii) gives the inequality, and under obedience every agent is paid M. □

Certification costs the learning that Assumption 4.2 declares for the class's trigger environments, which ignore payments: a class varying only payments costs nothing extra and need not be finite.

Corollary 5.3 (One elicitation pass certifies a payment polyhedron)**.** *Fix a recommendation rule μ and a set Q of nonnegative payment rules with $\sum_i q_i(r,z) \le B_Q$ for every (r,z). Invoke Assumption 4.2 once on $\mathcal{E}_{\mathrm{trig}}(\mu)$ and instantiate Assumption 4.3 through a certificate whose scalar labels and coefficients depend on the target only through μ; the reference-action certificate of Theorem E.12 is one. Let $\widehat{g}_{i,d}$ denote the resulting estimate of the utility-linear part of the gain, so that $\widehat{\Delta}_{i,d}(\mu,q) = \beta_{i,d}(\mu,q) + \widehat{g}_{i,d}$ by Lemma E.1. Then, on the success event of Assumption 4.2,*

$$\sup_{q\in Q} \max_{(i,d)\in A(\mu)} \left|\widehat{\Delta}_{i,d}(\mu,q) - \Delta_{i,d}^{u}(\mu,q)\right| \le L_{\mathrm{cov}}(\varepsilon_{\mathrm{el}} + \varepsilon_{\mathrm{sb}}) =: \varepsilon_{\mathrm{cov}},$$

with no dependence on Q. Consequently parts (i) and (ii) of Theorem 5.1 hold for the class $\{\mu\} \times Q$ with the same $(\varepsilon_{\mathrm{cov}}, \delta_{\mathrm{cov}})$ whenever the certified program attains its maximum, and part (iii) applies to the selected target. Moreover $\beta_{i,d}(\mu,\cdot)$ is linear and $V_0(\mu,\cdot)$ affine in q, so when Q is a polyhedron the certified set $\widehat{Q}_{\mathrm{sel}} := \{q \in Q : \beta_{i,d}(\mu,q) + \widehat{g}_{i,d} \le -\rho_{\mathrm{sel}} \text{ for every } (i,d) \in A(\mu)\}$ is a polyhedron and $\max_{q\in\widehat{Q}_{\mathrm{sel}}} V_0(\mu,q)$ is a linear program in the coordinates $q_i(r,z)$, whose maximum, whenever $\widehat{Q}_{\mathrm{sel}}$ is nonempty, is attained and can be selected by a deterministic vertex rule.

Section 6 records where our construction steers the selected candidate and where it stops.

6 Instantiation and Diagnostics

The interfaces of Section 4 are not empty. Gain recovery is satisfied by the algebraic reference-action construction with $L_{\mathrm{cov}} \le 2|\Gamma_{\mathrm{rec}}|L_{\mathrm{ref}}^{\max}$ (Theorem E.12). Sandboxing and steering are satisfied together by a *monitored construction* for compliance-bonus targets q^M with verifiable compliance and $M \ge 2U_{\max} + \rho$. It needs bounded utilities and three further premises beyond the base model: the principal observes full feedback during learning, so that it can evaluate the payment rules and verify the other players' obedience; it can start continuation episodes from any trigger context, each counted as a learning round (Assumption D.10); and every agent's regret against fixed continuation plans, in particular against the prescribed plans, holds blockwise (Assumption D.11). The construction sandboxes each trigger environment by paying the non-target players a bonus $M_{\mathrm{sb}} > 2U_{\max}$ for following the prescribed continuation, giving $\varepsilon_{\mathrm{sb}} \le \max_\eta \kappa_\eta \alpha_\eta$ with $\kappa_\eta \le 4U_{\max}|\mathcal{S}_{i,c}^{\mathrm{test}}|$ and α_η the observed mismatch fraction. It steers by playback of the target. Under $\widetilde{O}(\sqrt{t})$ blockwise rates, $T_{\mathrm{learn}} = \widetilde{O}(K_{\mathrm{trig}}\rho^{-2})$ rounds over the $K_{\mathrm{trig}} = \sum_i |\mathcal{C}_i^+(\mu)|$ trigger environments suffice for

$\rho_{\mathrm{cert}} \geq \rho/2$ (Corollary F.6). The complete statement, with payment bounds $B_{\mathrm{sb}} \leq (n-1)M_{\mathrm{sb}}$ and $B_{\mathrm{steer}} \leq nM$ and the end-to-end bound, is Corollary F.5; it is not a premise of Theorem 4.5.

The same construction marks the boundary. Its steering step, playback (Proposition D.14), needs a margin that holds along every path, which Lemma F.4 supplies for compliance-bonus rules with $M \geq 2U_{\max} + \rho_{\mathrm{cert}}$. A cheaper certified candidate carries only a certified expected margin, which is what Assumption 4.4 uses but is not by itself the pathwise condition (Remark D.15). The paper therefore certifies every member of the class and steers, with its own construction, the pathwise-margin members. Whether a certified expected margin suffices under terminal-record feedback with a bounded, non-vanishing payment below that level is open here, and Section 7 locates the two known-utility routes.

The experiments are theorem-facing diagnostics on toy games, not a benchmark study. Candidates are built offline in a simulator with the agents' utilities hidden from the online principal, and three components are tested separately: the elicitation game with the monitored sandbox, the downstream gain-estimation-to-steering pipeline, and the payment–value sweep read as a selection experiment. No experiment runs all components in one pipeline, implements resets from any trigger context, or tests adaptive online target discovery; full tables and figures are in Appendix G.

Component diagnostic. The frozen continuation is the two-action simplex with a one-dimensional gauge and coercivity $\gamma = 1$. Projected online gradient descent chooses the principal's payments and the target player runs exponential weights. The experiment computes both realized regrets exactly and checks $\|\widehat{h} - h\| \leq (R^P + R^A)/(\gamma T)$ at every prefix rather than injecting a synthetic error, and at $T = 4096$ the realized error is 0.022 against a bound of 0.035. The sandbox diagnostic adds a non-target exponential-weights learner over four complete plans on two continuation contexts, pays M_{sb} only when a full-policy audit verifies the prescribed plan, and forms the endogenous certificate $\varepsilon_{\mathrm{sb}}^{\mathrm{cert}} = \kappa_\eta \alpha_\eta^{\mathrm{reg}}$ from the learner's realized regret, with $\kappa_\eta = 4$. Across 90 certified runs (ten seeds, three horizons, three bonuses above $2U_{\max}$) every prefix check passes, and below-threshold bonuses serve as negative controls. These checks use direct realization-plan feedback and full-policy monitoring; they do not implement terminal-history estimators, coverage, or steering in the same run (Table 2).

Downstream pipeline. On a signaling/response and an entry/inspection family, a certification stage produces noisy active-gain estimates with recorded maximum error $\varepsilon_{\mathrm{gain}} = 0.003$, combines them with an externally declared $\varepsilon_{\mathrm{sb}} = 0.01$ and the toy certificate's $L_{\mathrm{cov}} = 8$ into $\rho_{\mathrm{cert}} = 0.167$ from $\rho = 0.25$. Playback then steers contextual multiplicative weights and contextual regret matching for 2000 rounds within 0.005–0.034 of the payment-subtracted benchmark, with bad-round fractions at most 0.011 (Appendix G). Monitoring, comparator coverage, and deterministic forfeit are hard-coded in the simulator; the sandbox output-error construction is not simulated.

Payment–value sweep and selection readout. Table 1 sweeps the compliance bonus M on both families. Low bonuses have negative certified margin and bad-round fractions near 0.27; larger bonuses certify and steer stably while the benchmark falls, because q^M is charged inside $V_0(\mu, q)$: reliability is purchased at benchmark cost. The tabulated ρ_{cert} is a fixed-target diagnostic from the simulator's true margin; the selection below uses the estimated margin instead. All rows share μ, so they form a payment family in the sense of Corollary 5.3. Read as a selection experiment, the rule of Theorem 5.1 at $\rho_{\mathrm{sel}} = 0.15$ selects $M = 0.8$ on the signaling family and $M = 0.95$ on the entry family. Their estimated margins are 0.248 and 0.300, and their benchmarks 2.40 and 1.95. The cheaper rows have higher benchmarks but are either the low bonuses above, which are not robust, or certified only below ρ_{sel}, like $M = 0.65$ on the signaling family at 0.100. Every larger certified bonus forgoes value (1.60 at $M = 1.2$; 1.25 at $M = 1.3$), which is what the a priori bonus certificate of Proposition 5.2 costs. Two caveats: each row ran its own certification stage, so the one-pass property of Corollary 5.3 is not exercised; and the toy utilities are deterministic per-context gains, so expected and pathwise margins coincide and the boundary above is not tested.

7 Related Work and Scope

Correlation and no-regret in extensive-form games. Regret matching and related no-regret procedures link adaptive play to correlated equilibrium in normal form (Hart & Mas-Colell, 2000).

family	M	target	realized	avg pay	ρ_{cert}	bad frac
signaling	0.550	2.900	1.713	0.953	-0.083	0.267
signaling	0.650	2.700	2.700	1.300	0.017	0.000
signaling	0.800	2.400	2.397	1.599	0.168	0.001
signaling	1.000	2.000	1.999	1.999	0.367	0.001
signaling	1.200	1.600	1.597	2.399	0.568	0.001
entry	0.650	2.550	1.148	1.126	-0.083	0.267
entry	0.950	1.950	1.956	1.900	0.217	0.000
entry	1.300	1.250	1.244	2.598	0.568	0.001

Table 1: Payment–value sweep for regret matching. Larger bonuses improve certified margins but are subtracted from the target benchmark.

EFCE extends correlated-equilibrium reasoning to sequential interaction through a mediator that reveals private recommendations only along the realized path (von Stengel & Forges, 2008), and uncoupled no-trigger-regret dynamics converge to it (Farina et al., 2022). We adopt the same incentive language but study a principal-controlled problem: certify and steer toward a class of recommendation/payment targets when the agents' utilities are unknown. Our three deviation roles (incentives and coverage, playback regret counting, the learner's comparator class) follow the view that deviation classes trade tractability against expressiveness (Morrill et al., 2021).

Steering with advice and payments. A mediator who knows the utilities can use advice and nonnegative payments to steer no-regret learners toward a desired equilibrium in an extensive-form game (Zhang et al., 2024). Two of its findings frame our steering interface. Under full strategy observation, a payment that hands each player the difference between its payoff against the target opponents and against the actual ones steers from an equilibrium margin with vanishing average payments. This is our sandboxing idea applied in the steering phase rather than the learning phase. Computing it requires the player's payoff on arbitrary opponent profiles, which with unknown utilities a finite spanning family of frozen environments would have to supply. Under trajectory feedback, the same work shows that bounded per-round payments cannot steer with vanishing average payment in general extensive-form games; our playback route lives on this side and accepts a non-vanishing compliance bonus, charged inside V_0, in exchange for a margin that holds pathwise. The model is also related to mechanism design with no-regret agents (Camara et al., 2020), to information design with known payoff primitives (Kamenica & Gentzkow, 2011), and to steering without payments, where knowing only that a learner has no regret does not suffice when its payoffs are unknown (Zhang et al., 2025).

Learning games through payments. In normal form each agent has one information set, and our information-set potentials reduce to the strategic equivalence up to which Zhang et al. (2026) learn a game: a term in each agent's utility that depends only on the other agents' actions. Our payment-subtracted benchmark uses the same accounting as their correlated equilibrium with payments, in which the principal's value is net of payments that need not vanish. We differ in two choices. They learn the whole game and steer to its principal-optimal outcome; we learn only what a known class of candidates needs, so our guarantee is relative to that class rather than to the best achievable value. They observe every agent's strategy each round and leave on-path observation in extensive form to future work; our steering construction uses only terminal histories, but our learning construction still uses full feedback, so learning from on-path observations remains open here too.

Scope. Selection is post-learning over a known class: a finite list, or a payment family over one recommendation rule; a continuous family of recommendation rules would need a covering or discretization argument. Steering beyond compliance-bonus targets is the open interface marked in Section 6. The monitored construction is one sufficient instance, under additional observability, conditional-continuation generative access, and blockwise fixed-plan regret; its episodes count as learning rounds, and its recommendation-augmented description can be exponentially larger than the game, so polynomial claims are relative to that description. The experiments are component checks on toy games: no run derives blockwise regret from contextual regret, runs coverage and steering together, or simulates the sandbox error.

AI USE STATEMENT

The authors conceived and led this work. We chose the research question, identified the literature on extensive-form game that it builds on, and decided the formulation, the results and the structure of the paper. Under our direction, generative AI tools helped refine the formulation by predicting outcomes of candidate formulations and testing them in exploratory simulations. GPT (OpenAI) assisted in the theory, by formulating intermediate claims, supplying proof steps and drafting proofs, which we reworked into their present form. Claude (Anthropic) assisted in designing the experiments with us, implementing and running them, and interpreting their results, as well as in drafting the text and figures and searching the literature. We verified the results. The proofs were also audited by the authors. We take full responsibility for the content of this work.

REPRODUCIBILITY STATEMENT

The model is defined in Section 2, and the main results are stated with their assumptions in Sections 3–5. The additional premises of the monitored construction are collected in Corollary F.5. Complete proofs are provided in the main text and Appendices A–F; the Appendix Roadmap links each main result to its proof. Section 6 and Appendix G describe the diagnostic experiments. The simulation code specifies the game instances, learning rules, parameters, and random seeds, and includes scripts for regenerating the reported tables and figures. The code will be made publicly available upon publication.

REFERENCES

Modibo K. Camara, Jason D. Hartline, and Aleck Johnsen. Mechanisms for a no-regret agent: Beyond the common prior. In *Proceedings of the 61st IEEE Symposium on Foundations of Computer Science*, pages 259–270, 2020.

Gabriele Farina, Andrea Celli, Alberto Marchesi, and Nicola Gatti. Simple uncoupled no-regret learning dynamics for extensive-form correlated equilibrium. *Journal of the ACM*, 69(6), 2022.

Sergiu Hart and Andreu Mas-Colell. A simple adaptive procedure leading to correlated equilibrium. *Econometrica*, 68(5):1127–1150, 2000.

Emir Kamenica and Matthew Gentzkow. Bayesian persuasion. *American Economic Review*, 101(6):2590–2615, 2011.

Daphne Koller, Nimrod Megiddo, and Bernhard von Stengel. Efficient computation of equilibria for extensive two-person games. *Games and Economic Behavior*, 14(2):247–259, 1996.

Harold W. Kuhn. Extensive games and the problem of information. In Harold W. Kuhn and Albert W. Tucker, editors, *Contributions to the Theory of Games II*, Annals of Mathematics Studies 28, pages 193–216. Princeton University Press, 1953.

Dustin Morrill, Ryan D'Orazio, Marc Lanctot, James R. Wright, Michael Bowling, and Amy R. Greenwald. Efficient deviation types and learning for hindsight rationality in extensive-form games. In *Proceedings of the 38th International Conference on Machine Learning*, PMLR 139:7818–7828, 2021.

Martin J. Osborne and Ariel Rubinstein. *A Course in Game Theory*. MIT Press, 1994.

Bernhard von Stengel and Francoise Forges. Extensive-form correlated equilibrium: Definition and computational complexity. *Mathematics of Operations Research*, 33(4):1002–1022, 2008.

Brian Hu Zhang, Gabriele Farina, Ioannis Anagnostides, Federico Cacciamani, Stephen Marcus McAleer, Andreas Alexander Haupt, Andrea Celli, Nicola Gatti, Vincent Conitzer, and Tuomas Sandholm. Steering no-regret learners to a desired equilibrium. In *ACM Conference on Economics and Computation (EC)*, 2024.

Brian Hu Zhang, Tao Lin, Yiling Chen, and Tuomas Sandholm. Learning a game by paying the agents. In *International Conference on Learning Representations*, 2026.

Yizhou Zhang, Yi-An Ma, and Eric Mazumdar. Learning to steer learners in games. In *Proceedings of the 42nd International Conference on Machine Learning*, 2025.

APPENDIX ROADMAP

Roadmap and verification paths. All headline conclusions and interface statements are in the main text; the appendices provide proofs, the concrete witnesses, accounting details, and explicitly marked secondary extensions. Path A (theorem audit): read the coverage route in Appendix E and the assembly in Appendix F, using Corollary D.4, Lemma D.6, and Corollaries D.7 and D.8 for the trigger-environment and margin handoffs. Path B (full dependency audit): Appendices A–C for quotient, gauge, and elicitation; Appendix D for interfaces, handoffs, and the certified-selection proofs; Appendix E for local coverage and its converse; Appendix F for assembly and the monitored construction; Appendix G for diagnostics.

Main text	Proof/support	Shortest path
Section 2, model and benchmark	Appendix A; Appendix D.1	Sequence-form primer and conventions; Definition A.1 and Lemma A.2 for why active proper trigger deviations suffice; observation and feedback regimes.
Section 3, what incentives reveal	Appendices B–C	Proposition B.3 (quotient); Lemma C.4 (value-zero payment); Lemmas C.5 and C.6 and the proof of Theorem 3.2; Lemma C.7 for the monitored construction; the terminal-history estimator is the optional last subsection.
Section 4, one target	Appendices D.1 and D.2; Appendix E.1; Appendix F	Lemmas E.1 and E.2 (gain linearization); Lemma E.5 and Theorem E.12 (coverage); the two-way handoffs above; the proof of Theorem 4.5 and Corollary F.1.
Section 5, a class of targets	Appendix D.6; Appendix F.3	Proofs of Theorem 5.1 and Corollary 5.3; Corollary D.18 (class-uniform event); Proposition D.17 (class-wide accounting); Lemma F.4 for Proposition 5.2.
Section 6, instantiation and diagnostics	Appendices D.3–D.5; Appendix F.3; Appendix E.2; Appendix G	Assumptions D.10 and D.11; Proposition D.12 (sandboxing) and Proposition D.14 with Remark D.15 (playback and its boundary); Corollary F.5 (complete ledger) and Corollary F.6 (learning horizon); Proposition E.15 (scalar-query obstruction); tables and figures for the diagnostics.

A SEQUENCE-FORM PRIMER AND NOTATION

Minimal route through Appendices A–C. Readers who only want the quotient/gauge and elicitation chain can read Appendix A for the sequence-form notation, Proposition 3.1 and Lemma B.2 for incentive invariance, Proposition B.3 for the gauge-fixed representative, and Theorem 3.2 for the centered frozen-environment elicitation guarantee.

Augmented sequence-form systems. The quotient arguments below require only perfect recall and the standard sequence-form flow equalities. For a player i, let Σ_i be the set of player-i sequences and let X_i be the perfect-recall sequence-form realization-plan polytope,

$$X_i = \{x_i \in \mathbb{R}_+^{|\Sigma_i|} : E_i x_i = e_i\}.$$

By convention throughout the paper, these are the sequences and information sets of the recommendation-augmented decision problem whenever a continuation rule may depend on the player's private recommendation history. Under Definition 2.2's rule those labels record the recommendations a player has actually received and whether recommendations are still being sent; obedient and deviating plans live in the same system, and recommendations already received stay observable after they stop. The same convention applies to every re-rooted continuation system $(X_{i,c}, E_{i,c}, e_{i,c})$. The algebra below uses only perfect recall and the displayed flow equalities, so it applies unchanged to each such augmented system.

Matrix and size conventions. The matrix E_i contains the root and information-set flow constraints. Information-set rows are oriented as parent sequence minus the sum of child sequences; the root row fixes the empty sequence. We use the standard move-tagged sequence encoding: every nonempty sequence is created at a unique player-i information-set row of E_i, and therefore appears as a child variable in exactly one such conservation row. Throughout the appendix, $|\Gamma_{\mathrm{rec}}|$ denotes the total

explicit size of the recommendation-augmented systems used by the certificate and is taken to dominate every listed $|\Sigma_{i,c}|$, $|\mathcal{I}_{i,c}|$, matrix dimension, and playback-context count. Polynomial claims are therefore explicit-representation-relative; they are polynomial in the base-game size when the augmented description has polynomial size. For nonnegative ledgers, maxima over empty certificate, environment, context, or atom families are defined as zero, as are the corresponding sums and round counts. If a continuation quotient space is zero-dimensional, its elicitation error is zero and no elicitation block or coercivity constant is required.

Payoff convention. The column space $\mathrm{Im}(E_i^\top)$ is the space of infoset-potential payoff vectors: adding a vector in this space shifts all feasible realization plans by the same constant payoff. This appendix uses the pairing $\langle c, x\rangle$ for the expected payoff of coefficient vector c against realization plan x.

Worked example: why only a quotient is identifiable. This is the two-move example of Section 1, worked through in sequence form. It is a sanity check for the quotient obstruction, not a proof burden for the main theorem. Consider one player i in a frozen environment. At the root, i chooses L or R. If L, the game ends. If R, the player reaches information set I and chooses a or b. In move-tagged sequence form, take

$$\Sigma_i = \{\emptyset, L, R, Ra, Rb\}.$$

A realization plan $x_i \in X_i$ satisfies $E_i x_i = e_i$ under the standard move-tagged encoding, where

$$E_i = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 1 & -1 & -1 & 0 & 0 \\ 0 & 0 & 1 & -1 & -1 \end{bmatrix}, \qquad e_i = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}.$$

The first row fixes $x_\emptyset = 1$; the other two rows are the flow constraints $x_\emptyset - x_L - x_R = 0$ and $x_R - x_{Ra} - x_{Rb} = 0$. For any vector of information-set potentials λ and any feasible x_i,

$$\left\langle E_i^\top \lambda, x_i \right\rangle = \langle \lambda, E_i x_i \rangle = \langle \lambda, e_i \rangle,$$

which is constant over X_i. Hence $c_i(\eta)$ and $c_i(\eta) + E_i^\top \lambda$ induce the same best-response comparisons. The trigger comparison between a and b at I is a kernel direction: $y = \mathbf{e}_{Ra} - \mathbf{e}_{Rb}$ satisfies $E_i y = 0$, and its gain is $\langle c_i(\eta), y\rangle = c_i(\eta)_{Ra} - c_i(\eta)_{Rb}$, so any potential term $E_i^\top \lambda$ cancels. This is why we elicit a gauge-fixed representative of the quotient $\mathbb{R}^{|\Sigma_i|}/\,\mathrm{Im}(E_i^\top)$ rather than a terminal utility table.

