THE ORTHOGONALITY FRONTIER: RELIABILITY-ADAPTIVE CAUSAL DECISIONS WITH MANY TREATMENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Large treatment libraries make dense prediction scalable while causal verification remains scarce. Even with a preliminary predictor and a doubly robust score in hand, a decision maker must determine how strongly causal discrepancies should revise the final normalized target. To formalize this choice, we introduce reliability-adaptive target orthogonalization (RATO), which represents the correction as an operator on the $K-1$ treatment-contrast quotient and distinguishes two criteria for choosing it: first-order orthogonality and local target risk. Full correction, $A = I$, uniquely removes first-order preliminary-predictor sensitivity within the target family. In contrast, minimum local target risk selects $A^\star = B(B+S)^{-1}$ under a zero cross-second-moment condition, where B and S capture both variance and persistent local bias. Under these local-risk conditions, whenever $A^\star$ is not a scalar multiple of I, the directional oracle has strictly lower local risk than every scalar correction. Because the required total error moments are not generically identified from routine deployment data, independent validation of the scalar family provides the default deployable route. Controlled experiments examine the frontier against exact nonlinear target error. In a retrospective participant-randomized study with 72 actions per context, increasing correction-and-tuning budget with preliminary fitting and test data fixed moves mean selected correction from 0.59 at 5% to 0.94 at 80%. A second, weak-signal study shows stronger correction without higher held-out value. The frontier allocates available causal information; it does not create treatment signal.

1 INTRODUCTION

A decision maker with many candidate treatments can often predict outcomes across the entire library, yet causally verify only a small fraction of those predictions. Shared treatment structure makes dense prediction economical; randomized exposure and outcome measurement remain costly. The practical tension is therefore not a lack of predictions, but a mismatch between their coverage and the causal evidence available for revising them. We study this offline, one-shot decision, with the treatment library and preliminary predictor fixed before correction.

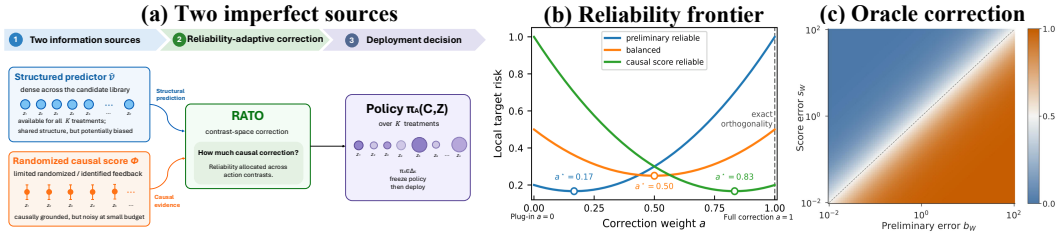

Figure 1: **Orthogonality and local target risk select different correction strengths.** The two-source calculation yields distinct scalar risk minima across preliminary-reliable, balanced, and score-reliable regimes. The local-risk minimizer increases as preliminary error grows and decreases as score error grows. Purple dots mark the analytic minima; this panel is a preview of Section 4, not experimental evidence.

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 two available sources are imperfect in different ways. A preliminary predictor $\hat{v}$ supplies dense, structured, outcome-scale estimates, but can be systematically wrong. A causally identified score Φ can be conditionally unbiased or doubly robust under appropriate nuisance conditions, yet remain noisy under weak overlap or a small randomized budget. When causal evidence is scarce, how strongly should it revise the preliminary decision, and should that revision differ across treatment comparisons? Figure 1 previews the resulting frontier.

Structured-treatment methods provide dense predictors (Kaddour et al., 2021; Shi et al., 2025). Orthogonal and doubly robust methods provide causal scores (Robins et al., 1994; Chernozhukov et al., 2018; Foster & Syrgkanis, 2023). Direct policy learning instead optimizes assignment value (Kitagawa & Tetenov, 2018; Kallus, 2018; Athey & Wager, 2021), while shrinkage and prediction-powered methods regulate the use of noisy predictive information (Su et al., 2020; Li & Ignatiadis, 2025; Sargent et al., 2026). Taken separately, these ingredients do not directly determine how a causal discrepancy should modify a globally normalized multi-action target.

Softmax is unchanged when the same constant is added to every treatment value. Hence only $K-1$ relative treatment contrasts affect the normalized decision. For $K \geq 3$, those pairwise comparisons must also be compatible with one joint probability vector. This motivates a *target-level correction operator on treatment-contrast space*: an operator that determines which parts of the discrepancy between the preliminary and causal sources enter the final target.