Causal deviation plans and first-divergence canonicalization. The following definition and lemma justify why the main text quantifies incentive constraints and local coverage over the active proper trigger deviations $\mathcal{D}_i^+(\mu)$ only.

Definition A.1 (Causal deviation plans)**.** *Fix (μ, q) and player i. Let $\mathcal{D}_i^{\mathrm{cau}}$ be the class of deterministic causal deviation plans for player i: recommendation-history-measurable action maps in the recommendation-augmented game. The obedient plan plays the recommended action at every decision. For $d \in \mathcal{D}_i^{\mathrm{cau}}$, define its unconditional gain by*

$$\Delta_i^u(\mu, q; d) := \mathbb{E}_{\mu,d}[u_i(z) + q_i(r, z)] - \mathbb{E}_{\mu,\mathrm{ob}}[u_i(z) + q_i(r, z)].$$

Proper trigger deviations in $\mathcal{D}_i^{\mathrm{IC}}$ are causal deviation plans that obey before their trigger context and then apply a continuation rule after the trigger. The larger class $\mathcal{D}_i^{\mathrm{cau}}$ is used only for canonicalization; incentive constraints and local coverage continue to quantify over $\mathcal{D}_i^+(\mu) \subseteq \mathcal{D}_i^{\mathrm{IC}}$. Randomized causal deviations are mixtures of deterministic plans and are handled by linearity of the gain.

Lemma A.2 (First-divergence canonicalization)**.** *Fix (μ, q) and player i. Any $d \in \mathcal{D}_i^{\mathrm{cau}}$ can be partitioned, under the obedient recommendation distribution, by the first private recommendation context at which it differs from obedience. Conditional on a positive-probability first-divergence context c, the continuation of that plan defines a proper trigger deviation $d_c \in \mathcal{D}_i^{\mathrm{IC}}$ with $\mathsf{C}_{i,d_c} = c$ that switches immediately at c, and*

$$\Delta_i^u(\mu, q; d) = \sum_c \Pr_\mu(c)\, \Delta_{i,d_c}^u(\mu, q),$$

where $\Pr_\mu(c)$ is the probability of the first-divergence event indexed by c, not merely the probability that the underlying information set is visited, and zero-probability first-divergence events contribute nothing. Consequently, imposing incentive constraints on all active proper trigger deviations in $\mathcal{D}_i^+(\mu)$ suffices to rule out profitable causal deviation plans.

Proof. Perfect recall makes the first context at which a causal deviation differs from obedience a well-defined private recommendation context. Before that context, the deviating and obedient processes coincide, so conditioning on the first-divergence event is the same in the obedient and deviating comparisons. At that context the original deviation chooses a nonrecommended action, and its subsequent continuation is exactly a proper trigger deviation with that first-divergence trigger. A rule that obeys at a candidate trigger and first deviates later is indexed by that later context instead. The first-divergence events are disjoint and cover all histories on which the deviation differs from obedience; on the complement, the gain is zero. Taking conditional expectations over this partition gives the displayed decomposition, and nonpositive active trigger-deviation gains make the weighted sum nonpositive. □

B Proofs for Quotient Identifiability

These proofs identify the quotient directions that do not affect incentive comparisons and then fix a concrete representative for the elicitation arguments that follow.

Definition B.1 (Sequence-form strategic equivalence)**.** *For a fixed environment η, two coefficient vectors $c_i(\eta), \widetilde{c}_i(\eta) \in \mathbb{R}^{|\Sigma_i|}$ are strategically equivalent if there is a vector λ such that*

$$\widetilde{c}_i(\eta) = c_i(\eta) + E_i^\top \lambda.$$

We write $[c_i(\eta)]$ for the equivalence class in $\mathbb{R}^{|\Sigma_i|}/\operatorname{Im}(E_i^\top)$.

Proof of Proposition 3.1. Since $E_i x_i = e_i$,

$$\left\langle E_i^\top \lambda, x_i \right\rangle = \langle \lambda, E_i x_i \rangle = \langle \lambda, e_i \rangle ,$$

which is independent of x_i. □

Lemma B.2 (Trigger-gain preservation)**.** *Fix an environment η and a proper incentive deviation $d \in \mathcal{D}_i^{\mathrm{IC}}$. Let $x_i^{\mathrm{ob}}(\eta)$ be the realization plan induced by obedience and let $x_i^d(\eta)$ be the realization plan induced by applying d, holding the recommendation environment fixed. If $\widetilde{c}_i(\eta) = c_i(\eta) + E_i^\top \lambda$, then*

$$\left\langle \widetilde{c}_i(\eta), x_i^d(\eta) - x_i^{\mathrm{ob}}(\eta) \right\rangle = \left\langle c_i(\eta), x_i^d(\eta) - x_i^{\mathrm{ob}}(\eta) \right\rangle .$$

Proof. Both $x_i^{\mathrm{ob}}(\eta)$ and $x_i^d(\eta)$ are feasible realization plans. Hence the contribution of $E_i^\top \lambda$ to the gain is

$$\left\langle E_i^\top \lambda, x_i^d(\eta) - x_i^{\mathrm{ob}}(\eta) \right\rangle = \left\langle \lambda, E_i x_i^d(\eta) - E_i x_i^{\mathrm{ob}}(\eta) \right\rangle = 0.$$

□

Gauge fixing. The invariance results identify the irrelevant row-space directions; we now choose one representative from each resulting equivalence class. Fix a linear complement

$$\mathbb{R}^{|\Sigma_i|} = H_i \oplus \operatorname{Im}(E_i^\top)$$

and let $Q_i : \mathbb{R}^{|\Sigma_i|} \to H_i$ be the projection with kernel $\operatorname{Im}(E_i^\top)$. The gauge-fixed representative is $h_i(\eta) = Q_i c_i(\eta)$.

Proposition B.3 (Unique gauge representative)**.** *For every coefficient vector $c_i(\eta)$, the vector $h_i(\eta)$ is the unique element of H_i in the equivalence class $[c_i(\eta)]$. Preserving $h_i(\eta)$ is therefore equivalent to preserving the sequence-form payoff quotient.*

Proof. Since Q_i has kernel $\operatorname{Im}(E_i^\top)$, the difference $c_i(\eta) - Q_i c_i(\eta)$ lies in $\operatorname{Im}(E_i^\top)$. Thus $Q_i c_i(\eta)$ is in the same equivalence class as $c_i(\eta)$. If $h, h' \in H_i$ are in the same class, then $h - h' \in H_i \cap \operatorname{Im}(E_i^\top) = \{0\}$, so $h = h'$. □

A basis of feasible differences. For quantitative bounds, fix a baseline action $b(I) \in A(I)$ at each information set. For a non-baseline action $a \neq b(I)$, let $z^{I,a}$ be the conditional baseline-continuation flow that enters I, chooses a, and thereafter follows the baseline continuation. Define

$$d_{I,a} = z^{I,a} - z^{I,b(I)}, \qquad \mathcal{U}_i = \{(I,a) : a \in A(I) \setminus \{b(I)\}\}.$$

Lemma B.4 (Local deviation basis)**.** *The collection $\{d_{I,a} : (I,a) \in \mathcal{U}_i\}$ is a basis of $\ker(E_i)$. Moreover, for each $u = (I,a)$ the non-baseline child sequence $\sigma(I)a$ is a unique pivot coordinate for d_u.*

Proof. The number of vectors is $\sum_{I\in\mathcal{I}_i}(|A(I)|-1) = \dim\ker(E_i)$. To prove independence, order information sets topologically and take the earliest I with a nonzero coefficient in a putative dependence. Inspect a non-baseline pivot coordinate $\sigma(I)a$ with nonzero coefficient. No later local deviation can affect this ancestor coordinate, no different action at I touches this child, and no earlier deviation appears by the choice of I. The pivot coordinate is therefore exactly that coefficient, a contradiction. Hence the vectors are independent and so form a basis. □

Lemma B.5 (Pure endpoint representation)**.** *For every $u \in \mathcal{U}_i$, there exist globally feasible pure realization plans $x_u^+, x_u^- \in X_i$ such that*

$$d_u = x_u^+ - x_u^-.$$

Proof. Let $u = (I, a)$. Construct both pure plans to follow the same predecessor sequence to I. At I, let x_u^+ choose a and let x_u^- choose the baseline action $b(I)$. After I both plans follow the baseline continuation, and outside this branch they coincide. Their difference is exactly the conditional continuation flow $z^{I,a} - z^{I,b(I)}$. □

Lemma C.5 uses these pure endpoints to turn the range of a feasible payoff difference into the coercivity bound required by the elicitation theorem.

C Proofs for Frozen-Environment Elicitation

Local reading guide. This appendix section has one theorem-facing elicitation route and one implementation-side witness. For the shortest theorem-support path, read the gauge-budget and zero-value chain and then the coercivity proof of Theorem 3.2. The optional terminal-history implementation is placed last because it is not needed for the shortest theorem audit.

C.1 Gauge Budget and Zero-Value Characterization

Fix a frozen environment η for player i, with payoff vector $c_i(\eta)$. The centered-elicitation interface is represented by a nonnegative linear perturbation $q_p(x_i) = \langle p, x_i\rangle$ over player i's realization plan. This notation records the expected perturbation used in the zero-sum elicitation game. A terminal-history-only implementation must supply payments or estimators whose conditional expectation and observable losses match these linear forms; a monitored implementation may instead provide the required feedback directly and record that monitoring in the corresponding interface witness.

Lemma C.1 (Exact flattening certificate)**.** *If there exist $p \geq 0$ and λ such that*

$$c_i(\eta) + p = E_i^\top \lambda,$$

then every feasible realization plan $x_i \in X_i$ gives the same perturbed payoff. Moreover, $-p$ is a representative of the quotient class $[c_i(\eta)]$.

Proof. For every $x_i \in X_i$,

$$\langle c_i(\eta) + p, x_i\rangle = \left\langle E_i^\top \lambda, x_i\right\rangle = \langle \lambda, e_i\rangle .$$

Also $c_i(\eta) - (-p) = E_i^\top \lambda$, so $-p$ is equivalent to $c_i(\eta)$. □

The exact certificate identifies the desired quotient representative. The affine slice below selects it uniquely while enforcing nonnegative bounded payments. Choose a row-space offset $a_i \in \mathrm{Im}(E_i^\top)$ and define the affine gauge slice $\widetilde{\mathcal{P}}_i = a_i + H_i$. The implementable payment set is

$$\mathcal{P}_i = \widetilde{\mathcal{P}}_i \cap [0, C_i]^{|\Sigma_i|}.$$

Under the budget condition $a_i - h_i(\eta) \in [0, C_i]^{|\Sigma_i|}$, the unique flattening payment in this slice is

$$p_i^\star(\eta) = a_i - h_i(\eta).$$

Lemma C.2 (Gauge budget reduction)**.** *Suppose $\|h_i(\eta)\|_\infty \leq M_i$ for every frozen environment used by the learning stage, and suppose $a_i(\sigma) \geq 2M_i$ for every $\sigma \in \Sigma_i$. With $C_i = \|a_i\|_\infty + M_i$, the canonical payment $p_i^\star(\eta) = a_i - h_i(\eta)$ lies in $\mathcal{P}_i$ for every such environment.*

Proof. Since $p_i^\star(\eta) - a_i = -h_i(\eta) \in H_i$, the payment lies in the affine section $\widetilde{\mathcal{P}}_i$. Coordinatewise, $p_i^\star(\eta)(\sigma) \geq 2M_i - M_i \geq 0$ and $p_i^\star(\eta)(\sigma) \leq \|a_i\|_\infty + M_i = C_i$. □

Lemma C.3 (Constructive row-space offset)**.** *For every $M_i \geq 0$, there is $a_i \in \mathrm{Im}(E_i^\top)$ such that $a_i(\sigma) \geq 2M_i$ for every sequence $\sigma \in \Sigma_i$ and $\|a_i\|_\infty \leq 2M_i S_i$ for a size parameter $S_i = \mathrm{poly}(|\Sigma_i|, |\mathcal{I}_i|)$.*

Proof. Use the standard sequence-form sign convention $x_i(\sigma(I)) - \sum_{a \in A(I)} x_i(\sigma(I)a) = 0$. Define subtree weights by

$$s(I) = 1 + \max_{a \in A(I)} \sum_{J : \sigma(J) = \sigma(I)a} s(J),$$

and set $S_i = 1 + \sum_I s(I)$. Let information-set potentials be $\lambda_I = -2M_i s(I)$ and choose the root potential so that the root coordinate equals $2M_i$. Then $a_i = E_i^\top \lambda$ satisfies, for a nonroot sequence $\sigma = \sigma(I)a$,

$$a_i(\sigma) = 2M_i \left(s(I) - \sum_{J : \sigma(J) = \sigma(I)a} s(J) \right) \geq 2M_i.$$

The same construction gives $|a_i(\sigma)| \leq 2M_i S_i$ for all coordinates. □

Let $x_i^\circ \in \mathrm{ri}(X_i)$ be a reference realization plan and define

$$\phi_i(p; \eta) = \max_{x_i \in X_i} \langle c_i(\eta) + p, x_i - x_i^\circ \rangle .$$

Lemma C.4 (Value-zero certificate)**.** *For any $p \in \mathcal{P}_i$, the centered value is nonnegative. It is zero if and only if $c_i(\eta) + p \in \mathrm{Im}(E_i^\top)$. Under the budget condition from Lemmas C.2 and C.3 in Appendix C, the unique value-zero payment is $p_i^\star(\eta) = a_i - h_i(\eta)$.*

Proof of Lemma C.4. The maximum is at least zero because $x_i = x_i^\circ$ is feasible. If $c_i(\eta) + p \in \mathrm{Im}(E_i^\top)$, infoset-potential invariance gives zero centered payoff for every feasible x_i. Conversely, suppose the maximum is zero and let $w = c_i(\eta) + p$. Since $x_i^\circ \in \mathrm{ri}(X_i)$, every $v \in \ker(E_i)$ is locally feasible in both signs: for some $\tau > 0$, $x_i^\circ \pm \tau v \in X_i$. The value-zero condition gives $\langle w, \tau v \rangle \leq 0$ and $\langle w, -\tau v \rangle \leq 0$, so $\langle w, v \rangle = 0$ for all $v \in \ker(E_i)$. Hence $w \in (\ker E_i)^\perp = \mathrm{Im}(E_i^\top)$. The uniqueness follows from the transversality of the affine slice. □

C.2 Coercivity and Proof of the Elicitation Theorem

Lemma C.5 (Slice coercivity)**.** *Fix any norm $\|\cdot\|$ on H_i and define*

$$\gamma_i := \min_{\substack{h \in H_i \\ \|h\| = 1}} \max_{x_i \in X_i} \langle h, x_i - x_i^\circ \rangle .$$

Then $\gamma_i > 0$. For every $p \in \widetilde{\mathcal{P}}_i$,

$$\phi_i(p; \eta) \geq \gamma_i \|p - p_i^\star(\eta)\|.$$

Proof. The unit sphere in H_i is compact and the displayed support function is continuous. If $\gamma_i = 0$, there is a nonzero $h \in H_i$ whose centered support function is zero. The same two-sided local-feasibility argument as above implies $h \in \mathrm{Im}(E_i^\top)$, contradicting $H_i \cap \mathrm{Im}(E_i^\top) = \{0\}$.

For the inequality, note that $p - p_i^\star(\eta) \in H_i$ and $c_i(\eta) + p_i^\star(\eta) \in \mathrm{Im}(E_i^\top)$. The row-space component drops from centered payoffs, so

$$\phi_i(p; \eta) = \max_{x_i \in X_i} \langle p - p_i^\star(\eta), x_i - x_i^\circ \rangle .$$

If $p \neq p_i^\star(\eta)$, apply the definition of γ_i to the normalized vector $(p - p_i^\star(\eta))/\|p - p_i^\star(\eta)\|$. □

Lemma C.6 (Endpoint reference lower bound)**.** *Let x_i^{bl} be the pure realization plan induced by the baseline actions. For every $u \in \mathcal{U}_i$, let x_u^+, x_u^- be the endpoint plans from the pure endpoint representation lemma and set*

$$\mathcal{S}_i^{\mathrm{test}} = \{x_i^{\mathrm{bl}}\} \cup \{x_u^+, x_u^- : u \in \mathcal{U}_i\}.$$

If x_i° is the uniform average of $\mathcal{S}_i^{\mathrm{test}}$, then $x_i^\circ \in \mathrm{ri}(X_i)$ and, for $\|h\|_{\mathrm{adv},\infty} := \max_{u \in \mathcal{U}_i} |\langle h, d_u \rangle|$,

$$\gamma_i \geq \frac{1}{|\mathcal{S}_i^{\mathrm{test}}|}.$$

Proof. By Lemma B.4, the vectors $\{d_u : u \in \mathcal{U}_i\}$ span $\ker(E_i)$. Hence $\|h\|_{\mathrm{adv},\infty} = 0$ implies $h \perp \ker(E_i)$, so $h \in \mathrm{Im}(E_i^\top)$. Since $h \in H_i$ and $H_i \cap \mathrm{Im}(E_i^\top) = \{0\}$, this implies $h = 0$. Thus $\|\cdot\|_{\mathrm{adv},\infty}$ is a norm on H_i.

By Lemma B.5, these endpoint plans are feasible and satisfy $x_u^+ - x_u^- = d_u$. The average is therefore feasible and gives positive mass to every nonroot sequence: baseline-only sequences are present in x_i^{bl}, and any sequence with a last non-baseline action is present in the corresponding endpoint plan. Thus $x_i^\circ \in \mathrm{ri}(X_i)$. If $\|h\|_{\mathrm{adv},\infty} = 1$, then for some u the two endpoint plans have payoff difference $|\langle h, x_u^+ - x_u^-\rangle| = 1$. Among a finite set of $n = |\mathcal{S}_i^{\mathrm{test}}|$ real numbers with range at least one, the maximum exceeds the average by at least $1/n$. This gives the displayed lower bound. □

Proof of Theorem 3.2. Define the notional zero-sum payoff

$$g_i(p, x; \eta) = \langle c_i(\eta) + p, x - x_i^\circ \rangle .$$

The principal's observable loss differs from $g_i(p, x_t; \eta)$ by a term independent of p, and the player's regret is unchanged by subtracting the round-dependent constant $\langle c_i(\eta) + p_t, x_i^\circ \rangle$. Combining the two regret inequalities and using bilinearity gives

$$\max_{x_i \in X_i} g_i(\bar{p}, x_i; \eta) - \min_{p \in \mathcal{P}_i} g_i(p, \bar{x}; \eta) \le \frac{R_i^P(T) + R_i^A(T)}{T},$$

where $\bar{x} = T^{-1} \sum_t x_t \in X_i$. Since $p_i^\star(\eta)$ is feasible and has zero centered payoff against every x_i, the minimum term is at most zero. Thus $\phi_i(\bar{p}; \eta) \le (R_i^P(T) + R_i^A(T))/T$, and Lemma C.5 gives the claim. □

Lemma C.7 (Corruption stability of centered elicitation)**.** *Fix a requested environment η and a block of T centered-elicitation rounds. Let $c_{i,t}$ be the conditional expected base-utility coefficient induced by the actual continuation environment faced by player i in round t, so $\langle c_{i,t}, x\rangle$ is the conditional expected base utility under candidate plan x. The candidate plan, including the actual choice x_t, is the argument of this functional rather than part of the environment defining $c_{i,t}$. The conditioning fixes the other players' realized policies as maps and averages the rest: writing $\Pi_{-i,t}$ for the complete contingent policies the other players realize in round t, and $\mathcal{G}_t$ for the pre-round history together with p_t and $\Pi_{-i,t}$,*

$$\langle c_{i,t}, x\rangle = \mathbb{E}\big[u_i\big(z(x, \Pi_{-i,t}, W_t)\big) \mid \mathcal{G}_t\big],$$

where W_t collects the remaining exogenous episode variables: the conditionally sampled prefix, the recommendation randomness, and chance. Conditioning on a policy map is not the same as conditioning on its value along one realized path, which would fix randomness that must stay averaged. Chance and other within-round randomness are integrated into the coefficient. Let $p_t \in \mathcal{P}_i$, $x_t \in X_i$, and let $\mathcal{B}_\eta \subseteq \{1, \dots, T\}$ be a set covering the coefficient-corrupted rounds:

$$\{t : c_{i,t} \ne c_i(\eta)\} \subseteq \mathcal{B}_\eta, \qquad \alpha_\eta := \frac{|\mathcal{B}_\eta|}{T}.$$

The coefficient $c_{i,t}$ excludes the centered perturbation p_t, displayed separately below. In the retained monitored route, the centered perturbation p_t is the only target-player transfer whose realization-plan-dependent coefficient varies during this block; every other transfer to the target player is suppressed, and the known target-payment contribution is reinserted analytically through $\beta_{i,d}(\mu, q)$. The monitored route also supplies the conditional-expected linear feedback directly, so $q_{p_t}(x) = \langle p_t, x\rangle$ remains exact on corrupted rounds. Suppose the principal has regret $R_i^P(T)$ for the observable centered losses from Theorem 3.2, while the target player has regret $R_i^A(T)$ for the corresponding sequence of round-t conditional-expected linear utilities $x_i \mapsto \langle c_{i,t} + p_t, x_i\rangle$. If

$$G_\eta \ge \sup_{t \in \mathcal{B}_\eta} \sup_{x, x' \in X_i} |\langle c_{i,t} - c_i(\eta), x - x'\rangle| ,$$

with the supremum over an empty $\mathcal{B}_\eta$ interpreted as zero, then the retained centered estimate $\widehat{h}_i(\eta) = a_i - \bar{p}$ satisfies

$$\|\widehat{h}_i(\eta) - h_i^u(\eta)\| \le \frac{R_i^P(T) + R_i^A(T)}{\gamma_i T} + \frac{G_\eta}{\gamma_i}\alpha_\eta.$$

Thus this implementation has corruption-sensitivity ledger G_η/γ_i. If $|u_i(z)| \le U_{\max}$, then one may take $G_\eta \le 4U_{\max}$, so the certificate-visible choice $\kappa_\eta := 4U_{\max}/\gamma_i$ is valid:

$$\kappa_\eta = \frac{4U_{\max}}{\gamma_i}.$$

Under the endpoint-reference norm of Lemma C.6, this further gives the certificate-visible magnitude bound

$$\kappa_\eta \le 4U_{\max}|\mathcal{S}_i^{\text{test}}|.$$

It is polynomial in the explicit sequence count when the declared payoff magnitude $U_{\max}$ is polynomially bounded.

Proof. Write $x_i^\circ = x^\circ$. Fix a realization of the coefficient and play sequences on which both stated regret inequalities hold; the argument below is deterministic in those coefficients and uses no distributional property of them. Adding the principal and player regret inequalities gives the time-varying saddle-gap bound

$$\max_{x\in X_i}\frac{1}{T}\sum_{t=1}^{T}\langle c_{i,t}+p_t, x-x^\circ\rangle - \min_{p\in\mathcal{P}_i}\frac{1}{T}\sum_{t=1}^{T}\langle c_{i,t}+p, x_t-x^\circ\rangle \le \frac{R_i^P(T)+R_i^A(T)}{T}.$$

Let $x^\star$ maximize the first term of the analogous target-environment gap, in which every $c_{i,t}$ is replaced by $c_i(\eta)$. The first maximum in the time-varying gap is at least its value at $x^\star$, while the two minimum terms differ only in their base-utility coefficients. Consequently, the target gap minus the time-varying gap is at most

$$\frac{1}{T}\sum_{t=1}^{T}\langle c_i(\eta)-c_{i,t}, x^\star-x_t\rangle\,.$$

The summands vanish outside $\mathcal{B}_\eta$ and are at most G_η on $\mathcal{B}_\eta$. The target-environment saddle gap is therefore at most $(R_i^P(T)+R_i^A(T))/T + G_\eta\alpha_\eta$. The target flattening payment $p_i^\star(\eta)$ is feasible and makes the centered payoff zero against every x, so the minimum side of this target gap is at most zero. Hence

$$\phi_i(\bar{p};\eta) \le \frac{R_i^P(T)+R_i^A(T)}{T} + G_\eta\alpha_\eta.$$

Lemma C.5 gives the claimed estimate.

Finally, if terminal utilities lie in $[-U_{\max}, U_{\max}]$, every feasible realization plan has expected base utility in that interval under both the actual round-t continuation environment and η. For any $x, x' \in X_i$, the displayed coefficient difference is therefore bounded by $4U_{\max}$. The final bound follows from Lemma C.6. □

Remark C.8 (Additional varying target-player transfers)**.** *The specialization $G_\eta \le 4U_{\max}$ uses the retained payment separation in Lemma C.7. If another target-player transfer has a realization-plan-dependent coefficient that varies across corrupted rounds, its coefficient-difference range must be added explicitly. Writing that range as G_η^{pay}, the valid ledger is*

$$G_\eta \le 4U_{\max} + G_\eta^{\text{pay}}, \qquad \kappa_\eta = \frac{4U_{\max}+G_\eta^{\text{pay}}}{\gamma_i}.$$

The retained monitored witness sets $G_\eta^{\text{pay}} = 0$ rather than hiding this quantity in the base-utility bound. The optional terminal-history estimator in Lemma C.9 establishes unbiasedness only; using it under corruption would additionally require its distortion and sampling residual in the declared ledger.

C.3 Optional Terminal-History Implementation

The following implementation-side witness realizes the sequence-coordinate perturbations from move-tagged terminal histories when all queried coordinates have positive environment reach.

Lemma C.9 (Terminal-history implementation for positive-reach sequence coordinates)**.** *Suppose terminal histories are move-tagged, so that from a realized terminal history z the principal can recover the player-i information sets reached on the path and the action chosen by player i at each such information set. This is an encoding requirement on the terminal record of the realized path, not observation of unreached information sets or additional online monitoring of counterfactual actions. For a nonempty player-i sequence $\sigma = (I, a)$, let $\rho_\eta(I)$ be the environment reach probability of I conditional on player i's predecessor sequence to I. Assume the positive reach probabilities used below are available to the implementation, for example as part of the frozen-environment certificate: the certificate may expose either a certificate-visible representation of the terminal-history distribution induced by η, from which the $\rho_\eta(I)$ values are computable on the known tree, or the required $\rho_\eta(I)$ values for the queried coordinates. This lemma does not cover additional error from*

estimating those probabilities; an implementation that estimates them must declare a conservative lower bound and record the resulting estimation error in its elicitation or sandboxing ledger. For every coordinate with $\rho_\eta(I) > 0$, define

$$\widehat{x}^{\eta}_{i,\sigma}(z) := \frac{\mathbf{1}\{I \text{ is reached in } z \text{ and player } i \text{ chooses } a \text{ at } I\}}{\rho_\eta(I)}, \qquad \widehat{x}^{\eta}_{i,\emptyset}(z) := 1.$$

Then, under the frozen environment η and any player realization plan x_i,

$$\mathbb{E}_{\eta,x_i}[\widehat{x}^{\eta}_{i,\sigma}(z)] = x_i(\sigma).$$

Consequently, if a nonnegative perturbation p is supported on the empty sequence and positive-reach nonempty coordinates, the terminal-history-dependent payment

$$\widehat{q}^{\eta}_{p,i}(r, z) := \langle p, \widehat{x}^{\eta}_{i}(z)\rangle$$

is nonnegative and has expectation $\langle p, x_i\rangle$, while the centered observable loss $\langle p, \widehat{x}^{\eta}_{i}(z) - x^{\circ}_i\rangle$ has expectation $\langle p, x_i - x^{\circ}_i\rangle$. Moreover, with

$$B^{\eta}_{\mathrm{el}}(p) := p_\emptyset + \sum_{\substack{\sigma=(I,a)\in\mathrm{supp}(p)\\ \sigma\neq\emptyset}} \frac{p_\sigma}{\rho_\eta(I)},$$

the one-round payment is at most $B^{\eta}_{\mathrm{el}}(p)$ and the centered loss magnitude is at most $B^{\eta}_{\mathrm{el}}(p) + \langle p, x^{\circ}_i\rangle \le B^{\eta}_{\mathrm{el}}(p) + \|p\|_1$. This quantity is the centered-elicitation payment/loss ledger declared by this implementation route. Coordinates with zero environment reach are not implemented by this route; they must be covered by another frozen environment or by an explicitly monitored implementation.

Proof. For the empty sequence the claim is immediate. Fix a nonempty sequence $\sigma = (I, a)$. Under the frozen environment, the probability that player i's predecessor sequence to I is played, the environment brings the play to I, and player i chooses a at I is $x_i(\sigma)\rho_\eta(I)$. Dividing the path indicator by $\rho_\eta(I)$ therefore gives the displayed unbiased coordinate estimator. Linearity gives the expectation identities for $\widehat{q}^{\eta}_{p,i}$ and for the centered loss. Nonnegativity follows from $p \ge 0$ and the coordinate estimators being nonnegative. The displayed payment bound follows by upper-bounding each path indicator by one, and the centered-loss bound follows from the triangle inequality and $0 \le x^{\circ}_i(\sigma) \le 1$ for feasible realization plans. □

D Interfaces, Witnesses, and Certified Selection

Minimal route through Appendix D. Readers should separate the theorem-facing interface route from the narrower witness routes. The interface route consists of Assumption 4.2 and Corollary D.4 for frozen environments, Lemma D.6 and Corollaries D.7 and D.8 for true/estimated margin handoffs, and Assumption 4.4 for the known-utility steering interface. Proposition D.12 and Proposition D.14 are narrower sufficient witness routes, not additional premises of the main theorem.

The proof isolates sandboxing, local coverage, and known-utility steering as explicit interfaces. This separates the unknown-utility reduction from the implementation details of each interface.

D.1 Interface Declarations in Full

The main-text statements of the three interfaces give each one's input, output, and declared ledgers. This subsection records the conventions and declarations that the witnesses rely on but that Theorem 4.5 does not consume.

Remark D.1 (Per-context versus aggregate contextual regret)**.** *The definition above uses a per-context regret bound. When a proof sums bad events over all private recommendation contexts, a factor $|\mathcal{C}_i| \le |\Gamma_{\mathrm{rec}}|$ appears. If one instead assumes an aggregate contextual regret bound, for example $\sum_{c\in\mathcal{C}_i} \mathrm{Reg}_{i,c}(T) \le \mathrm{Reg}^{\mathrm{agg}}_i(T)$, the same arguments replace this factor by the aggregate regret term. This convention is the source of the context-summing factor in the later playback bounds.*

Terminal-history-dependent payments: scope. In Definition 2.1, the observed terminal history means the terminal action/history information needed to evaluate $q_i(r, z)$. This term only concerns the dependence of $q_i(r, z)$ on terminal histories; it is distinct from the additional monitoring assumptions used by the monitored full-feedback witness in Appendix F. In particular, it does not assume additional online monitoring beyond the realized terminal record, nor observation of unreached information sets. Any compliance gating is part of the rule q itself; the playback procedure does not add an extra payment filter after choosing (μ, q).

Observation and feedback regimes. The base target/payment model and the later learning interfaces use different feedback regimes. In Definition 2.1, the fixed target rule $q_i(r, z)$ is evaluated from recommendation-terminal information. Later centered-elicitation and known-utility steering statements are module interfaces. When those interfaces mention sequence-coordinate perturbations, observable losses, or full-feedback advice-and-payment, the required feedback belongs to that interface rather than to Definition 2.1. A concrete route must either realize the displayed linear perturbations and loss observations from terminal-history-dependent payments or terminal-history estimators, or explicitly provide additional monitoring as the monitored witness in Appendix F does. Thus the base model does not by itself grant observation of full realization plans, counterfactual actions, unreached information sets, or feedback beyond the realized terminal record. The interface does not fix a feedback regime: the playback witness of Proposition D.14 runs under the base terminal-record feedback, while the monitored witness uses full feedback upstream in its learning blocks (Appendix D.1).

Remark D.2 (Sandboxed elicitation: indices and ledgers)**.** *In Assumption 4.2 the list $\mathcal{E}$ may carry additional certificate-visible indices: a context-indexed entry $(i, c, \eta_{i,c})$ is identified with $(i, (c, \eta_{i,c}))$ in the generic notation, and its estimate and gauge are written $\widehat{h}_{i,c}$ and $h_{i,c}$. The protocol also declares any separate observable-loss magnitude bound required by its elicitation implementation. When a protocol first produces environment-specific ledgers, the theorem-facing scalars are*

$$\varepsilon_{\mathrm{el}} := \max_{(i,\eta)\in\mathcal{E}} \varepsilon_{\mathrm{el},i,\eta}, \qquad \varepsilon_{\mathrm{sb}} := \max_{(i,\eta)\in\mathcal{E}} \varepsilon_{\mathrm{sb},i,\eta}.$$

At the theorem level, this interface contributes the sandboxing discrepancy $\varepsilon_{\mathrm{sb}}$, the associated failure probability, and its declared round and magnitude ledgers; it does not by itself grant stronger observability. A concrete mismatch-counting witness must also declare how a behavioral-mismatch fraction controls coefficient-corrupted rounds and hence the displayed gauge-error ledger; Proposition D.12 does so through Lemma C.7.

Remark D.3 (Known-utility steering: regret mapping, ledgers, and feedback)**.** *The aggregate $\mathrm{Reg}(T_{\mathrm{steer}})$ in Assumption 4.4 may be value-scaled and need not equal a raw sum of agent regrets; each concrete implementation declares its own mapping, and the notation may suppress the declared dependence on the consumed margin and confidence level. The aggregate regret term, payment domain, and value guarantee must all refer to the same consumed margin and implementation. In a concrete end-to-end instantiation, the declared payment bound is polynomial in $|\Gamma_{\mathrm{rec}}|$ and $1/\rho_{\mathrm{cert}}$, and in the separately declared numerical payoff and payment-magnitude parameters. The steering feedback regime is implementation-specific: the interface does not itself require full feedback. The retained compliance-bonus playback rule uses an obedience event determined by the recommendation and observed terminal history, so its steering phase is implementable under the base recommendation–terminal-record feedback; the full-feedback interface of the monitored witness is used upstream by its sandboxed learning and coefficient-corruption accounting, not by Proposition D.14. The playback witness has no additional surrogate-to-true loss: it executes the fixed target rule itself, and its actual signed pathwise margin is at least every consumed margin $\rho_{\mathrm{e}} \le \rho$.*

D.2 Trigger Environments and Margin Handoffs

Corollary D.4 (Trigger environments are frozen-elicitable targets)**.** *Fix a target recommendation-and-payment pair (μ, q). For each player i, and each active trigger context $c \in \mathcal{C}_i^+(\mu)$, the frozen continuation environment $\eta_{i,c}^{\mu,q}$ used in Lemma E.2 may be included in the finite elicitation list of Assumption 4.2. Running the frozen-environment elicitation protocol in this sandboxed episode yields an estimate $\widehat{h}_{i,c}(\eta_{i,c}^{\mu,q})$ with the same aggregate error ledger used by local coverage:*

$$\|\widehat{h}_{i,c}(\eta_{i,c}^{\mu,q}) - h_{i,c}^{u}(\eta_{i,c}^{\mu,q})\| \le \varepsilon_{\mathrm{el}} + \varepsilon_{\mathrm{sb}}.$$

For a fixed target, the trigger-environment list is

$$\mathcal{E}_{\mathrm{trig}} := \{(i, c, \eta_{i,c}^{\mu,q}) : c \in \mathcal{C}_i^+(\mu)\}, \qquad K_{\mathrm{trig}} := |\mathcal{E}_{\mathrm{trig}}| = \sum_i |\mathcal{C}_i^+(\mu)|.$$

This count is finite and certificate-visible; it is polynomial whenever the explicitly represented recommendation-augmented continuation description has polynomial size.

Proof. Assumption 4.2 applies to every frozen environment used by the certificate, and $\eta_{i,c}^{\mu,q}$ is one such environment for each active pair (i, c). Within its sandboxed episode, Theorem 3.2 gives the

elicitation error $\varepsilon_{\mathrm{el}}$ relative to the frozen environment; the sandboxing discrepancy contributes $\varepsilon_{\mathrm{sb}}$ in the same norm ledger. Listing the active augmented contexts gives the displayed exact environment count; no polynomial bound relative to the unaugmented base game is used without a corresponding representation-size premise. □

Remark D.5 (Trigger environments versus sandbox construction)**.** *Corollary D.4 does not add a separate obedience wrapper to define $\eta_{i,c}^{\mu,q}$: the environment is the continuation law induced by the fixed advice-and-payment target (μ, q) conditional on the active trigger context c. The sandboxing module is responsible only for approximating this context-specific frozen environment when other learning agents are present. In the retained centered block, the known target-payment contribution is suppressed and reinserted analytically through $\beta_{i,d}(\mu, q)$; it is not an additional varying transfer hidden inside the base-utility coefficient.*

Structured local-coverage witnesses. Assumption 4.3 asks only for uniform estimates of the active incentive-deviation gains $\Delta_{i,d}^{u}(\mu, q)$ with $d \in \mathcal{D}_i^{+}(\mu)$. Theorem E.12 records a purely algebraic reference-action baseline that discharges this interface. This is the concrete local-coverage witness used by the retained construction.

Lemma D.6 (Robust transfer)**.** *Suppose an estimated utility profile $\widehat{u}$ satisfies*

$$\left|\Delta_{i,d}^{\widehat{u}}(\mu, q) - \Delta_{i,d}^{u}(\mu, q)\right| \le \varepsilon$$

for every player i and active incentive deviation $d \in \mathcal{D}_i^{+}(\mu)$. If (μ, q) is ρ-robustly feasible for $\widehat{u}$ over these active constraints, then

$$\Delta_{i,d}^{u}(\mu, q) \le -\rho + \varepsilon \qquad \textit{for all } i,\ d \in \mathcal{D}_i^{+}(\mu).$$

Proof. For every i and $d \in \mathcal{D}_i^{+}(\mu)$,

$$\Delta_{i,d}^{u}(\mu, q) \le \Delta_{i,d}^{\widehat{u}}(\mu, q) + \varepsilon \le -\rho + \varepsilon.$$

□

Corollary D.7 (Estimated margin inherited from true robustness)**.** *Suppose estimates $\widehat{\Delta}_{i,d}$ on the active incentive-deviation set satisfy*

$$|\widehat{\Delta}_{i,d} - \Delta_{i,d}^{u}(\mu, q)| \le \varepsilon_{\mathrm{cov}} \qquad \textit{for every } i,\ d \in \mathcal{D}_i^{+}(\mu).$$

If (μ, q) is ρ-robustly feasible for the true utilities over these active constraints, then

$$\widehat{\Delta}_{i,d} \le -\rho + \varepsilon_{\mathrm{cov}} \qquad \textit{for every } i,\ d \in \mathcal{D}_i^{+}(\mu).$$

In particular, if $\varepsilon_{\mathrm{cov}} \le \rho/2$, then the estimated gains certify margin at least $\rho/2$.

Proof. For every i and $d \in \mathcal{D}_i^{+}(\mu)$,

$$\widehat{\Delta}_{i,d} \le \Delta_{i,d}^{u}(\mu, q) + \varepsilon_{\mathrm{cov}} \le -\rho + \varepsilon_{\mathrm{cov}}.$$

□

The next corollary is the gain-level instantiation of Lemma D.6 for the estimates produced by the local coverage module. It records the same slack erosion directly in the $L_{\mathrm{cov}}(\varepsilon_{\mathrm{el}} + \varepsilon_{\mathrm{sb}})$ ledger.

Corollary D.8 (Local-coverage margin condition)**.** *Suppose the local coverage module estimates every active incentive-deviation gain $d \in \mathcal{D}_i^{+}(\mu)$ with uniform error*

$$\varepsilon_{\mathrm{cov}} = L_{\mathrm{cov}}(\varepsilon_{\mathrm{el}} + \varepsilon_{\mathrm{sb}}).$$

If a nonnegative payment rule q and recommendation distribution μ satisfy the estimated incentive constraints with margin ρ_{est},

$$\widehat{\Delta}_{i,d} \le -\rho_{\mathrm{est}} \qquad \textit{for every } i,\ d \in \mathcal{D}_i^{+}(\mu),$$

then their true gains satisfy

$$\Delta_{i,d}^{u}(\mu, q) \le -\rho_{\mathrm{est}} + L_{\mathrm{cov}}(\varepsilon_{\mathrm{el}} + \varepsilon_{\mathrm{sb}}) \qquad \textit{for every } i,\ d \in \mathcal{D}_i^{+}(\mu).$$

In particular, if $L_{\mathrm{cov}}(\varepsilon_{\mathrm{el}} + \varepsilon_{\mathrm{sb}}) \le \rho_{\mathrm{est}}/2$, then (μ, q) remains payment-implementable for the true utilities with slack at least $\rho_{\mathrm{est}}/2$.

Proof. Apply the triangle inequality coordinatewise on the active set: for every i and $d \in \mathcal{D}_i^+(\mu)$,

$$\Delta_{i,d}^u(\mu, q) \le \widehat{\Delta}_{i,d} + \varepsilon_{\mathrm{cov}} \le -\rho_{\mathrm{est}} + \varepsilon_{\mathrm{cov}},$$

and substitute $\varepsilon_{\mathrm{cov}} = L_{\mathrm{cov}}(\varepsilon_{\mathrm{el}} + \varepsilon_{\mathrm{sb}})$. If $\varepsilon_{\mathrm{cov}} \le \rho_{\mathrm{est}}$ the right-hand side is nonpositive, and if $\varepsilon_{\mathrm{cov}} \le \rho_{\mathrm{est}}/2$ it is at most $-\rho_{\mathrm{est}}/2$. Nonnegativity is preserved because the statement only transfers a payment rule q that is already nonnegative. □

Remark D.9 (Two-way handoff between true and estimated margins)**.** *Corollary D.7 is the direction used by Theorem 4.5 as stated: a target that is truly ρ-robust still has estimated margin at least $\rho - \varepsilon_{\mathrm{cov}}$, so under the half-slack condition the learned gains certify margin at least $\rho/2$. Corollary D.8 is the reverse direction: a target certified on estimated gains with margin ρ_{est} has retained true slack $\rho_{\mathrm{est}} - \varepsilon_{\mathrm{cov}}$. Thus the same local-coverage ledger governs both ways of feeding a fixed target into the steering theorem.*

Stop point for the interface route. Readers checking only the theorem-facing handoff chain can stop here and move to Appendix E for local coverage and Appendix F for assembly. The next three subsections state the retained learning-block premises and record two sufficient monitored implementations of selected interfaces; none is an additional assumption of Theorem 4.5.