Reliability-adaptive target orthogonalization (RATO) makes this operator A explicit. The choices $A = 0$, $A = I$, and $A = aI$ correspond to no correction, full first-order correction, and one common correction fraction. A directional A allows different contrasts to receive different correction strengths. Two criteria then lead to different questions. First-order orthogonality asks whether small preliminary-predictor errors have zero first-order effect on the target; within this family it selects $A = I$. Local target risk asks which operator minimizes the quadratic target error after accounting for errors in both information sources. Under the stated zero cross-second-moment condition, its population oracle is $A^\star = B(B+S)^{-1}$, where B and S are total second moments of contrast errors relative to the unknown true values. Stronger correction can suppress preliminary error while transmitting more score noise, so the two criteria need not agree.

The formula for $A^\star$ is an oracle characterization, not a generic estimator. Because the true value vector is unknown, routine deployment data do not generally identify the total error moments B and S. External calibration or another valid anchor can make directional estimation possible. Without such information, our directly selectable procedure restricts to $A = aI$ and chooses (a, λ) by independent validation.

The experiments are organized around these distinctions. Controlled simulations check the local-risk predictions and the gap between local and exact nonlinear target error. Retrospective reanalyses of two randomized message studies examine how the validation-selected scalar rule changes with randomized-information budget (Hackenburg et al., 2025; 2026).

Our contributions are threefold: **(i)** a target-level orthogonality frontier and basis-invariant oracle characterization; **(ii)** finite-grid validation and target-to-regret guarantees for the scalar restriction; and **(iii)** controlled and randomized-data analyses of the reliability response and its limits.

2 Related Work

We position RATO along three lines, organized by the statistical object being protected and the geometry in which uncertainty is resolved. Appendix C provides a fuller object-level comparison.

(1) Orthogonal estimation and policy learning. Orthogonal and doubly robust methods protect a causal estimand, score, or loss against nuisance perturbations (Robins et al., 1994; Chernozhukov et al., 2018; Foster & Syrgkanis, 2023), while direct policy learning optimizes assignment value over a policy class (Kitagawa & Tetenov, 2018; Kallus, 2018; Athey & Wager, 2021). RATO starts after a causal score already exists: its question is how much of that score's discrepancy from a preliminary predictor should enter a globally normalized multi-action target. This separates nuisance orthogonality from the downstream question of target-level correction strength. **(2) Shrinkage and prediction-powered estimation.** Forecast combination, shrinkage, and prediction-powered methods adapt the use of noisy predictive information (Bates & Granger, 1969; Su et al., 2020; Li & Ignatiadis,

2025). In RATO, reliability adaptation is defined on the $K-1$ treatment-contrast quotient induced by the decision target. The same geometry carries an independently defined orthogonality endpoint: $A = I$ answers a first-order robustness question, whereas $A^\star$ answers a local target-risk question. **(3) Many, structured, and generated treatments.** Structured-treatment models exploit shared treatment representations (Kaddour et al., 2021; Shi et al., 2025), while generated-treatment settings motivate large candidate libraries (Hackenburg et al., 2025; 2026; Brooks et al., 2025). Pairwise ranking methods instead protect treatment ordering (Arno et al., 2026). A jointly normalized multi-action target imposes a global compatibility constraint across treatment comparisons when $K \geq 3$, formalized by Proposition 8 in Appendix D.1. Our question therefore begins after the treatment library and preliminary predictor already exist: which decision-relevant causal discrepancies should revise the joint target?

3 Causal Decision Problem

A correction is meaningful only after the decision object and information boundary are fixed. We first separate deployment from learning information, then define the smooth decision target and the treatment-contrast geometry on which correction acts.

Notation. Let $[K] = \{1, \dots, K\}$, let $\mathbf{1}$ denote the all-ones vector, and let I denote an identity matrix of appropriate dimension. Define $\Delta_K = \{p \in \mathbb{R}^K : p_a \geq 0, \sum_a p_a = 1\}$. We use $\mathbb{E}$ and $\mathbb{P}$ for expectation and probability; $\|\cdot\|_2$, $\|\cdot\|_{\mathrm{op}}$, and $\|\cdot\|_F$ for Euclidean, operator, and Frobenius norms; and $\mathrm{diag}(\cdot)$ and $\mathrm{tr}(\cdot)$ for the diagonal and trace operators. Superscript $\top$ denotes transpose, and subscript c denotes coordinates in $\mathbf{1}^\perp$.