Gain handoffs. The reduction uses the two triangle-inequality directions recorded in Corollaries D.7 and D.8: true robustness implies an estimated margin after one uniform-error erosion, while estimated feasibility implies true feasibility after the same erosion. The fixed-target theorem uses the first direction; certified selection (Theorem 5.1) uses the second.

D.3 Learning-Block Access, Regret, and Feedback Premises

This subsection declares the two additional premises consumed only by the retained monitored elicitation witness: conditional-continuation access and blockwise fixed-plan regret with the corresponding feedback interface. They discharge the interaction assumptions used by the centered-elicitation and corruption-stability arguments; they are not premises of the abstract sandboxing interface.

Assumption D.10 (Conditional-continuation generative access for the retained witness)**.** *For every active pair (i, c), the retained monitored witness can run a conditional-continuation learning episode for $\eta = (i, c, \eta_{i,c}^{\mu,q})$. The episode samples a complete prefix from the obedient target law conditional on the private recommendation context c, including the latent node, chance history, other players' private recommendation histories, and any other hidden variables needed to define $\eta_{i,c}^{\mu,q}$. It reveals to each player only the information that player would possess in the original game, starts the actual continuation at c, and concatenates the sampled prefix with the realized continuation to form a complete terminal record.*

Each such episode counts as one learning round and is charged to the repeated interaction horizon and to the learning-round payment and value ledgers. It is not a free offline oracle call. This access premise belongs only to the retained monitored witness; Assumption 4.2 permits other implementations to meet its declared output by different means.

Assumption D.11 (Blockwise fixed-plan regret for monitored elicitation)**.** *For every retained conditional-continuation block $\eta = (i, c, \eta_{i,c}^{\mu,q})$ of length t_η, the target player's continuation plans $x_\tau \in X_{i,c}$ satisfy*

$$R_{i,c}^A(t_\eta) \ge \sup_{x \in X_{i,c}} \sum_{\tau=1}^{t_\eta} \langle c_{i,c,\tau} + p_\tau, x - x_\tau \rangle .$$

Here $c_{i,c,\tau}$ is the conditional-expected base-utility coefficient in episode τ of the context-conditioned block, in the realized-policy conditioning of Lemma C.7; the displayed inequality is assumed for that coefficient sequence, not for coefficients averaged over the other players' policy draws. The retained full-feedback route also directly observes the conditional-expected linear losses $\ell_\tau(p) = \langle p, x_\tau - x_{i,c}^\circ \rangle$ and declares principal regret $R_{i,c}^P(t_\eta)$ for this loss sequence. For this retained witness, the declared full-feedback observations augment the base terminal record and belong to the sigma-field used to evaluate learning-phase transfers and losses. An implementation restricted to terminal-record feedback must instead use the estimator route; if either side learns from sampled terminal-history estimators, the corresponding martingale or estimation residual must be included in its declared regret ledger before applying the centered-elicitation theorem.

A learner-specific causal-regret decomposition may be used to instantiate $R^A_{i,c}$, but this paper does not assume that the ordinary contextual causal regret in Definition 2.3 automatically implies the displayed fixed-plan bound. This additional premise is used only by the retained monitored elicitation witness.

D.4 A Monitored Sandboxing Witness

We now convert monitored non-target obedience into a bound on coefficient-corrupted rounds and then use Lemma C.7 to obtain the output-error ledger required by Assumption 4.2.

Proposition D.12 (Dominant-obedience wrapper is a sufficient sandboxing route)**.** *Assume agent terminal utilities are uniformly bounded, $|u_j(z)| \le U_{\max}$, and that during a sandboxed elicitation block of length t_η the principal has full-feedback monitoring of whether each non-target player $j \neq i$ follows the recommendation policy that defines the frozen environment η. The retained wrapper runs the chance law, principal recommendation/randomization rule, and other exogenous continuation components as specified by η. For each non-target player, every transfer other than the sandbox compliance bonus is suppressed during the block; hence a unilateral deviation cannot gain through another payment component. For the target player, the only varying realization-plan-dependent transfer is the centered perturbation p_t, as required by Lemma C.7. Its monitored event certifies each non-target player's prescribed continuation policy, not merely a sampled-path action. If the principal pays each non-target player a compliance bonus $M_{\mathrm{sb}} > 2U_{\max}$ for following that policy in the round, and zero otherwise, then compliance is a stagewise dominant response with margin at least $M_{\mathrm{sb}} - 2U_{\max}$. Assume that each non-target player's sandbox-phase regret guarantee is evaluated against this prescribed-policy obedience comparator, against a class containing it, or otherwise upper-bounds its cumulative gain.*

Let $\xi^\eta_{\mathrm{sb}}(t_\eta, \delta_\eta) \ge 0$ be the explicit residual declared for monitoring, confidence, or lower-order terms not already included in the regret bounds. On the corresponding success event, the fraction of rounds in which at least one non-target player is noncompliant is at most

$$\alpha_\eta(t_\eta) := \sum_{j \neq i} \frac{\mathrm{Reg}^{\mathrm{sb}}_{j,\eta}(t_\eta)}{(M_{\mathrm{sb}} - 2U_{\max})t_\eta} + \xi^\eta_{\mathrm{sb}}(t_\eta, \delta_\eta).$$

Assume also that the wrapper preserves the requested environment's conditional law, in the notation of Lemma C.7: $\mathcal{L}(W_t \mid \mathcal{G}_t) = \mathcal{L}_\eta(W)$ almost surely, where the right-hand side is η's full joint law and already includes its conditioning on the trigger context. The monitored event audits the same realized policy maps that $\mathcal{G}_t$ fixes. On a compliant round those maps equal the prescribed ones and the remaining exogenous law is η's, so the same coefficient construction gives $c_{i,t} = c_i(\eta)$. Consequently,

$$\{t : c_{i,t} \neq c_i(\eta)\} \subseteq \bigcup_{j \neq i} \{t : j \text{ is noncompliant in round } t\}.$$

Thus Lemma C.7 may take $\mathcal{B}_\eta$ to be this union, whose fraction is at most $\alpha_\eta(t_\eta)$ on the displayed success event. For the retained deterministic full-feedback wrapper, all lower-order terms may be included in $\mathrm{Reg}^{\mathrm{sb}}_{j,\eta}$, and one may take $\xi^\eta_{\mathrm{sb}} = 0$ on the success event. For the retained centered-elicitation implementation, Lemma C.7 permits the certificate-visible choice

$$\kappa_\eta := \frac{4U_{\max}}{\gamma_{i,c}},$$

where $\eta = (i, c, \eta^{\mu,q}_{i,c})$ and $\gamma_{i,c}$ is the coercivity constant of the context-conditioned continuation system. Then this wrapper instantiates Assumption 4.2 with

$$\varepsilon_{\mathrm{sb},\eta} \le \kappa_\eta \alpha_\eta(t_\eta), \qquad B_{\mathrm{sb}} \le (n-1)M_{\mathrm{sb}}.$$

Proof. For any non-target player, changing the terminal history can change terminal utility by at most $2U_{\max}$, while the retained route suppresses all other deviation-dependent transfers. A noncompliant round forfeits the bonus M_{sb}, so following the prescribed policy improves utility-plus-payment by at least $M_{\mathrm{sb}} - 2U_{\max}$ relative to any one-round deviation. The phase-specific obedience comparator is included in the contextual regret class used by the sandboxing witness, or its cumulative gain is upper-bounded by the declared regret guarantee. Hence player j's noncompliance count is at most

$$\frac{\mathrm{Reg}^{\mathrm{sb}}_{j,\eta}(t_\eta)}{M_{\mathrm{sb}} - 2U_{\max}},$$

apart from the explicitly declared residual. By the displayed coefficient-corruption inclusion, a deterministic union-counting bound over non-target players gives $\alpha_\eta(t_\eta)$. Multiplying this fraction by the explicit corruption sensitivity from Lemma C.7 gives the output-level sandboxing error consumed by Assumption 4.2. At most $n-1$ compliance bonuses are paid in a round, which gives the payment ledger. □

Remark D.13 (Scope of the obedience wrapper)**.** *Proposition D.12 is only a coarse existence witness for Assumption 4.2. If the game information structure does not let the principal verify non-target compliance with the prescribed recommendation policy, or if the sandbox-phase regret guarantees do not cover the prescribed-policy comparator, this proposition does not satisfy the output-level sandboxing interface. Other elicitation implementations must supply their own mismatch-to-output stability analysis.*

D.5 A Playback Steering Witness and Payment Bookkeeping

Proposition D.14 (Fixed-target playback is a sufficient steering route)**.** *Fix a known-utility target (μ, q), a steering horizon $T_{\mathrm{steer}} \geq 1$, a concentration budget $\delta_{\mathrm{conc}} \in (0,1)$, and a joint playback-regret failure budget $\delta_{\mathrm{reg}} \in [0,1)$ such that*

$$\delta_{\mathrm{play}} := \delta_{\mathrm{conc}} + \delta_{\mathrm{reg}} < 1.$$

The playback procedure samples recommendations from μ with fresh randomization in every steering round, uses the stage game's fresh chance randomness, and executes the fixed terminal-history-dependent payment rule $q_i(r,z)$ after observing the terminal history; no additional compliance-conditioned gating is applied beyond what is encoded in q. For each player i, let $\mathcal{C}_i$ contain every playback context whose occurrence together with nonobedience by player i can make the realized terminal history, or the payment-subtracted principal value, differ from the full-round obedience counterfactual. In the retained playback route these are move-tagged information-set/recommended-action pairs. These playback contexts index regret subsequences and are distinct from the augmented first-divergence contexts used by trigger elicitation and local coverage. The margin condition below is asserted at every $c \in \mathcal{C}_i$. Within-round behavioral randomization is realized before the pathwise comparator gains are evaluated. Actual and counterfactual executions start from the same pre-round state and share the latent recommendation draw and the complete exogenous random tapes, including off-path choices. A comparator replaces only the named player's full-round policy, holding the other players' contingent policies and private tapes fixed, so their realized actions may still differ; within-round histories, recommendation deliveries, and whether recommendations are still being sent are recomputed along each execution's own path rather than copied from the actual record. The same convention with every player replaced by obedience defines the all-obedient counterfactual used in the proof. Any conversion from an expected-regret statement to the joint realized event $\mathcal{E}_{\mathrm{reg}}$ below is charged to δ_{reg}. Assume:

1. *q is nonnegative and $\sum_i q_i(r,z) \leq B_{\mathrm{steer}}$ for every recommendation-terminal pair (r,z);*
2. *the realized one-round principal value $u_0(z) - \sum_i q_i(r,z)$ has absolute value at most B_{play};*
3. *for every player i, context $c \in \mathcal{C}_i$, and steering round t in which c occurs, let $g^{\mathrm{ob}}_{i,c,t}$ be the one-round utility-plus-payment gain of the context-indexed full-round obedience comparator associated with c. Thus c selects the actual-round subsequence, while the comparator replaces player i's entire round plan by obedience. The target satisfies the signed playback-margin condition*

$$g^{\mathrm{ob}}_{i,c,t} \geq \rho_{\mathrm{e}}\, \mathbf{1}\{\text{player } i \text{ is nonobedient at } c \text{ in round } t\}$$

for some consumed margin $\rho_{\mathrm{e}} > 0$.

Define the joint playback-regret event

$$\mathcal{E}_{\mathrm{reg}} := \bigcap_i \bigcap_{c \in \mathcal{C}_i} \left\{ \sum_{t:\, c \text{ occurs}} g^{\mathrm{ob}}_{i,c,t} \leq R_i^{\mathrm{play}}(T_{\mathrm{steer}}) \right\}.$$

Assume $\Pr(\mathcal{E}_{\mathrm{reg}}) \geq 1 - \delta_{\mathrm{reg}}$. Here $R_i^{\mathrm{play}}(T_{\mathrm{steer}})$ is the bound declared for this steering block; it is a separate quantity from the whole-horizon $\mathrm{Reg}_i(T)$ of Definition 2.3, which does not imply it. A pathwise deterministic regret guarantee is the special case $\delta_{\mathrm{reg}} = 0$. Then the playback procedure

uses the steering domain $\mathcal{Q}_{\mathrm{steer}} = \{q\}$ with one-round payment bound B_{steer}, and with probability at least $1 - \delta_{\mathrm{play}}$ its realized average principal value satisfies

$$\bar{V}_{T_{\mathrm{steer}}} \geq V_0(\mu, q) - \frac{2B_{\mathrm{play}}|\Gamma_{\mathrm{rec}}| \sum_i R_i^{\mathrm{play}}(T_{\mathrm{steer}})}{\rho_{\mathrm{e}} T_{\mathrm{steer}}} - B_{\mathrm{play}} \sqrt{\frac{2\log(1/\delta_{\mathrm{conc}})}{T_{\mathrm{steer}}}}.$$

Equivalently, this playback witness instantiates the generic steering-regret term in Assumption 4.4 with

$$\mathrm{Reg}_{\mathrm{play}}(T, \rho_{\mathrm{e}}, \delta_{\mathrm{conc}}) := 2B_{\mathrm{play}}|\Gamma_{\mathrm{rec}}| \sum_i R_i^{\mathrm{play}}(T) + \rho_{\mathrm{e}} B_{\mathrm{play}} \sqrt{2T\log(1/\delta_{\mathrm{conc}})}$$

under the per-context regret convention. If the target has signed pathwise margin ρ, the same playback procedure and regret event support every consumed margin $0 < \rho_{\mathrm{e}} \leq \rho$, using the displayed aggregate at that margin; decreasing ρ_{e} weakens only the bad-round term in the value bound, while the separately displayed concentration term is unchanged. Thus the signed pathwise margin is one sufficient route for Assumption 4.4.

Proof. Counting nonobedient contexts. Work on the joint regret event $\mathcal{E}_{\mathrm{reg}}$. Fix a player i. In the playback route the contexts in $\mathcal{C}_i$ are move-tagged information-set/recommended-action pairs and are included in the explicit playback-context list, so $|\mathcal{C}_i| \leq |\Gamma_{\mathrm{rec}}|$. For $c \in \mathcal{C}_i$, let $N_{i,c}$ be the number of steering rounds in which context c occurs and player i is nonobedient at c. The signed margin is nonnegative on every round in which c occurs and at least ρ_{e} on the counted rounds. Hence

$$\rho_{\mathrm{e}} N_{i,c} \leq \sum_{t:\, c \text{ occurs}} g_{i,c,t}^{\mathrm{ob}} \leq R_i^{\mathrm{play}}(T_{\mathrm{steer}}).$$

The last inequality is the assumed joint playback-regret event $\mathcal{E}_{\mathrm{reg}}$, whose comparators are the context-indexed full-round comparators collected in $\mathcal{D}_i^{\mathrm{ob}}$; we take $R_i^{\mathrm{play}}(T_{\mathrm{steer}})$ to include the standard lower-order terms in the per-context regret guarantee. Definition 2.3 bounds the sum over all T rounds carrying that context and does not by itself imply the steering-block bound displayed here. Thus

$$N_{i,c} \leq \frac{R_i^{\mathrm{play}}(T_{\mathrm{steer}})}{\rho_{\mathrm{e}}}.$$

From context counts to bad rounds. Summing over c gives

$$N_i^{\mathrm{bad}} := \sum_{c \in \mathcal{C}_i} N_{i,c} \leq \frac{|\mathcal{C}_i| R_i^{\mathrm{play}}(T_{\mathrm{steer}})}{\rho_{\mathrm{e}}} \leq \frac{|\Gamma_{\mathrm{rec}}| R_i^{\mathrm{play}}(T_{\mathrm{steer}})}{\rho_{\mathrm{e}}}.$$

A steering round is bad if at least one player has a nonobedient context occurrence in that round, so the number of bad rounds is at most $\sum_i N_i^{\mathrm{bad}}$. Because each $\mathcal{C}_i$ exhausts every playback context whose nonobedience can change the terminal history or payment-subtracted principal value, a good round agrees with the full-round obedience counterfactual in both terminal history and principal value.

Value accounting and concentration. Let Y_t be the counterfactual principal value in steering round t under the common coupling declared in the proposition, but with every player obeying. For the natural history filtration $\mathcal{F}_{t-1}$,

$$|Y_t| \leq B_{\mathrm{play}}, \qquad \mathbb{E}[Y_t \mid \mathcal{F}_{t-1}] = V_0(\mu, q).$$

If round t is good, its actual value v_t equals Y_t; if it is bad, both values lie in $[-B_{\mathrm{play}}, B_{\mathrm{play}}]$. Pathwise,

$$v_t \geq Y_t - 2B_{\mathrm{play}} \mathbf{1}\{\text{round } t \text{ is bad}\}.$$

Conditional Hoeffding–Azuma gives an event $\mathcal{E}_{\mathrm{conc}}$ of probability at least $1 - \delta_{\mathrm{conc}}$ on which

$$\frac{1}{T_{\mathrm{steer}}} \sum_{t=1}^{T_{\mathrm{steer}}} Y_t \geq V_0(\mu, q) - B_{\mathrm{play}} \sqrt{\frac{2\log(1/\delta_{\mathrm{conc}})}{T_{\mathrm{steer}}}}.$$

On $\mathcal{E}_{\mathrm{reg}} \cap \mathcal{E}_{\mathrm{conc}}$, combining this inequality with the bad-round count proves the displayed bound. A union bound gives this intersection probability at least $1 - (\delta_{\mathrm{reg}} + \delta_{\mathrm{conc}}) = 1 - \delta_{\mathrm{play}}$. The payment-domain claim is immediate because playback uses the fixed nonnegative rule q and no additional steering payments. □

Remark D.15 (Scope of playback steering). *Proposition D.14 is a sufficient known-utility steering witness, not a replacement for the general steering module. Its margin condition is stronger and more pathwise than mere payment-implementability of (μ, q), and the paper does not claim that every robust payment-implementable target satisfies this playback condition. More sophisticated known-utility steering algorithms may discharge Assumption 4.4 without satisfying this playback margin. When the observed terminal record identifies every reached action, the recommendation–terminal pair (r, z) determines whether each reached recommendation was obeyed. Thus sufficiently large compliance bonuses encoded inside q provide one coarse base-feedback route to the pathwise margin when utilities are bounded. These bonuses are part of the target rule q and are therefore charged in $V_0(\mu, q)$. Full feedback remains an additional premise of the retained sandboxed learning witness, not of this playback argument.*

Remark D.16 (Payment magnitude bookkeeping). *The parameter B in Theorem 4.5 dominates $|V_0(\mu, q)|$ and the absolute base principal payoff plus total payment magnitude on every learning and sandboxing round. Because centered elicitation and sandboxing run in the same block, their magnitudes add. Let $B_{\mathrm{el}}^{\max}$ bound the one-round centered-elicitation payment, and let $B_{\mathrm{other}}^{\max}$ bound any other simultaneous learning transfer. A valid witness ledger is*

$$B_{\mathrm{learn}} := U_0^{\max} + B_{\mathrm{el}}^{\max} + B_{\mathrm{sb}} + B_{\mathrm{other}}^{\max}, \qquad B := \max\{|V_0(\mu, q)|, B_{\mathrm{learn}}\}.$$

For the retained route, transfers excluded by the payment-separation premise give $B_{\mathrm{other}}^{\max} = 0$. If the terminal-history estimator route of Lemma C.9 is used, then $B_{\mathrm{el}}^{\max}$ must include the queried-route maximum of $B_{\mathrm{el}}^{\eta}(p)$ and the centered-loss bound $B_{\mathrm{el}}^{\eta}(p) + \|p\|_1$ over the declared frozen environments and perturbations, including any $1/\rho_\eta$ reach-probability blow-up.

The steering-phase payment domain is recorded separately as $B_{\mathrm{steer}}(\rho_{\mathrm{cert}})$ in Assumption 4.4. The abstract reduction bound depends on the steering interface through $\mathrm{Reg}(T_{\mathrm{steer}})$, but the payment bound remains a certificate-visible domain ledger; in playback it is reflected in B_{play} and in the target-value accounting through $V_0(\mu, q)$. The algebraic reference-action baseline changes only L_{cov}; it does not change the payment obligations of either phase.

D.6 Proofs and Accounting for Certified Selection

This subsection proves Theorem 5.1 and its two corollaries and records the class-wide learning and steering accounting. The selection uses simultaneous active-gain estimates produced before steering.

Proof of Theorem 5.1, parts (i) and (ii). Fix any selected target $\widehat{s}$ and active constraint $(i, d) \in A(\widehat{s})$. On $\mathcal{E}_{\mathrm{cov}}$,

$$\Delta_{i,d}^{u}(\widehat{s}) \le \widehat{\Delta}_{i,d}(\widehat{s}) + \varepsilon_{\mathrm{cov}} \le -\rho_{\mathrm{sel}} + \varepsilon_{\mathrm{cov}},$$

which proves the retained true-margin claim. Now fix any comparator $s^\star$ satisfying the stronger true-margin condition in the theorem. Again on $\mathcal{E}_{\mathrm{cov}}$,

$$\widehat{\Delta}_{i,d}(s^\star) \le \Delta_{i,d}^{u}(s^\star) + \varepsilon_{\mathrm{cov}} \le -\rho_{\mathrm{sel}}$$

for every active constraint of $s^\star$. Thus $s^\star$ is feasible for the same certified selection program, and optimality of $\widehat{s}$ gives $V_0(\widehat{s}) \ge V_0(s^\star)$. □

The proposition below charges T_{learn} and B to the full class-wide learning procedure; a class-uniform certificate is not treated as free or as the cost of certifying only one fixed target.

Proposition D.17 (Class-wide learning and steering accounting). *Let $\mathcal{T}$ be a nonempty finite class of candidate targets $s = (\mu, q)$ with q nonnegative, and write*

$$\Delta_{i,d}^{u}(s) := \Delta_{i,d}^{u}(\mu, q), \qquad A(s) := \{(i, d) : d \in \mathcal{D}_i^{+}(\mu)\}.$$

Suppose the learning phase outputs estimates $\widehat{\Delta}_{i,d}(s)$ and a class-uniform active-gain error ledger $\varepsilon_{\mathrm{cov}} \ge 0$. Define the simultaneous event

$$\mathcal{E}_{\mathrm{cov}} := \left\{ \left| \widehat{\Delta}_{i,d}(s) - \Delta_{i,d}^{u}(s) \right| \le \varepsilon_{\mathrm{cov}} \text{ for all } s \in \mathcal{T},\ (i, d) \in A(s) \right\},$$

and suppose

$$\Pr(\mathcal{E}_{\mathrm{cov}}) \ge 1 - \delta_{\mathrm{cov}}$$

for some $\delta_{\mathrm{cov}} \in [0, 1)$. On this event, every active gain in the class is estimated within $\varepsilon_{\mathrm{cov}}$. Let $\mathcal{F}_{\mathrm{learn}}$ be the learning-phase sigma-field; the learning outputs and $\mathcal{E}_{\mathrm{cov}}$ are $\mathcal{F}_{\mathrm{learn}}$-measurable. Let T_{learn} denote the total number of learning rounds used to produce this event, and let B satisfy the class-wide counterpart of the magnitude requirement in Theorem 4.5, including $\max_{s \in \mathcal{T}} |V_0(s)|$. The

declared failure probability of $\mathcal{E}_{\mathrm{cov}}$ must include the confidence allocation or finite-class union bound over the candidate targets and their active constraints; this probability accounting does not remove the additional rounds or magnitude cost of producing the simultaneous estimates.

For a selection margin $\rho_{\mathrm{sel}} > 0$, fix an arbitrary default target $s_0 \in \mathcal{T}$. Whenever the certified feasible set is nonempty, define a selected target $\widehat{s} = (\widehat{\mu}, \widehat{q})$ by

$$\widehat{s} \in \operatorname*{argmax}_{s \in \mathcal{T}} V_0(s) \qquad \textit{subject to} \qquad \widehat{\Delta}_{i,d}(s) \le -\rho_{\mathrm{sel}} \textit{ for all } (i,d) \in A(s),$$

using a fixed deterministic tie-breaking rule. When the feasible set is empty, the algorithm declares selection failure but sets $\widehat{s} = s_0$ solely to define the random variable on the whole sample space. Thus $\widehat{s}$ is globally $\mathcal{F}_{\mathrm{learn}}$-measurable. Since the game tree, chance process, principal payoff u_0, and every candidate pair (μ, q) are known to the principal, $V_0(s)$ is a known offline objective over $\mathcal{T}$.

Suppose the remaining interfaces of Theorem 4.5 are discharged at the selected target. For each $s \in \mathcal{T}$, let $\mathcal{E}_{\mathrm{steer}}(s)$ denote the event on which its declared steering guarantee holds when that candidate is invoked with its certified margin. Fix $\delta_{\mathrm{steer}} \in [0, 1)$ with $\delta_{\mathrm{cov}} + \delta_{\mathrm{steer}} < 1$. Assume the certified selection program is nonempty throughout $\mathcal{E}_{\mathrm{cov}}$. It suffices that one of the following holds:

(i) the selected-target guarantee holds conditionally,

$$\mathbf{1}_{\mathcal{E}_{\mathrm{cov}}} \Pr(\mathcal{E}_{\mathrm{steer}}(\widehat{s})^c \mid \mathcal{F}_{\mathrm{learn}}) \le \delta_{\mathrm{steer}} \mathbf{1}_{\mathcal{E}_{\mathrm{cov}}} \qquad \textit{a.s.};$$

(ii) there is a simultaneous class-uniform event $\mathcal{E}^{\mathrm{unif}}_{\mathrm{steer}}$ satisfying

$$\mathcal{E}^{\mathrm{unif}}_{\mathrm{steer}} \subseteq \bigcap_{s \in \mathcal{T}} \mathcal{E}_{\mathrm{steer}}(s), \qquad \Pr(\mathcal{E}^{\mathrm{unif}}_{\mathrm{steer}}) \ge 1 - \delta_{\mathrm{steer}}.$$

Fresh steering randomness, together with the fixed-target module's declared guarantee uniformly over realized learning outputs, is one sufficient way to establish route (i), but neither route requires independence from the learning phase. Define

$$\mathcal{E}_{\mathrm{joint}} := \begin{cases} \mathcal{E}_{\mathrm{cov}} \cap \mathcal{E}_{\mathrm{steer}}(\widehat{s}), & \textit{under route (i)}, \\ \mathcal{E}_{\mathrm{cov}} \cap \mathcal{E}^{\mathrm{unif}}_{\mathrm{steer}}, & \textit{under route (ii)}. \end{cases}$$

Either route gives

$$\Pr(\mathcal{E}_{\mathrm{joint}}) \ge 1 - \delta_{\mathrm{cov}} - \delta_{\mathrm{steer}}.$$

Let $T_{\mathrm{steer}} \ge 1$. If

$$\rho_{\mathrm{sel}} \ge 3\varepsilon_{\mathrm{cov}},$$

then the selected target has true robustness margin

$$\rho_{\mathrm{true}} := \rho_{\mathrm{sel}} - \varepsilon_{\mathrm{cov}},$$

and the certified steering margin may be taken as

$$\rho_{\mathrm{cert}} = \rho_{\mathrm{true}} - \varepsilon_{\mathrm{cov}} = \rho_{\mathrm{sel}} - 2\varepsilon_{\mathrm{cov}}.$$

Consequently, on $\mathcal{E}_{\mathrm{joint}}$, the resulting learning-then-steering strategy satisfies the following bound, with $T = T_{\mathrm{learn}} + T_{\mathrm{steer}} > T_{\mathrm{learn}}$. Here the steering module, its payment domain, and $\mathrm{Reg}(T_{\mathrm{steer}})$ are those certified for the selected target $\widehat{s}$, whereas T_{learn} and B account for the full class-wide learning phase that produced $\mathcal{E}_{\mathrm{cov}}$:

$$\bar{V}_T \ge V_0(\widehat{\mu}, \widehat{q}) - O\left(\frac{\mathrm{Reg}(T_{\mathrm{steer}})}{\rho_{\mathrm{cert}} T_{\mathrm{steer}}}\right) - \frac{2BT_{\mathrm{learn}}}{T}.$$

Proof. Safety and oracle comparison. Fix any selected target $\widehat{s} = (\widehat{\mu}, \widehat{q})$ and any active constraint $(i,d) \in A(\widehat{s})$. On $\mathcal{E}_{\mathrm{cov}}$, feasibility and the uniform error bound give

$$\Delta^u_{i,d}(\widehat{s}) \le \widehat{\Delta}_{i,d}(\widehat{s}) + \varepsilon_{\mathrm{cov}} \le -\rho_{\mathrm{sel}} + \varepsilon_{\mathrm{cov}}.$$

Thus the selected target is $(\rho_{\mathrm{sel}} - \varepsilon_{\mathrm{cov}})$-robust for the true utilities, and is payment-implementable whenever $\rho_{\mathrm{sel}} > \varepsilon_{\mathrm{cov}}$.

Certified steering margin. If $\rho_{\mathrm{sel}} \ge 3\varepsilon_{\mathrm{cov}}$, then

$$\varepsilon_{\mathrm{cov}} \le \frac{\rho_{\mathrm{sel}} - \varepsilon_{\mathrm{cov}}}{2} = \frac{\rho_{\mathrm{true}}}{2}.$$

So the displayed half-slack condition of Theorem 4.5 holds for the selected target with true robustness margin ρ_{true}. The estimated-gain handoff then certifies steering margin

$$\rho_{\mathrm{cert}} = \rho_{\mathrm{true}} - \varepsilon_{\mathrm{cov}} = \rho_{\mathrm{sel}} - 2\varepsilon_{\mathrm{cov}}.$$

Substituting this target into Theorem 4.5 gives the displayed value bound.

Conditional selected-target route. By construction, $\mathcal{E}_{\mathrm{cov}} \in \mathcal{F}_{\mathrm{learn}}$, and the default-target extension together with fixed tie-breaking makes $\widehat{s}$ globally $\mathcal{F}_{\mathrm{learn}}$-measurable. Under route (i), the conditional steering guarantee and the tower property give

$$\Pr(\mathcal{E}_{\mathrm{cov}} \cap \mathcal{E}_{\mathrm{steer}}(\widehat{s})^c) = \mathbb{E}[\mathbf{1}_{\mathcal{E}_{\mathrm{cov}}} \Pr(\mathcal{E}_{\mathrm{steer}}(\widehat{s})^c \mid \mathcal{F}_{\mathrm{learn}})] \le \delta_{\mathrm{steer}}.$$

Together with $\Pr(\mathcal{E}_{\mathrm{cov}}^c) \le \delta_{\mathrm{cov}}$, this proves the displayed bound for route (i).

Simultaneous class-uniform route. Under route (ii), the simultaneous inclusion implies that $\mathcal{E}_{\mathrm{steer}}^{\mathrm{unif}}$ supplies the steering guarantee for the realized $\widehat{s}$, and the ordinary union bound gives

$$\Pr\big(\mathcal{E}_{\mathrm{cov}} \cap \mathcal{E}_{\mathrm{steer}}^{\mathrm{unif}}\big) \ge 1 - \delta_{\mathrm{cov}} - \delta_{\mathrm{steer}}.$$

Thus neither route uses an independence assumption. □

Part (iii) of Theorem 5.1 is the displayed conclusion of Proposition D.17. The next two proofs supply the event $\mathcal{E}_{\mathrm{cov}}$ that the theorem and the proposition presuppose.

Corollary D.18 (Class-uniform certificate from the fixed-target interfaces)**.** *For a candidate (μ, q), write $\mathcal{E}_{\mathrm{trig}}(\mu)$ for its trigger-environment list from Corollary D.4; it depends on the candidate only through μ. Let $\mathcal{E}_{\mathcal{T}} := \bigcup_{(\mu,q)\in\mathcal{T}} \mathcal{E}_{\mathrm{trig}}(\mu)$. Invoke Assumption 4.2 once on the finite list $\mathcal{E}_{\mathcal{T}}$ with tolerances $(\varepsilon_{\mathrm{el}}, \varepsilon_{\mathrm{sb}})$ and failure budget δ_{sb}, and let Assumption 4.3 hold for each $s \in \mathcal{T}$ with constant $L_{\mathrm{cov}}(s)$. Then the success event of Assumption 4.2 is a class-uniform event $\mathcal{E}_{\mathrm{cov}}$ for Theorem 5.1 with*

$$\varepsilon_{\mathrm{cov}} = \max_{s\in\mathcal{T}} L_{\mathrm{cov}}(s)\,(\varepsilon_{\mathrm{el}} + \varepsilon_{\mathrm{sb}}), \qquad \delta_{\mathrm{cov}} = \delta_{\mathrm{sb}}.$$

Under the algebraic baseline of Theorem E.12 one may take the uniform value $L_{\mathrm{cov}}(s) \le 2|\Gamma_{\mathrm{rec}}|L_{\mathrm{ref}}^{\max}$. The learning-round and payment ledgers are those the protocol declares for $\mathcal{E}_{\mathcal{T}}$; candidates sharing a recommendation rule share every entry, so $|\mathcal{E}_{\mathcal{T}}| \le \sum_{\mu\in\mathcal{M}(\mathcal{T})} \sum_i |\mathcal{C}_i^+(\mu)|$, where $\mathcal{M}(\mathcal{T})$ is the set of distinct recommendation rules in $\mathcal{T}$.

Proof of Corollary D.18. Apply Assumption 4.2 once to the finite list $\mathcal{E}_{\mathcal{T}}$. On its success event, which has probability at least $1 - \delta_{\mathrm{sb}}$, every environment in the list carries a gauge estimate with error at most $\varepsilon_{\mathrm{el}} + \varepsilon_{\mathrm{sb}}$; in particular this holds for every environment in $\mathcal{E}_{\mathrm{trig}}(\mu)$ for every $(\mu, q) \in \mathcal{T}$. For each $s \in \mathcal{T}$, Assumption 4.3, through Lemma E.5, is a deterministic map from these gauge estimates to gain estimates with error at most $L_{\mathrm{cov}}(s)(\varepsilon_{\mathrm{el}} + \varepsilon_{\mathrm{sb}}) \le \varepsilon_{\mathrm{cov}}$. Taking $\mathcal{E}_{\mathrm{cov}}$ to be the success event of Assumption 4.2 gives the class-uniform bound with $\delta_{\mathrm{cov}} = \delta_{\mathrm{sb}}$. The environment count follows because the trigger environment $\eta_{i,c}^{\mu,q}$ and the active context set $\mathcal{C}_i^+(\mu)$ depend on the target only through μ: the environment is the continuation law under obedient play from c with target payments suppressed (Remark D.5), and activity is a property of $\Pr_\mu(\mathsf{C}_{i,d})$. □

Proof of Corollary 5.3. By Lemma E.1, $\Delta_{i,d}^u(\mu, q) = \beta_{i,d}(\mu, q) + \sum_z m_{i,d}^{\mu,q}(z)u_i(z)$, where $\beta_{i,d}(\mu, q)$ is the conditional expected difference of the known payments $q_i(r, z)$ under deviation and obedience, hence linear in q, and the signed measure $m_{i,d}^{\mu,q}$ is the difference of the obedient and deviating terminal-history laws, hence determined by (μ, d) alone. By Lemma E.2 the utility-linear part equals $\langle h_{i,c}^u(\eta_{i,c}^{\mu,q}), y_{i,d}^{\mu,q}\rangle$, and both the environment and the trigger-flow vector are likewise determined by (μ, d): the environment suppresses target payments (Remark D.5) and $y_{i,d}^{\mu,q}$ is a difference of two realization plans induced by behavior, not by payments. A certificate whose labels and coefficients depend on the target only through μ therefore produces one estimate $\widehat{g}_{i,d}$ of this part, and

$$\widehat{\Delta}_{i,d}(\mu, q) - \Delta_{i,d}^u(\mu, q) = \widehat{g}_{i,d} - \langle h_{i,c}^u(\eta_{i,c}^{\mu,q}), y_{i,d}^{\mu,q}\rangle$$

does not depend on q. Lemma E.5 bounds it by $L_{\mathrm{cov}}(\varepsilon_{\mathrm{el}}+\varepsilon_{\mathrm{sb}})$ on the success event of Assumption 4.2 for the single list $\mathcal{E}_{\mathrm{trig}}(\mu)$, so the supremum over Q equals the single-target bound and no union bound over Q is used. The reference-action certificate of Theorem E.12 has the required property: its labels are (i, c, I, a) with environment $\eta_{i,c}^{\mu,q}$, and its coefficients are coordinates of $y_{i,d}^{\mu,q}$. Parts (i) and (ii) of Theorem 5.1 use only the class-uniform bound and the optimality of $\widehat{s}$, so their proofs apply verbatim to $\{\mu\} \times Q$ whenever the certified program attains its maximum, and part (iii) applies to the selected target as before. Finally, $A(\mu)$ does not depend on q; each constraint $\beta_{i,d}(\mu, q) + \widehat{g}_{i,d} \le -\rho_{\mathrm{sel}}$ is a closed half-space in q; $V_0(\mu, q) = \mathbb{E}_{\mu,\mathrm{ob}}[u_0(z)] - \sum_i \mathbb{E}_{\mu,\mathrm{ob}}[q_i(r, z)]$ is affine in q; and nonnegativity together with $\sum_i q_i(r, z) \le B_Q$ is a system of linear constraints. Hence for polyhedral Q the certified set is a polyhedron, and the selection program is a linear program over a compact polyhedron, which attains its maximum and admits a deterministic vertex rule whenever that polyhedron is nonempty. □

Remark D.19 (Scope of certified selection)**.** *Theorem 5.1 and Proposition D.17 concern a post-learning certified target-selection procedure over a known class. The plug-in margin $\rho_{\mathrm{sel}} - 2\varepsilon_{\mathrm{cov}}$ is conservative, with no sharpness claim, and the oracle comparison is margin-shifted rather than relative to every truly feasible candidate. This is a statistical safety and oracle-comparison guarantee. If a finite class is represented implicitly, the result does not separately assert polynomial-time optimization of the displayed selection program, whereas for a polyhedral payment family over a fixed recommendation rule Corollary 5.3 makes the program a linear program. A continuous family of recommendation rules would still need an additional uniform convergence, covering, or Lipschitz-discretization argument, because its environments and trigger-flow vectors vary with μ. The result is therefore not a statement about adaptive online target discovery.*

E Local-Coverage Proofs

This appendix gives the required theorem-facing route supporting Assumption 4.3: an abstract scalar-query certificate, a route-neutral error-propagation lemma, and the algebraic reference-action baseline. Its role is narrow: to give certificate-level sufficient conditions under which active trigger-deviation gains can be estimated with error $L_{\mathrm{cov}}(\varepsilon_{\mathrm{el}} + \varepsilon_{\mathrm{sb}})$. The main theorem uses only an abstract scalar-query certificate, the active-gain decomposition, and the resulting certified error ledger. The algebraic reference-action route below supplies a coarse explicit-representation-relative discharge by recovering every active trigger gain from finitely many trigger-environment scalar queries with one uniform error ledger.

After the sufficient route, a query-model-relative converse explains why some form of local coverage is necessary for the scalar quotient-query interface used here.

E.1 Sufficient Scalar-Certificate and Algebraic Route

The route converts active deviations into kernel directions, propagates the scalar-query error ledger through an abstract certificate, and instantiates that certificate with reference-action spanning. Readers who only need this coarse discharge can stop after Theorem E.12 and the assembly in Appendix F.

Theorem-facing trigger-flow bridge.

Lemma E.1 (Trigger gains are utility-linear)**.** *Fix a target recommendation-and-payment pair (μ, q), a player i, and a proper incentive deviation $d \in \mathcal{D}_i^{\mathrm{IC}}$. There are a utility-independent signed terminal-history measure $m_{i,d}^{\mu,q}$ and a utility-independent scalar $\beta_{i,d}(\mu, q)$ such that, for every utility profile u,*

$$\Delta_{i,d}^{u}(\mu, q) = \beta_{i,d}(\mu, q) + \sum_{z \in Z} m_{i,d}^{\mu,q}(z)\, u_i(z).$$

Consequently the unknown-utility part of every trigger gain is a linear functional of the sequence-form payoff quotient. The known scalar $\beta_{i,d}$ contains the payment contribution and contributes no elicitation error.

Proof. If $\Pr_\mu(\mathsf{C}_{i,d}) = 0$, set $m_{i,d}^{\mu,q} = 0$ and $\beta_{i,d}(\mu, q) = 0$; the claim is then immediate. Otherwise, conditional on the trigger context $\mathsf{C}_{i,d}$, the target (μ, q) and the deviation d induce two distributions over recommendation-terminal pairs and payment transfers: one under obedience and one after applying d. Their difference is independent of the unknown utility values. Let $m_{i,d}^{\mu,q}$ be the signed difference of the induced terminal-history marginals, and let $\beta_{i,d}(\mu, q)$ be the corresponding signed difference of the known terminal-history-dependent payments $q_i(r, z)$. Taking the conditional expected utility-plus-payment difference gives the displayed formula. Infoset-potential invariance then shows that the utility-linear part depends only on the payoff quotient. □

Lemma E.2 (Sequence-form trigger linearization)**.** *In the setting of Lemma E.1, all trigger-flow objects below are defined under the corresponding trigger-context conditioning. If $\Pr_\mu(\mathsf{C}_{i,d}) = 0$, set $y_{i,d}^{\mu,q} = 0$ and generate no local-coverage obligation for this deviation. Otherwise, write $c = \mathsf{C}_{i,d}$ and re-root player i's recommendation-augmented continuation sequence form at c:*

$$X_{i,c} := \{x \geq 0 : E_{i,c}x = e_{i,c}\}.$$

Let $\Sigma_{i,c}$ and $\mathcal{I}_{i,c}$ denote its continuation sequences and augmented information sets, whose labels retain the private recommendations player i has actually received. There exist a frozen conditional

continuation environment $\eta^{\mu,q}_{i,c}$ and two feasible realization plans $x^{\mathrm{ob}}_{i,c}(\eta^{\mu,q}_{i,c})$ and $x^{d}_{i,c}(\eta^{\mu,q}_{i,c})$ such that, with

$$y^{\mu,q}_{i,d} = x^{d}_{i,c}(\eta^{\mu,q}_{i,c}) - x^{\mathrm{ob}}_{i,c}(\eta^{\mu,q}_{i,c}) \in \ker(E_{i,c}),$$

the unknown-utility part of the trigger gain satisfies

$$\sum_{z\in Z} m^{\mu,q}_{i,d}(z)\, u_i(z) = \left\langle c^{u}_{i,c}(\eta^{\mu,q}_{i,c}), y^{\mu,q}_{i,d} \right\rangle = \left\langle h^{u}_{i,c}(\eta^{\mu,q}_{i,c}), y^{\mu,q}_{i,d} \right\rangle.$$

The vector $y^{\mu,q}_{i,d}$ is fixed by the tree, target, and deviation; it does not depend on the unknown utility values. Both realization plans satisfy the same context-conditioned sequence-form constraints $E_{i,c}x = e_{i,c}$, so the kernel membership is purely algebraic.

Proof. If $\Pr_\mu(\mathsf{C}_{i,d}) = 0$, the utility-linear part was set to zero in Lemma E.1, and the convention $y^{\mu,q}_{i,d} = 0$ makes the claim immediate. Otherwise, the target recommendation rule, chance process, opponents, and the trigger context define a frozen environment for player i relevant to the conditional comparison. Obedience and applying d give two feasible realization plans in that same environment. Thus $E_{i,c}x^{\mathrm{ob}}_{i,c} = e_{i,c}$ and $E_{i,c}x^{d}_{i,c} = e_{i,c}$, so their difference is in $\ker(E_{i,c})$. The expected unknown utility difference is the utility-only sequence-form payoff coefficient paired with this signed realization-flow difference; the known payment contribution has already been placed in $\beta_{i,d}$. Since $c^{u}_{i,c}(\eta^{\mu,q}_{i,c}) - h^{u}_{i,c}(\eta^{\mu,q}_{i,c}) \in \mathrm{Im}(E^{\top}_{i,c})$ and $y^{\mu,q}_{i,d} \in \ker(E_{i,c})$, the quotient representative gives the same pairing. □

For every active pair (i, c), apply the quotient and centered-elicitation construction of Section 3 to the re-rooted system $E_{i,c}x = e_{i,c}$: choose a complement $H_{i,c}$, reference plan $x^{\circ}_{i,c}$, gauge representative $h_{i,c}$, and coercivity constant $\gamma_{i,c} > 0$. The proofs in Appendices B–C depend only on the sequence-form equalities and therefore apply verbatim to each continuation system. In particular, the endpoint-reference construction gives a set $\mathcal{S}^{\mathrm{test}}_{i,c}$ satisfying

$$\gamma_{i,c} \geq 1/|\mathcal{S}^{\mathrm{test}}_{i,c}|.$$

Lemma E.3 (Trigger-flow base norm bound)**.** *For every target (μ, q) and causal deviation d,*

$$\|y^{\mu,q}_{i,d}\|_1 \leq 2|\Sigma_{i,c}| \leq 2|\Gamma_{\mathrm{rec}}|, \qquad c = \mathsf{C}_{i,d}.$$

Proof. If $\Pr_\mu(\mathsf{C}_{i,d}) = 0$, the convention $y^{\mu,q}_{i,d} = 0$ gives the claim. Otherwise, in a feasible realization plan, the root coordinate is one and every nonempty sequence coordinate is a nonnegative flow descendant of the root. Thus each coordinate lies in $[0, 1]$, and every feasible realization plan has ℓ_1 norm at most $|\Sigma_{i,c}|$. Lemma E.2 writes $y^{\mu,q}_{i,d}$ as the difference of two feasible realization plans in the same frozen environment, so the triangle inequality gives the displayed bound. □

Definition E.4 (Abstract local-coverage certificate)**.** *Fix a target (μ, q). An abstract local-coverage certificate consists of the following utility-independent data, together with the scalar values induced by the unknown utilities.*

1. *A finite scalar-label set Λ_{cov}. Each $\lambda \in \Lambda_{\mathrm{cov}}$ declares a frozen or sandboxed environment η_λ, a utility-dependent scalar Ψ^{u}_{λ}, an estimate $\widehat{\Psi}_\lambda$, and a local error ledger $L_\lambda \geq 0$. Several labels may use the same environment; the environment ledger is the number of distinct environments in $\{\eta_\lambda : \lambda \in \Lambda_{\mathrm{cov}}\}$, while the scalar-query ledger is $|\Lambda_{\mathrm{cov}}|$.*

2. *For every player i and active incentive deviation $d \in \mathcal{D}^{+}_{i}(\mu)$, a known scalar $\beta_{i,d}(\mu, q)$, a finite support $K_{i,d} \subseteq \Lambda_{\mathrm{cov}}$, and coefficients $\alpha_{i,d,\lambda}$, all independent of the utility profile, such that*

$$\Delta^{u}_{i,d}(\mu, q) = \beta_{i,d}(\mu, q) + \sum_{\lambda\in K_{i,d}} \alpha_{i,d,\lambda}\, \Psi^{u}_{\lambda} \qquad \textit{for every } u.$$

3. *A certificate error guarantee*

$$|\widehat{\Psi}_\lambda - \Psi^{u}_{\lambda}| \leq L_\lambda(\varepsilon_{\mathrm{el}} + \varepsilon_{\mathrm{sb}}) \qquad \textit{for every } \lambda \in \Lambda_{\mathrm{cov}}.$$

The certificate declares the weighted local-coverage ledger

$$L^{\mathrm{cert}}_{\mathrm{cov}} := \max_{\substack{i \\ d\in\mathcal{D}^{+}_{i}(\mu)}} \sum_{\lambda\in K_{i,d}} |\alpha_{i,d,\lambda}|\, L_\lambda.$$

For the coarser product ledger, write

$$W_{\mathrm{cov}} := \max_{i} \sum_{d \in \mathcal{D}_i^+(\mu)} \sum_{\lambda \in K_{i,d}} |\alpha_{i,d,\lambda}|, \qquad L_{\mathrm{loc}} := \max_{\lambda \in \Lambda_{\mathrm{cov}}} L_\lambda,$$

so that $L_{\mathrm{cov}}^{\mathrm{cert}} \le W_{\mathrm{cov}} L_{\mathrm{loc}}$. The environment and scalar-query ledgers above are part of the certificate description and support the non-vacuity accounting for a concrete route. Implementation-specific payment, loss, or importance-weight magnitudes are declared by the route that realizes the labels; they are not part of the abstract error inequality but remain part of the route's ledger.

Lemma E.5 (Abstract coverage certificate implies local coverage)**.** *Suppose an abstract local-coverage certificate is given for (μ, q), and define*

$$\widehat{\Delta}_{i,d}(\mu, q) = \beta_{i,d}(\mu, q) + \sum_{\lambda \in K_{i,d}} \alpha_{i,d,\lambda}\, \widehat{\Psi}_\lambda.$$

Then

$$|\widehat{\Delta}_{i,d}(\mu, q) - \Delta_{i,d}^u(\mu, q)| \le \left(\sum_{\lambda \in K_{i,d}} |\alpha_{i,d,\lambda}|\, L_\lambda \right) (\varepsilon_{\mathrm{el}} + \varepsilon_{\mathrm{sb}}) \le L_{\mathrm{cov}}^{\mathrm{cert}} (\varepsilon_{\mathrm{el}} + \varepsilon_{\mathrm{sb}})$$

for all i and $d \in \mathcal{D}_i^+(\mu)$. Hence the local coverage module holds with any certified bound $L_{\mathrm{cov}} \ge L_{\mathrm{cov}}^{\mathrm{cert}}$; in particular one may take $L_{\mathrm{cov}} := W_{\mathrm{cov}} L_{\mathrm{loc}}$.

Proof. Subtract the certificate decomposition from the definition of $\widehat{\Delta}_{i,d}$ and use the triangle inequality:

$$|\widehat{\Delta}_{i,d} - \Delta_{i,d}^u| \le \sum_{\lambda \in K_{i,d}} |\alpha_{i,d,\lambda}|\, |\widehat{\Psi}_\lambda - \Psi_\lambda^u| \le \sum_{\lambda \in K_{i,d}} |\alpha_{i,d,\lambda}|\, L_\lambda (\varepsilon_{\mathrm{el}} + \varepsilon_{\mathrm{sb}}).$$

Taking the maximum over active deviations gives the displayed $L_{\mathrm{cov}}^{\mathrm{cert}}$ bound, and the product bound follows from the definitions of W_{cov} and L_{loc}. □

Remark E.6 (Full-vector certificate and the role of the baseline)**.** *The simplest abstract certificate uses one label per active deviation $d \in \mathcal{D}_i^+(\mu)$, with environment $\eta_{i,c}^{\mu,q}$, scalar $\Psi_\lambda^u = \left\langle h_{i,c}^u(\eta_{i,c}^{\mu,q}), y_{i,d}^{\mu,q} \right\rangle$, coefficient 1, and local ledger $L_\lambda = \|y_{i,d}^{\mu,q}\|_*$; its error guarantee is the dual-norm inequality applied to $\widehat{h}_{i,c} - h_{i,c}^u$, exactly as in Corollary E.9. Lemma E.5 then gives $L_{\mathrm{cov}} = \max_{i,d} \|y_{i,d}^{\mu,q}\|_*$. The reference-action decomposition of Lemma E.11 yields*

$$\|y_{i,d}^{\mu,q}\|_* \le \sum_{I,\, a \ne a_i^\star(I)} |y_{i,d}^{\mu,q}(p(I)a)|\, \|b_{I,a}^\star\|_* \le 2|\Sigma_{i,c}| L_{\mathrm{ref}}^{\max},$$

so Theorem E.12 is a certificate-visible upper bound on this dual norm that holds for any gauge norm. It is the route to use when the dual norm of $y_{i,d}^{\mu,q}$ is not itself available in closed form for the chosen gauge norm.

Reference-action spanning for the required route. The following reference-action atoms are the algebraic objects behind the coarse local-coverage discharge. They provide a direct spanning family for the trigger-flow kernel directions and are the required-route ingredients for Theorem E.12. They are constructed separately inside each active continuation system (i, c). Within the next definitions and algebraic lemmas we abbreviate $(E_{i,c}, \Sigma_{i,c}, \mathcal{I}_{i,c})$ by $(E_i, \Sigma_i, \mathcal{I}_i)$; when scalar labels are assembled, the suppressed context index is restored explicitly.

Definition E.7 (Reference-action global remainder basis)**.** *Fix an active pair (i, c) and choose a reference action $a_{i,c}^\star(I) \in A(I)$ at every continuation information set $I \in \mathcal{I}_{i,c}$. Using the local abbreviations above, for a nonempty sequence $\sigma \in \Sigma_i$, define the reference-completion vector $R_i^\star(\sigma) \in \mathbb{R}^{\Sigma_i}$ by the following recursion. It puts unit mass on σ. Whenever a supported sequence τ is the parent sequence of a later player-i information set J, it puts unit mass on the reference child $\tau a_{i,c}^\star(J)$ and zero mass on all other children of J. If the same sequence is the parent of several later information-set rows, the recursion follows the reference child in each such row. Hence $|\operatorname{supp} R_i^\star(\sigma)| \le |\Sigma_i|$.*

For every information set I and every non-reference action $a \ne a_{i,c}^\star(I)$, define

$$b_{i,c,I,a}^\star := R_i^\star(p(I)a) - R_i^\star(p(I)a_{i,c}^\star(I)),$$

where $p(I)$ is the parent sequence of I. Both starting sequences are nonempty, so the construction never puts mass on the root coordinate. The reference-action global remainder family is

$$\mathcal{B}^{\mathrm{ref}}_{i,c} := \{b^\star_{i,c,I,a} : I \in \mathcal{I}_i, a \in A(I) \setminus \{a^\star_{i,c}(I)\}\}.$$

This family is certificate-visible once the reference actions are fixed, and has size at most $|\Sigma_i|$. When (i, c) is fixed, we continue to write $b^\star_{I,a}$ and $a^\star_i(I)$ for these context-specific objects, and $\mathcal{B}^{\mathrm{ref}}_i$ for $\mathcal{B}^{\mathrm{ref}}_{i,c}$, to keep the algebraic proofs readable; the certificate theorem restores the full (i, c, I, a) index.

The vector $R^\star_i(\sigma)$ is only a deterministic completion vector used to build kernel atoms. It is not asserted to be a feasible realization plan from the root; feasibility is needed only for the differences $b^\star_{I,a}$, which the next lemma verifies to lie in $\ker(E_i)$.

Lemma E.8 (Reference-action atoms span sequence-form kernel directions)**.** *After adjoining the reference-action atoms as the global-remainder atom family $\mathcal{B}^{\mathrm{rem}}_i = \mathcal{B}^{\mathrm{ref}}_i$, every vector $\kappa \in \ker(E_i)$ admits the decomposition*

$$\kappa = \sum_{I \in \mathcal{I}_i} \sum_{a \neq a^\star_i(I)} \kappa_{p(I)a}\, b^\star_{I,a},$$

with coefficient cost

$$\sum_{I, a \neq a^\star_i(I)} |\kappa_{p(I)a}| \le \|\kappa\|_1.$$

Each atom has support size at most $|\Sigma_i|$.

Proof. Kernel membership. First, $b^\star_{I,a} \in \ker(E_i)$. At row I, the two completions enter through two children of the same parent and cancel in the conservation row. Below those children, each completion follows reference actions, so every downstream row is locally conservative. No ancestor row is touched.

Spanning and coefficient cost. For spanning, use the displayed coefficients. We prove equality by induction along the sequence partial order. A non-reference child sequence $p(I)a$ is created only by the atom $b^\star_{I,a}$, with coefficient $\kappa_{p(I)a}$, so the expansion matches κ on every non-reference child. At the reference child $p(I)a^\star_i(I)$, upstream reference completions contribute the already-matched parent value $\kappa_{p(I)}$, and the atoms at I subtract the non-reference child masses. Since $\kappa \in \ker(E_i)$, the conservation row at I gives

$$\kappa_{p(I)a^\star_i(I)} = \kappa_{p(I)} - \sum_{a \neq a^\star_i(I)} \kappa_{p(I)a},$$

so the reference child also matches. This proves the expansion. The coefficient-cost bound is immediate because the coefficients are a subset of the coordinates of κ, and the support bound follows from Definition E.7. □

Corollary E.9 (Eliciting reference-action scalar queries)**.** *Fix an active pair (i, c). Suppose the trigger environment $\eta^{\mu,q}_{i,c}$ is included among the frozen environments whose gauge representatives are elicited. If the sandboxed elicitation stage returns $\widehat{h}_{i,c}(\eta^{\mu,q}_{i,c})$ with gauge-norm error at most $\varepsilon_{\mathrm{el}} + \varepsilon_{\mathrm{sb}}$, then every context-specific reference-action atom $b^\star_{I,a}$ induces a scalar estimate*

$$\widehat{\Psi}_{I,a} := \big\langle \widehat{h}_{i,c}(\eta^{\mu,q}_{i,c}), b^\star_{I,a} \big\rangle$$

for

$$\Psi^u_{I,a} := \big\langle h^u_{i,c}(\eta^{\mu,q}_{i,c}), b^\star_{I,a} \big\rangle$$

with

$$|\widehat{\Psi}_{I,a} - \Psi^u_{I,a}| \le \|b^\star_{I,a}\|_*(\varepsilon_{\mathrm{el}} + \varepsilon_{\mathrm{sb}}).$$

These reference-action scalar labels use the trigger environment and do not create recorded-excursion sandbox environments.

Proof. By Lemma E.8, $b^\star_{I,a} \in \ker(E_i)$. Hence the pairing with the payoff representative is invariant under adding an infoset potential in $\mathrm{Im}(E_i^\top)$ and is a well-defined linear functional of the elicited gauge representative. The displayed error bound is the dual-norm inequality applied to $\widehat{h}_{i,c}(\eta^{\mu,q}_{i,c}) - h^u_{i,c}(\eta^{\mu,q}_{i,c})$. □

Remark E.10 (Why no reconvergence condition is needed for this certificate)**.** *An arbitrary local atom family need not stably span all kernel remainders. The designated reference-action family instead uses the pivot coordinates $\kappa_{p(I)a}$, avoiding a generic propagation inverse. If a sequence is the parent of several later information-set rows, the reference completion follows the reference child in each such row; this can enlarge an atom's support, but by at most $|\Sigma_i|$. It does not duplicate the coefficient charged to that atom. Thus reconvergence affects the atom support/query ledger, a purely combinatorial quantity that is polynomial in the explicit sequence-form encoding, but it does not create an additional coefficient-norm loss in Lemma E.8.*

Lemma E.11 (Reference-action spanning of active trigger-flow directions)**.** *Fix $d \in \mathcal{D}_i^+(\mu)$ and write $c = \mathsf{C}_{i,d}$. Within the continuation system (i, c), after adjoining the reference-action atoms as $\mathcal{B}_i^{\mathrm{rem}} = \mathcal{B}_i^{\mathrm{ref}}$, every trigger-flow kernel vector $y_{i,d}^{\mu,q}$ has a constructive atomic decomposition using only global-remainder labels:*

$$y_{i,d}^{\mu,q} = \sum_{I \in \mathcal{I}_i} \sum_{a \neq a_i^\star(I)} y_{i,d}^{\mu,q}(p(I)a)\, b_{I,a}^\star.$$

The coefficient cost is bounded by

$$\|\mathsf{Dec}_{i,d}^{\mathrm{ref}}\|_1 \le \|y_{i,d}^{\mu,q}\|_1 \le 2|\Sigma_{i,c}| \le 2|\Gamma_{\mathrm{rec}}|,$$

and the number of distinct global-remainder labels used is at most $|\Sigma_{i,c}|$. This algebraic baseline adds no recorded-excursion sandbox environment; its scalar queries are the trigger-environment gauge-linear queries of Corollary E.9.

Proof. Apply Lemma E.8 to $\kappa = y_{i,d}^{\mu,q}$, which lies in $\ker(E_i)$ by Lemma E.2. The coefficient bound is the lemma's ℓ_1 coefficient bound followed by Lemma E.3. The label count is the size of the reference-action basis. The scalar-query statement is exactly Corollary E.9. □

Theorem E.12 (Algebraic reference-action baseline supplies local coverage)**.** *Fix a target (μ, q) and choose reference actions $a_{i,c}^\star(I)$ for every continuation information set in every active pair (i, c). Adjoin the reference-action basis as the global-remainder atom family $\mathcal{B}_{i,c}^{\mathrm{rem}} = \mathcal{B}_{i,c}^{\mathrm{ref}}$ for every active pair. Assume every trigger environment $\eta_{i,c}^{\mu,q}$ is included in the elicitation list and that, in the chosen gauge norms,*

$$L_{\mathrm{ref}}^{\max} := \max_{\substack{i,\ c \in \mathcal{C}_i^+(\mu) \\ I \in \mathcal{I}_{i,c},\ a \neq a_{i,c}^\star(I)}} \|b_{i,c,I,a}^\star\|_* \le \mathrm{poly}(|\Gamma_{\mathrm{rec}}|).$$

Then the local coverage module holds using only trigger-environment global-remainder scalar queries for this algebraic baseline. It uses no recorded-excursion sandbox environment for the decomposition labels, uses at most

$$\sum_i \sum_{c \in \mathcal{C}_i^+(\mu)} |\Sigma_{i,c}| \le |\Gamma_{\mathrm{rec}}|$$

distinct reference-action scalar labels, and has

$$W_{\mathrm{cov}} \le 2|\Gamma_{\mathrm{rec}}|, \qquad L_{\mathrm{loc}} \le L_{\mathrm{ref}}^{\max}, \qquad L_{\mathrm{cov}} \le 2|\Gamma_{\mathrm{rec}}|\, L_{\mathrm{ref}}^{\max}.$$

Proof. Scalar labels and their errors. We directly build the abstract certificate of Definition E.4. For each player i, active context $c \in \mathcal{C}_i^+(\mu)$, and reference-action atom $b_{i,c,I,a}^\star \in \mathcal{B}_{i,c}^{\mathrm{ref}}$, create one scalar label $\lambda = (i, c, I, a)$, set its environment to $\eta_{i,c}^{\mu,q}$, and set

$$\Psi_\lambda^u := \big\langle h_{i,c}^u(\eta_{i,c}^{\mu,q}), b_{i,c,I,a}^\star \big\rangle.$$

Since the trigger environment is included in the elicitation list, Corollary E.9 gives an estimate $\widehat{\Psi}_\lambda$ with

$$|\widehat{\Psi}_\lambda - \Psi_\lambda^u| \le \|b_{i,c,I,a}^\star\|_*(\varepsilon_{\mathrm{el}} + \varepsilon_{\mathrm{sb}}) \le L_{\mathrm{ref}}^{\max}(\varepsilon_{\mathrm{el}} + \varepsilon_{\mathrm{sb}}).$$

Thus these labels have local ledger $L_\lambda \le L_{\mathrm{ref}}^{\max}$ and can be implemented as trigger-environment global-remainder scalar queries, without recorded-excursion sandbox environments.

Active-gain decomposition and aggregate ledger. For each active deviation $d \in \mathcal{D}_i^+(\mu)$, write $c = \mathsf{C}_{i,d}$. Then Lemma E.2 gives

$$\Delta_{i,d}^u(\mu, q) = \beta_{i,d}(\mu, q) + \big\langle h_{i,c}^u(\eta_{i,c}^{\mu,q}), y_{i,d}^{\mu,q} \big\rangle, \qquad y_{i,d}^{\mu,q} \in \ker(E_{i,c}).$$

Lemma E.11 expands $y_{i,d}^{\mu,q}$ over the reference-action atoms using at most $|\Sigma_{i,c}|$ distinct labels and coefficient cost at most

$$\|y_{i,d}^{\mu,q}\|_1 \le 2|\Sigma_{i,c}| \le 2|\Gamma_{\mathrm{rec}}|,$$

where the last inequality is Lemma E.3. This is exactly the active-gain decomposition required by the abstract certificate, with the known payment term $\beta_{i,d}(\mu, q)$ carried unchanged. Therefore the certificate has $W_{\mathrm{cov}} \le 2|\Gamma_{\mathrm{rec}}|$, $L_{\mathrm{loc}} \le L_{\mathrm{ref}}^{\max}$, and at most $\sum_i \sum_{c \in \mathcal{C}_i^+(\mu)} |\Sigma_{i,c}|$ distinct scalar labels. Applying Lemma E.5 yields $L_{\mathrm{cov}} \le 2|\Gamma_{\mathrm{rec}}| L_{\mathrm{ref}}^{\max}$. □

Remark E.13 (Checklist for the algebraic baseline)**.** *The baseline closure has three certificate-visible inputs: context-specific reference actions $a^\star_{i,c}(I)$, inclusion of the trigger environments $\eta^{\mu,q}_{i,c}$ in the elicitation list, and a dual-norm ledger $L_{\mathrm{ref}}^{\max}$ for the chosen gauge norm. It does not require recorded-excursion sandbox environments for decomposition labels.*

Remark E.14 (Size of the reference-action dual-norm ledger)**.** *The assumption $L_{\mathrm{ref}}^{\max} \le \mathrm{poly}(|\Gamma_{\mathrm{rec}}|)$ is an explicit norm ledger rather than a hidden constant. For common coordinate gauge norms it is immediate. If the gauge norm is the restriction of ambient ℓ_1 to $H_{i,c}$, ambient Hölder's inequality gives*

$$\|b^\star_{I,a}\|_* \le \|b^\star_{I,a}\|_\infty \le 1.$$

If it is the restricted ambient ℓ_2 norm, then

$$\|b^\star_{I,a}\|_* \le \|b^\star_{I,a}\|_2 \le \sqrt{|\operatorname{supp}(b^\star_{I,a})|} \le \sqrt{|\Gamma_{\mathrm{rec}}|}.$$

Stop point for the coarse non-vacuity route. Theorem E.12, together with Lemma E.5 and the assembly in Appendix F, completes the explicit algebraic local-coverage discharge. Sharper certificate constructions are not needed for this theorem-supporting route.

E.2 Proof of the Scalar-Query Converse

The formal obstruction is stated in the main text; this subsection supplies its proof and scope audit.

Proposition E.15 (Scalar-query span obstruction)**.** *Let H be the stacked gauge space for a fixed finite family of queried environments. Suppose a certificate observes only*

$$Q_m(\theta) = \langle \theta, s_m \rangle, \qquad m = 1, \dots, M,$$

for $\theta \in \Theta \subseteq H$, and write $S := \mathrm{span}\{s_1, \dots, s_M\}$. Consider an active gain

$$G_y(\theta) := \beta + \langle \theta, y \rangle,$$

where β is known and $y \in H$ is the active trigger-flow direction. If there exist $\theta_0 \in \Theta$, $v \in S^\perp$, and $\alpha > 0$ such that

$$\langle v, y \rangle > 0, \qquad \theta_0 + \alpha v \in \Theta, \qquad \theta_0 - \alpha v \in \Theta,$$

then, writing $\theta^{\mathrm{rob}} := \theta_0 - \alpha v$ and $\theta^{\mathrm{bad}} := \theta_0 + \alpha v$, the two representatives produce identical query transcripts while their active gains differ by $2\alpha \langle v, y \rangle$. In particular, if

$$G_y(\theta_0) = -\rho/2, \qquad \alpha \langle v, y \rangle \ge \rho/2,$$

then one representative has gain at most $-\rho$ and the other has gain at least zero. Hence no rule depending only on the scalar-query transcript can uniformly decide whether this active constraint has margin at least ρ over Θ.

Proof of Proposition E.15. Because $v \in S^\perp$ and each $s_m \in S$, for every m,

$$Q_m(\theta_0 + \alpha v) = \langle \theta_0 + \alpha v, s_m \rangle = \langle \theta_0, s_m \rangle = \langle \theta_0 - \alpha v, s_m \rangle = Q_m(\theta_0 - \alpha v).$$

Hence θ^{rob} and θ^{bad} are observationally indistinguishable under the given scalar queries. Their active gains satisfy

$$G_y(\theta^{\mathrm{bad}}) - G_y(\theta^{\mathrm{rob}}) = \langle \theta_0 + \alpha v, y \rangle - \langle \theta_0 - \alpha v, y \rangle = 2\alpha \langle v, y \rangle.$$

If $G_y(\theta_0) = -\rho/2$ and $\alpha \langle v, y \rangle \ge \rho/2$, then

$$G_y(\theta^{\mathrm{rob}}) = -\rho/2 - \alpha \langle v, y \rangle \le -\rho, \qquad G_y(\theta^{\mathrm{bad}}) = -\rho/2 + \alpha \langle v, y \rangle \ge 0.$$

Any rule depending only on the shared transcript must output the same answer on the two representatives, so it cannot uniformly decide whether the active constraint has margin at least ρ over Θ. □

Remark E.16 (Scope of the converse obstruction)**.** *Proposition E.15 is relative to the scalar quotient-query model used by the paper. It does not rule out stronger observation models, adaptive query selection, or richer certificates that query additional directions. It establishes a conditional obstruction rather than unconditional span necessity, and the condition has two layers. Nonempty interior of Θ in the stacked gauge space makes $y^{\mu,q}_{i,d} \in S$ necessary for recovering the gain exactly, but not for*

deciding robustness at a fixed threshold: a query family can decide robustness without recovering the gain, for instance when every admissible representative has gain at most $-\rho$. Deciding robustness is obstructed only by admissible representatives with identical query transcripts and different robustness labels, which is what the threshold-crossing hypotheses of Proposition E.15 construct. Two further limits: the representatives of several environments are induced by one terminal-utility vector, so Θ may lie in a proper subspace; and the proposition establishes no quantitative optimality of the reference-action decomposition, coefficient budget, or L_{cov} ledger used by the sufficient route.

F Proof of the Fixed-Target Steering Theorem

Minimal route through Appendix F. Readers who only want the theorem-assembly chain for the main text can read the proof of Theorem 4.5, the proof of Corollary F.1 for the local-coverage instantiation, and the proof of Corollary F.5 for the monitored witness. The rest records one concrete monitored instantiation of the modular interfaces and its witness-specific learning-horizon corollary.

F.1 Modular Theorem Assembly

Proof of Theorem 4.5. Allocate the declared failure budget across sandboxed elicitation, local coverage, and known-utility steering, and let $\mathcal{E}$ be the intersection of their success events. By the theorem premise and a union bound,

$$\Pr(\mathcal{E}) \geq 1 - \delta.$$

We prove the value guarantee on $\mathcal{E}$.

Uniform gain estimation. Assumption 4.2 supplies, simultaneously over the finite environment list required by the certificate, gauge estimates with error at most $\varepsilon_{\mathrm{el}} + \varepsilon_{\mathrm{sb}}$. Applying Assumption 4.3 gives estimates of every active gain satisfying

$$|\widehat{\Delta}_{i,d} - \Delta^{u}_{i,d}(\mu^\star, q^\star)| \leq \varepsilon_{\mathrm{cov}} := L_{\mathrm{cov}}(\varepsilon_{\mathrm{el}} + \varepsilon_{\mathrm{sb}})$$

for every i and $d \in \mathcal{D}_i^+(\mu^\star)$.

True-to-estimated margin handoff. The target is truly ρ-robust, so Corollary D.7 yields

$$\widehat{\Delta}_{i,d} \leq -\rho + \varepsilon_{\mathrm{cov}} = -\rho_{\mathrm{cert}}.$$

The half-slack hypothesis implies $\rho_{\mathrm{cert}} \geq \rho/2 > 0$. Hence the learned gains meet the premise of Assumption 4.4 at exactly the declared margin and payment domain. On its success event,

$$\bar{V}_{\mathrm{steer}} \geq V_0(\mu^\star, q^\star) - L_{\mathrm{steer}}, \qquad L_{\mathrm{steer}} = O\left(\frac{\mathrm{Reg}(T_{\mathrm{steer}})}{\rho_{\mathrm{cert}} T_{\mathrm{steer}}}\right).$$

Full-horizon accounting. Let $\mathcal{L}$ be the set of learning and sandboxing rounds, so $|\mathcal{L}| = T_{\mathrm{learn}}$, and write $v_t = u_0(z^t) - \sum_i q_i^t$ for the realized principal value on such a round. By the definition of B, $v_t \geq -B$. Splitting the total average into its steering and learning parts gives

$$\bar{V}_T = \frac{T_{\mathrm{steer}}}{T}\bar{V}_{\mathrm{steer}} + \frac{1}{T}\sum_{t\in\mathcal{L}} v_t.$$

Therefore

$$\begin{aligned}
\bar{V}_T &\geq \frac{T_{\mathrm{steer}}}{T}\left(V_0(\mu^\star, q^\star) - L_{\mathrm{steer}}\right) - \frac{BT_{\mathrm{learn}}}{T} \\
&\geq V_0(\mu^\star, q^\star) - L_{\mathrm{steer}} - \frac{(|V_0(\mu^\star, q^\star)| + B)T_{\mathrm{learn}}}{T} \\
&\geq V_0(\mu^\star, q^\star) - L_{\mathrm{steer}} - \frac{2BT_{\mathrm{learn}}}{T},
\end{aligned}$$

where the second inequality uses $T_{\mathrm{steer}}/T = 1 - T_{\mathrm{learn}}/T$ together with $V_0(\mu^\star, q^\star) - L_{\mathrm{steer}} \leq |V_0(\mu^\star, q^\star)|$, and the last inequality uses the theorem's requirement $B \geq |V_0(\mu^\star, q^\star)|$. Steering payments require no additional full-horizon term: the steering interface measures value relative to $V_0(\mu^\star, q^\star)$, in which the target payments are already subtracted. Combining this deterministic implication on $\mathcal{E}$ with $\Pr(\mathcal{E}) \geq 1 - \delta$ proves the theorem. □

F.2 Algebraic Baseline Instantiation

Corollary F.1 (Fixed-target steering with the algebraic baseline)**.** *Assume all hypotheses of Theorem 4.5 with total failure probability at most δ, except that the local-coverage module is instantiated by the algebraic reference-action baseline of Theorem E.12. Then*

$$L_{\mathrm{cov}} \le 2|\Gamma_{\mathrm{rec}}|L_{\mathrm{ref}}^{\max} \le \mathrm{poly}(|\Gamma_{\mathrm{rec}}|).$$

If

$$2|\Gamma_{\mathrm{rec}}|L_{\mathrm{ref}}^{\max}(\varepsilon_{\mathrm{el}} + \varepsilon_{\mathrm{sb}}) \le \rho/2,$$

then the same fixed-target steering guarantee holds.

Proof of Corollary F.1. Theorem E.12 supplies Assumption 4.3 with the displayed L_{cov}. Substitute this module into Theorem 4.5. □

F.3 Monitored Full-Feedback Witness and Bookkeeping

Verification route for the monitored construction. This subsection assembles the retained monitored witness, combining the sandboxing, local-coverage, playback, probability, and magnitude ledgers used by Corollary F.5. To check that assembly, begin with the statement of Corollary F.5. For learning, read Corollary D.4 and Assumptions D.10 and D.11, then Lemma C.7 and Proposition D.12 for the error and payment ledgers; Lemma C.5 converts the centered gap into gauge error, and Lemma C.6, using Lemmas B.4 and B.5, supplies the explicit coercivity lower bound. The coverage step is Theorem E.12 with Lemma E.5. Independently, read Definition F.2 and Lemma F.4 for the expected and pathwise compliance-bonus margins, then Proposition D.14 for playback under its own steering-block regret premise. Return to the proof of Corollary F.5 for the joint failure and magnitude accounting through Theorem 4.5. Corollary F.6 adds the learning-horizon bound and is not needed for this assembly.

Definition F.2 (Monitored obedience events and deterministic forfeit)**.** *Given a recommendation realization r and observed terminal history z, let $\mathsf{E}_i(r,z) \in \{0,1\}$ indicate that player i obeyed every reached private recommendation context at which that player's action could affect the realized path. This reached-path event is measurable from the base recommendation–terminal record and does not itself require full feedback. The retained witness covers all such reached decisions, not a selected subset, and $\mathsf{E}_i(r,z) = 1$ if and only if all of them were obeyed. The event is read off the recommendations delivered in the execution at hand, and a first nonrecommended action sets it to zero for the rest of the round: stopping later recommendations does not erase the violation. The event family has deterministic forfeit for the active incentive deviations of (μ, q) if, for every i and $d \in \mathcal{D}_i^+(\mu)$, where d is indexed by its first-divergence context and chooses a nonrecommended action there,*

$$\Pr_{\mu,\mathrm{ob}}\big(\mathsf{E}_i(r,z) = 1 \mid \mathsf{C}_{i,d}\big) = 1, \qquad \Pr_{\mu,d}\big(\mathsf{E}_i(r,z) = 1 \mid \mathsf{C}_{i,d}\big) = 0.$$

Remark F.3 (Expected forfeit)**.** *Lemma F.4 also gives the auditable expected-forfeit variant with pass-probability gap $1 - 2\varepsilon_f > 0$; the monitored corollary uses the deterministic case $\varepsilon_f = 0$.*

Lemma F.4 (Compliance bonus: expected and signed pathwise margins)**.** *Assume $|u_i(z)| \le U_{\max}$ for every player i and terminal history z. Define the terminal-history-dependent payment rule*

$$q_i^M(r,z) = M\,\mathbf{1}\{\mathsf{E}_i(r,z) = 1\}$$

which is nonnegative and has one-round total payment at most nM.

First, suppose that for every active first-divergence trigger deviation d,

$$\Pr_{\mu,\mathrm{ob}}(\mathsf{E}_i = 1 \mid \mathsf{C}_{i,d}) \ge 1 - \varepsilon_f, \qquad \Pr_{\mu,d}(\mathsf{E}_i = 1 \mid \mathsf{C}_{i,d}) \le \varepsilon_f.$$

Then

$$\Delta_{i,d}^u(\mu, q^M) \le 2U_{\max} - (1 - 2\varepsilon_f)M.$$

In particular, if $0 \le \varepsilon_f < 1/2$ and

$$M \ge \frac{2U_{\max} + \rho}{1 - 2\varepsilon_f},$$

then (μ, q^M) has expected active-deviation margin at least ρ.

Second, suppose $\mathsf{E}_i(r,z) = 1$ exactly when player i obeys every reached decision covered by Definition F.2, and suppose $M \ge 2U_{\max}$. For every playback context c, let $g_{i,c,t}^{\mathrm{ob}}$ denote the one-round gain of the context-indexed full-round obedience comparator in a round in which c occurs. Then

$$g_{i,c,t}^{\mathrm{ob}} \ge (M - 2U_{\max})\mathbf{1}\{\text{player } i \text{ is nonobedient at } c \text{ in round } t\}.$$

Hence $M \ge 2U_{\max} + \rho$ supplies the signed playback-margin condition of Proposition D.14.

Proof. The nonnegativity and payment bound are immediate. For the expected claim, fix $d \in \mathcal{D}_i^+(\mu)$ and condition on $\mathsf{C}_{i,d}$. A deviation can increase terminal utility by at most $2U_{\max}$, while the expected bonus drops by at least $(1-2\varepsilon_f)M$. Hence

$$\Delta_{i,d}^u(\mu, q^M) \le 2U_{\max} - (1-2\varepsilon_f)M,$$

which gives the stated margin under the displayed choice of M.

For the pathwise claim, if player i is fully obedient, the comparator and actual play coincide and $g_{i,c,t}^{\mathrm{ob}} = 0$. If player i is nonobedient at any reached monitored context, its actual bonus is zero, whereas the context-indexed full-round obedience comparator receives M: on its own path that comparator never takes a nonrecommended action, so it keeps receiving recommendations and its obedience event holds. The comparator can lose at most $2U_{\max}$ in base utility, so its gain is at least $M - 2U_{\max}$. Under $M \ge 2U_{\max}$ this lower bound is nonnegative, so $g_{i,c,t}^{\mathrm{ob}} \ge 0$ in every round, and in a round in which player i is nonobedient at the indexed context c the same bound gives $g_{i,c,t}^{\mathrm{ob}} \ge M - 2U_{\max}$. The two cases are the signed inequality. □

Corollary F.5 (Monitored witness: complete ledger)**.** *Assume $|u_i(z)| \le U_{\max}$ and that the principal's learning-phase full-feedback observations can evaluate the relevant payment rules and verify the prescribed-policy obedience events used by the sandboxing witness. For this retained witness, the declared full-feedback observations are also part of the learning-phase transfer- and loss-evaluation sigma-field. The playback context list covers every move-tagged information-set/recommended-action context whose nonobedience can change the realized terminal history or payment-subtracted principal value. Assume also that the sandbox-phase regret guarantees cover the prescribed-policy obedience comparators, and that the retained learning blocks satisfy Assumptions D.10 and D.11. The retained sandbox blocks use the coefficient-conditioning and conditional-law convention of Proposition D.12. Assume further that the actual steering block, starting from the agents' post-learning states, satisfies the joint playback-regret event of Proposition D.14 with failure probability at most δ_{reg}; this blockwise premise is a separate assumption on the agents and is not inferred from Definition 2.3. Thus each context-conditioned episode is counted as one learning round rather than treated as a free oracle call. All non-target transfers other than the sandbox bonus are suppressed, and the only varying target-player realization-plan-dependent transfer is the centered perturbation. For the retained centered-elicitation implementation, Lemma C.7 supplies*

$$\kappa_\eta := \frac{4U_{\max}}{\gamma_{i,c}}, \qquad \kappa_\eta \le 4U_{\max}|\mathcal{S}_{i,c}^{\mathrm{test}}|$$

for $\eta = (i, c, \eta_{i,c}^{\mu,q})$, under the endpoint-reference norm. If the dominant-obedience sandbox uses a bonus $M_{\mathrm{sb}} > 2U_{\max}$ and the algebraic reference-action baseline is used for local coverage, then

$$K_{\mathrm{trig}} = \sum_i |\mathcal{C}_i^+(\mu)|, \qquad \varepsilon_{\mathrm{sb}} = \max_\eta \kappa_\eta \alpha_\eta, \qquad B_{\mathrm{sb}} \le (n-1)M_{\mathrm{sb}}, \qquad L_{\mathrm{cov}} \le 2|\Gamma_{\mathrm{rec}}|L_{\mathrm{ref}}^{\max}.$$

The environment count K_{trig} is finite and certificate-visible. The count K_{trig} and the purely combinatorial size factors are polynomial in the base-game size whenever the explicitly represented recommendation-augmented continuation description has polynomial size. The numerical payment, error, and value ledgers are polynomial only under the separately stated bounds on $L_{\mathrm{ref}}^{\max}$, the payoff and payment-magnitude parameters, the inverse coercivity constants, the regret-rate constants, and the elicitation domains. Here α_η is the monitored mismatch-fraction bound from Proposition D.12. If, in addition, the monitored obedience events satisfy deterministic forfeit, write $\mathsf{E}_i(r,z) \in \{0,1\}$ for the verified obedience event and define the compliance-bonus target rule q^M by

$$q_i^M(r,z) := M\mathbf{1}\{\mathsf{E}_i(r,z) = 1\}.$$

For $M \ge 2U_{\max} + \rho$, this rule gives both the expected active-deviation margin and the signed pathwise playback margin required by Proposition D.14, and has $B_{\mathrm{steer}} \le nM$. Because $\mathsf{E}_i(r,z)$ is determined from the recommendation and observed terminal path, this playback step uses the base feedback model; the full-feedback premise above is needed by the sandboxed learning blocks, not by the playback proposition. Writing $U_0^{\max} := \max_{z\in Z}|u_0(z)|$ and

$$B_{\mathrm{play}} := \max_{r,z}\left|u_0(z) - \sum_i q_i^M(r,z)\right|,$$

the same route has

$$B_{\mathrm{play}} \le U_0^{\max} + nM.$$

Let $B_{\mathrm{el}}^{\max}$ be the largest one-round centered payment over the declared learning blocks. Because the retained payment separation has $B_{\mathrm{other}}^{\max} = 0$, a valid simultaneous learning-round ledger is

$$B_{\mathrm{learn}} = U_0^{\max} + B_{\mathrm{el}}^{\max} + B_{\mathrm{sb}}, \qquad B = \max\{|V_0(\mu, q^M)|, B_{\mathrm{learn}}\}.$$

For any steering horizon $T_{\mathrm{steer}} \geq 1$, consumed margin $0 < \rho_{\mathrm{e}} \leq \rho$, concentration budget $\delta_{\mathrm{conc}} \in (0,1)$, and joint playback-regret failure budget $\delta_{\mathrm{reg}} \in [0,1)$ with $\delta_{\mathrm{play}} := \delta_{\mathrm{conc}} + \delta_{\mathrm{reg}} < 1$, it instantiates the generic steering quantity with

$$\mathrm{Reg}_{\mathrm{play}}(T_{\mathrm{steer}}, \rho_{\mathrm{e}}, \delta_{\mathrm{conc}}) = 2B_{\mathrm{play}}|\Gamma_{\mathrm{rec}}| \sum_i R_i^{\mathrm{play}}(T_{\mathrm{steer}}) + \rho_{\mathrm{e}} B_{\mathrm{play}} \sqrt{2T_{\mathrm{steer}} \log(1/\delta_{\mathrm{conc}})}$$

under the per-context regret convention for this retained playback witness. The pathwise deterministic-regret case has $\delta_{\mathrm{reg}} = 0$. At the main-theorem handoff, $\rho_{\mathrm{e}} = \rho_{\mathrm{cert}}$. If the context-conditioned learning blocks have individual failure budgets δ_η satisfying $\sum_{\eta \in \mathcal{E}_{\mathrm{trig}}} \delta_\eta \leq \delta_{\mathrm{learn}}$, and if all remaining witness events have total failure budget at most δ_{other}, choose the component budgets so that

$$\delta_{\mathrm{mon}} := \delta_{\mathrm{learn}} + \delta_{\mathrm{conc}} + \delta_{\mathrm{reg}} + \delta_{\mathrm{other}} < 1.$$

Then the joint witness succeeds with probability at least $1 - \delta_{\mathrm{mon}}$. Equivalently, $\delta_{\mathrm{mon}} = \delta_{\mathrm{learn}} + \delta_{\mathrm{play}} + \delta_{\mathrm{other}}$. Thus the retained monitored route jointly discharges the remaining interfaces of Theorem 4.5 with polynomial combinatorial factors and numerical ledgers whenever their respective explicit-description, magnitude, stability, regret-rate, and elicitation-domain bounds are polynomially bounded, conditional on its declared conditional-continuation generative access and blockwise fixed-plan-regret premises. Under the half-slack condition $2|\Gamma_{\mathrm{rec}}|L_{\mathrm{ref}}^{\max}(\varepsilon_{\mathrm{el}} + \varepsilon_{\mathrm{sb}}) \leq \rho/2$, the resulting end-to-end bound is

$$\bar{V}_T \geq V_0(\mu, q^M) - O\left(\frac{(U_0^{\max} + nM)|\Gamma_{\mathrm{rec}}| \sum_i R_i^{\mathrm{play}}(T_{\mathrm{steer}})}{\rho T_{\mathrm{steer}}} \right)$$
$$- (U_0^{\max} + nM) \sqrt{\frac{2\log(1/\delta_{\mathrm{conc}})}{T_{\mathrm{steer}}}} - \frac{2BT_{\mathrm{learn}}}{T},$$

where $U_0^{\max} = \max_z |u_0(z)|$, $B = \max\{|V_0(\mu, q^M)|, U_0^{\max} + B_{\mathrm{el}}^{\max} + B_{\mathrm{sb}}\}$ with $B_{\mathrm{el}}^{\max}$ the largest one-round centered payment, and $\delta_{\mathrm{play}} = \delta_{\mathrm{conc}} + \delta_{\mathrm{reg}}$ splits the playback budget into its concentration and regret parts.

Proof of Corollary F.5.

Learning and coverage interfaces. Proposition D.12 supplies the sandboxing interface with the displayed sandboxing-payment bound, and Corollary D.4 includes the context-conditioned trigger environments needed for global-remainder scalar queries. Assumption D.10 supplies the retained context-conditioned episodes, and Assumption D.11 supplies exactly the target-player fixed-plan regret consumed by centered elicitation. The latter also supplies the principal's directly monitored conditional-expected loss sequence without an additional sampling residual. No implication from the ordinary contextual causal-regret model is used for this learning-block premise. Theorem E.12 supplies the local coverage interface with $L_{\mathrm{cov}} \leq 2|\Gamma_{\mathrm{rec}}|L_{\mathrm{ref}}^{\max}$.

Playback margin and value loss. The playback proposition supplies the known-utility steering interface and identifies the aggregate regret quantity under the per-context convention. Because playback has signed pathwise margin at least ρ, it also certifies every smaller consumed margin $0 < \rho_{\mathrm{cert}} \leq \rho$ by Proposition D.14. Under Theorem 4.5's half-slack condition, the steering interface may therefore be invoked with the certified margin ρ_{cert}, namely with $\rho_{\mathrm{e}} = \rho_{\mathrm{cert}}$ in Proposition D.14. Lemma F.4 gives the displayed expected and signed pathwise compliance-bonus margins.

Failure probability. The playback success event uses a concentration budget δ_{conc} for the obedient-counterfactual concentration event from Proposition D.14 and a joint playback-regret failure budget δ_{reg}. Thus $\delta_{\mathrm{play}} := \delta_{\mathrm{conc}} + \delta_{\mathrm{reg}}$; the pathwise deterministic-regret case has $\delta_{\mathrm{reg}} = 0$. The learning-block events have total failure at most δ_{learn}, and remaining witness events have total failure at most δ_{other}. A union bound therefore gives total failure at most $\delta_{\mathrm{mon}} = \delta_{\mathrm{learn}} + \delta_{\mathrm{conc}} + \delta_{\mathrm{reg}} + \delta_{\mathrm{other}} < 1$.

Magnitude accounting and assembly. Let $B_{\mathrm{el}}^{\max}$ bound the simultaneous centered payment and let $B_{\mathrm{other}}^{\max} = 0$ under the retained payment separation. The learning-phase magnitude parameter is chosen as

$$B_{\mathrm{learn}} = U_0^{\max} + B_{\mathrm{el}}^{\max} + B_{\mathrm{sb}} + B_{\mathrm{other}}^{\max}, \qquad B = \max\{|V_0(\mu, q^M)|, B_{\mathrm{learn}}\}.$$

Thus it dominates the total of payments that occur simultaneously, rather than only each component separately. The steering bonus is accounted for separately: the principal's benchmark $V_0(\mu, q^M)$ already subtracts these payments, so the bonus is a target cost rather than an additional unrecorded steering overhead. The parameter B_{steer} records the steering payment domain, while B_{play} is

the principal-value magnitude absorbed into the aggregate steering-regret term. Substituting these modules into Theorem 4.5 gives the claimed joint bound under the declared conditional-continuation generative access, blockwise fixed-plan regret, and monitoring premises. Finally, substitute $B_{\mathrm{play}} \le U_0^{\max} + nM$, $L_{\mathrm{cov}} \le 2|\Gamma_{\mathrm{rec}}|L_{\mathrm{ref}}^{\max}$, and the displayed per-context playback aggregate at $\rho_{\mathrm{e}} = \rho_{\mathrm{cert}}$. Under the half-slack condition, $\rho_{\mathrm{cert}} \ge \rho/2$, so $1/\rho_{\mathrm{cert}} \le 2/\rho$, which gives the end-to-end bound displayed in the statement on the same joint success event, with the factor of at most two absorbed into the universal constant in the $O(\cdot)$ term. This proves Corollary F.5. □

F.4 Learning-Horizon Consequence

This subsection translates the blockwise elicitation and sandboxing ledgers into a sufficient learning horizon. The conclusion is conditional on the retained conditional-continuation access and fixed-plan-regret premises; it does not assert the same rate for every implementation of the abstract sandboxing interface.

Corollary F.6 (Sufficient learning horizon under the retained monitored route)**.** *Fix $\rho > 0$ and a target (μ, q) that is truly ρ-robust, and use the retained monitored route of Corollary F.5. Let*

$$\mathcal{E}_{\mathrm{trig}} := \{(i, c, \eta_{i,c}^{\mu,q}) : c \in \mathcal{C}_i^+(\mu)\}, \qquad K_{\mathrm{trig}} := |\mathcal{E}_{\mathrm{trig}}| = \sum_i |\mathcal{C}_i^+(\mu)|.$$

The count K_{trig} is finite and certificate-visible; it is polynomial in the base-game size whenever the explicitly represented recommendation-augmented continuation description has polynomial size. For each $\eta = (i, c, \eta_{i,c}^{\mu,q}) \in \mathcal{E}_{\mathrm{trig}}$, run one sandboxed elicitation block of length t_η using Assumption D.10. Every conditional-continuation episode counts as one learning round and carries its payment and value charges; no episode is a free oracle call. The sandboxing wrapper and the centered payment game operate in the same block; they are not two sequential copies of those rounds. Thus

$$T_{\mathrm{learn}} = \sum_{\eta \in \mathcal{E}_{\mathrm{trig}}} t_\eta.$$

Allocate failure probabilities δ_η with $\sum_\eta \delta_\eta \le \delta_{\mathrm{learn}}$.

Suppose each block admits declared bounds

$$R_{i,\eta}^{P}(t_\eta) \le R_{i,\eta}^{P,\mathrm{ub}}(t_\eta), \qquad R_{i,\eta}^{A}(t_\eta) \le R_{i,\eta}^{A,\mathrm{ub}}(t_\eta),$$

where $R_{i,\eta}^{A,\mathrm{ub}}$ directly upper-bounds the blockwise fixed-plan regret declared in Assumption D.11, and $R_{i,\eta}^{P,\mathrm{ub}}$ is regret for the directly observed conditional-expected linear losses. Suppose also that, for every non-target player $j \ne i$,

$$\mathrm{Reg}_{j,\eta}^{\mathrm{sb}}(t_\eta) \le R_{j,\eta}^{\mathrm{sb,ub}}(t_\eta).$$

By Lemma C.7, set $\kappa_\eta := 4U_{\max}/\gamma_{i,c}$. Let $\xi_{\mathrm{sb}}^{\eta}(t_\eta, \delta_\eta)$ be the explicit residual from Proposition D.12. For each $\eta \in \mathcal{E}_{\mathrm{trig}}$, let $\mathcal{E}_\eta$ be a block success event with

$$\Pr(\mathcal{E}_\eta) \ge 1 - \delta_\eta,$$

on which the displayed principal-regret, target-player fixed-plan-regret, sandbox-regret, and residual bounds for that block hold jointly. No independence across the events $\mathcal{E}_\eta$ is assumed. Define

$$\varepsilon_{\mathrm{el}} := \max_{\eta=(i,c,\cdot)} \frac{R_{i,\eta}^{P,\mathrm{ub}}(t_\eta) + R_{i,\eta}^{A,\mathrm{ub}}(t_\eta)}{\gamma_{i,c} t_\eta}$$

and

$$\varepsilon_{\mathrm{sb}} := \max_{\eta=(i,c,\cdot)} \kappa_\eta \left[\sum_{j \ne i} \frac{R_{j,\eta}^{\mathrm{sb,ub}}(t_\eta)}{(M_{\mathrm{sb}} - 2U_{\max})t_\eta} + \xi_{\mathrm{sb}}^{\eta}(t_\eta, \delta_\eta) \right].$$

If

$$2|\Gamma_{\mathrm{rec}}|L_{\mathrm{ref}}^{\max}(\varepsilon_{\mathrm{el}} + \varepsilon_{\mathrm{sb}}) \le \rho/2,$$

then, conditional on the declared access and fixed-plan-regret premises and with probability at least $1 - \delta_{\mathrm{learn}}$, the certified margin satisfies $\rho_{\mathrm{cert}} \ge \rho/2$.

Proof. On $\mathcal{E}_\eta$ for block $\eta = (i, c, \eta_{i,c}^{\mu,q})$, Lemma C.7 gives an exact-environment regret contribution of at most

$$\frac{R_{i,\eta}^{P,\mathrm{ub}}(t_\eta) + R_{i,\eta}^{A,\mathrm{ub}}(t_\eta)}{\gamma_{i,c} t_\eta},$$

while Proposition D.12 and the same lemma's explicit sensitivity κ_η give the displayed sandboxing contribution. No independence is needed: the union bound gives

$$\Pr\left(\bigcap_{\eta\in\mathcal{E}_{\mathrm{trig}}}\mathcal{E}_\eta\right)\geq 1-\sum_{\eta\in\mathcal{E}_{\mathrm{trig}}}\delta_\eta\geq 1-\delta_{\mathrm{learn}}.$$

On this intersection the simultaneous error is at most $\varepsilon_{\mathrm{el}}+\varepsilon_{\mathrm{sb}}$. The algebraic baseline gives $L_{\mathrm{cov}}\leq 2|\Gamma_{\mathrm{rec}}|L_{\mathrm{ref}}^{\max}$, hence

$$\rho_{\mathrm{cert}}=\rho-L_{\mathrm{cov}}(\varepsilon_{\mathrm{el}}+\varepsilon_{\mathrm{sb}})\geq\rho/2.$$

The remaining monitored interfaces are discharged by Corollary F.5. □

Remark F.7 (If the blockwise regret rates are $\widetilde{O}(\sqrt{t})$)**.** *Suppose there are certificate-visible constants $C_{\mathrm{el}},C_{\mathrm{sb}},C_\xi\geq 0$ such that, uniformly over $\eta=(i,c,\cdot)$ and t,*

$$\frac{R_{i,\eta}^{P,\mathrm{ub}}(t)+R_{i,\eta}^{A,\mathrm{ub}}(t)}{\gamma_{i,c}}\leq\widetilde{O}(C_{\mathrm{el}}\sqrt{t}),$$

$$\kappa_\eta\sum_{j\neq i}\frac{R_{j,\eta}^{\mathrm{sb,ub}}(t)}{M_{\mathrm{sb}}-2U_{\max}}\leq\widetilde{O}(C_{\mathrm{sb}}\sqrt{t}),\qquad\kappa_\eta\xi_{\mathrm{sb}}^\eta(t,\delta_\eta)\leq\widetilde{O}(C_\xi t^{-1/2}).$$

The retained deterministic wrapper has $C_\xi=0$ when all lower-order terms are included in the regret bounds. With an equal block allocation $t_\eta=t$,

$$\varepsilon_{\mathrm{el}}+\varepsilon_{\mathrm{sb}}=\widetilde{O}\Big((C_{\mathrm{el}}+C_{\mathrm{sb}}+C_\xi)t^{-1/2}\Big).$$

Consequently, under the declared conditional-continuation generative access and blockwise fixed-plan-regret premises, it suffices to take

$$T_{\mathrm{learn}}=K_{\mathrm{trig}}t=\widetilde{O}\left(K_{\mathrm{trig}}\,\frac{\big(|\Gamma_{\mathrm{rec}}|L_{\mathrm{ref}}^{\max}(C_{\mathrm{el}}+C_{\mathrm{sb}}+C_\xi)\big)^2}{\rho^2}\right)$$

to ensure $\rho_{\mathrm{cert}}\geq\rho/2$. Holding the displayed certificate-visible constants K_{trig}, $L_{\mathrm{ref}}^{\max}$, C_{el}, C_{sb}, and C_ξ fixed as ρ varies, this is $\widetilde{O}(\rho^{-2})$ in the margin. When the recommendation-augmented description has polynomial size, K_{trig} and $|\Gamma_{\mathrm{rec}}|$ are polynomial combinatorial factors. The numerical horizon is polynomial only under the separately stated polynomial bounds on $L_{\mathrm{ref}}^{\max}$, C_{el}, C_{sb}, C_ξ, the inverse coercivity and other magnitude terms absorbed into those constants, and $1/\rho$. In particular, the explicit bound $\kappa_\eta=4U_{\max}/\gamma_{i,c}$ is independent of t and ρ whenever $U_{\max}$ and the fixed gauge/reference construction are held fixed.

G Additional Experimental Diagnostics

This appendix collects the full tables and figures behind the three diagnostics of Section 6, and one exploratory audit that is not used to support a main claim. All artifacts are generated from the current seed, horizon, and budget grids.

G.1 Component Diagnostic

Table 2 reports the two-action centered game and the monitored sandbox. The column $\bar{\varepsilon}_{\mathrm{el}}$ is the realized regret certificate $(R^P+R^A)/T$, not an externally supplied error. The theorem-facing sandbox quantity is $\varepsilon_{\mathrm{sb}}^{\mathrm{cert}}=\kappa_\eta\alpha_\eta^{\mathrm{reg}}$; the paired frozen/corrupted estimate shift also stored in the artifact is descriptive only. Because only the first context lies on the target path, an off-path deviation is behaviorally noncompliant without corrupting the coefficient, so the monitored bad set can strictly cover the corruption set, as Lemma C.7 allows.

T	$\|\widehat{h} - h\|$	$(R^P + R^A)/T$	slack
256	0.0535	0.1389	0.0855
1024	0.0377	0.0607	0.0230
4096	0.0223	0.0349	0.0126

M_{sb}	$\overline{\alpha}_{\mathcal{B}}$	$\overline{\alpha}_{\mathcal{C}}$	$\overline{\varepsilon}_{\mathrm{el}}$	$\overline{\varepsilon}^{\mathrm{cert}}_{\mathrm{sb}}$	$\overline{\|\widehat{h} - h\|}$	Lem. C.7 RHS
2.1	0.0008	0.0006	0.0374	0.0545	0.0160	0.0919
2.5	0.0006	0.0004	0.0376	0.0096	0.0180	0.0472
3.0	0.0005	0.0004	0.0399	0.0047	0.0172	0.0446

Table 2: Direct witness-component diagnostics. Top: the two-action frozen centered game, where the realized Theorem 3.2 bound holds at every prefix. Bottom: means over 10 seeds at $T = 4096$ for bonuses satisfying $M_{\mathrm{sb}} > 2U_{\max}$, with $U_{\max} = \gamma = 1$, $G_\eta = 1.5$, and $\kappa_\eta = 4$. Here $\mathcal{C} \subseteq \mathcal{B}$ are the coefficient-corruption and monitored noncompliance sets. The column $\overline{\varepsilon}_{\mathrm{el}}$ is the realized centered-game regret certificate $(R^P + R^A)/T$, not an externally supplied error. All reported certified runs satisfy the Proposition D.12 and Lemma C.7 inequalities at every prefix.

G.2 Downstream Pipeline

Table 3 summarizes steering on both toy families for both learners, at $M = 0.8$ on the signaling family and $M = 0.9$ on the entry family; the signaling configuration is the $M = 0.8$ row of Table 1 run with a different seed, while the entry bonus is not among that table's entry rows. Table 4 reports the certification stage; its two rows coincide because both families use the same gain-estimation budget, the same externally declared $\varepsilon_{\mathrm{sb}} = 0.01$, the same toy ledger $L_{\mathrm{cov}} = 8$, and the same true margin. Figure 1 shows the target gap over steering rounds and Figure 2 the two error components. The sandbox-output error is an input throughout, so these artifacts do not validate Proposition D.12 or Lemma C.7; the component diagnostic does that directly.

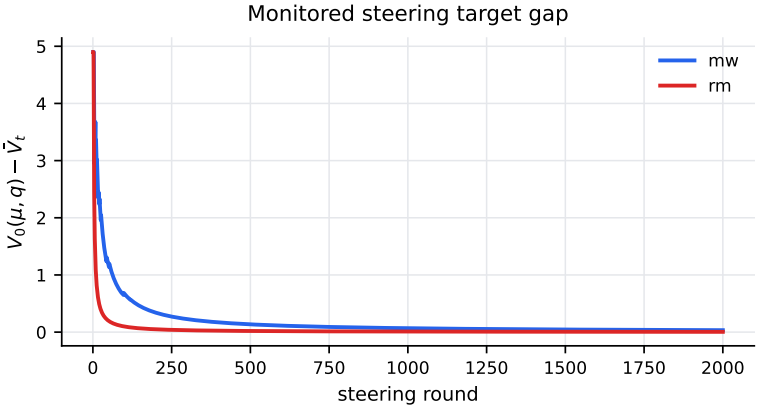


Figure 1: Monitored steering target gap on the signaling/response toy family. Smaller gaps indicate closer tracking of the payment-subtracted target benchmark after certification.

family	learner	target	realized	gap	avg pay	ρ_{cert}	bad frac
signaling	mw	2.400	2.366	0.034	1.589	0.167	0.011
signaling	rm	2.400	2.395	0.005	1.598	0.167	0.001
entry	mw	2.050	2.023	0.027	1.788	0.167	0.011
entry	rm	2.050	2.027	0.023	1.798	0.167	0.001

Table 3: Ledger-fed monitored gain-estimation-to-steering summary on two toy families. The rows report target tracking, average payment, certified margin, and bad-round fraction under the fixed-target protocol.

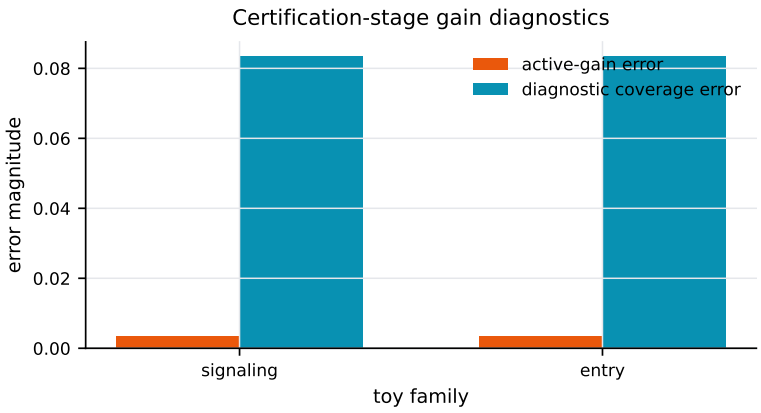


Figure 2: Certification-stage diagnostics for active trigger-deviation gains. The orange bars show $\varepsilon_{\text{gain}}$; the blue bars add the coverage-scaled sandbox-output term and give $\varepsilon_{\text{cov}}^{\text{diag}}$.

family	$\varepsilon_{\text{gain}}$	ε_{sb}	$\varepsilon_{\text{cov}}^{\text{diag}}$	true margin	ρ_{cert}
signaling	0.003	0.010	0.083	0.250	0.167
entry	0.003	0.010	0.083	0.250	0.167

Table 4: Certification-stage diagnostics for the monitored fixed-target protocol. Here $\varepsilon_{\text{gain}}$ is already an active-gain error and ε_{sb} is an externally declared pre-coverage input; the diagnostic error is $\varepsilon_{\text{gain}} + L_{\text{cov}}\varepsilon_{\text{sb}}$. The reported toy algebraic ledger uses $L_{\text{cov}} = 8$.

G.3 Payment–Value Tradeoff

Figure 3 plots the tradeoff behind Table 1.

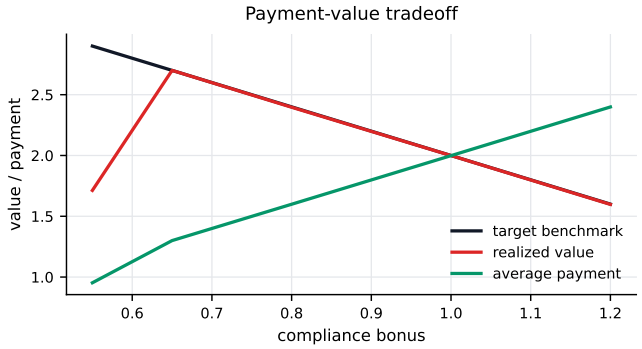


Figure 3: Payment–value tradeoff for regret matching on the signaling/response toy family. Larger bonuses improve steering reliability but reduce the payment-subtracted target benchmark.