3.1 Decision Target and Treatment Contrasts

An observation is $O = (C, L, Z, D, Y)$. Here C is the deployment context, $Z = (z_1, \dots, z_K)$ is the candidate set, $D \in [K]$ is the assigned treatment, and $Y \in [0, 1]$ is the observed outcome. The pre-treatment variables L may be used for adjustment during learning, but deployment uses only (C, Z). Write $H = (C, L, Z)$.

Let $Y(a)$ denote the potential outcome under treatment a (Acharki et al., 2023). We define $v_a(C, Z) = \mathbb{E}[Y(a) \mid C, Z]$, averaging over learning-only information. With $v = (v_1, \dots, v_K)^\top$ and $v_* = \max_a v_a$, a policy $\pi(C, Z) \in \Delta_K$ has

$$V(\pi) = \mathbb{E}\langle \pi, v \rangle, \qquad R_{\mathrm{st}}(\pi) = \mathbb{E}[v_* - \langle \pi, v \rangle], \qquad R_{\mathrm{dec}}(\pi) = \mathbb{E}[v_* - v_{d_\pi}], \tag{1}$$

where $d_\pi = \arg\max_a \pi_a$ with fixed tie-breaking. Thus $V(\pi)$ is the value of the stochastic assignment distribution, R_{st} is its stochastic regret, and R_{dec} is the regret after decoding the largest policy probability to a single action.

Hard maximization changes discontinuously near ties. We therefore use softmax as a smooth, entropy-regularized reference target. The temperature λ controls how concentrated the target is: for $t \in \mathbb{R}^K$ and $\lambda > 0$,

$$q_\lambda(t)_a = \frac{\exp(t_a/\lambda)}{\sum_{j=1}^K \exp(t_j/\lambda)}, \qquad J_\lambda(t) = \lambda^{-1}\left\{\mathrm{diag}(q_\lambda(t)) - q_\lambda(t) q_\lambda(t)^\top\right\}. \tag{2}$$

The reference target is $q_\lambda(v)$; smaller λ concentrates mass on better treatments, while $J_\lambda(t)$ converts value perturbations into first-order changes of the normalized target. Because $q_\lambda(t + c\mathbf{1}) = q_\lambda(t)$, only relative treatment values affect the target, so vectors differing by $c\mathbf{1}$ are decision-equivalent. We therefore represent their $(K-1)$-dimensional quotient by $U \in \mathbb{R}^{K \times (K-1)}$, whose orthonormal columns span $\mathbf{1}^\perp$, with $U^\top U = I$ and $U^\top \mathbf{1} = 0$. Writing $P = UU^\top = I - \mathbf{1}\mathbf{1}^\top/K$ and $v_c = U^\top v$, the map $U^\top$ removes the common level while retaining all decision-relevant contrasts.

For $K \geq 3$, these contrasts are not independent pairwise decisions. A single softmax vector imposes compatibility among all pairwise probabilities; equivalently, pairwise log-odds around a cycle must agree with one common set of treatment scores. Proposition 8 in Appendix D.1 gives the formal characterization. Since $J_\lambda(t)\mathbf{1} = 0$, first-order changes of the normalized target depend only on this contrast space.

3.2 Preliminary Prediction and Causal Scores

Neither the contrast geometry nor normalization reveals the true value vector v. The fixed preliminary predictor $\hat{v}(C, Z)$ provides a dense outcome-scale approximation, which may remain systematically misspecified. Causal feedback provides a second value source.

Let $\mathcal{F}$ contain the data and randomness used to fit $\hat{v}$ and the auxiliary models entering the causal score. These auxiliary models are the *nuisance functions*. We analyze score observations that are independent of $\mathcal{F}$. With cross-fitting, each observation is scored using nuisance estimates trained on other folds; clustered data are separated at the independent-unit level (Chernozhukov et al., 2018).

Define the outcome regression $m_a(H) = \mathbb{E}[Y \mid D = a, H]$ and the propensity score $e_a(H) = \mathbb{P}(D = a \mid H)$. The usual identification conditions have concrete meanings. Consistency links the observed outcome to the potential outcome under the received treatment. Conditional exchangeability rules out residual treatment confounding given H. Positivity requires each relevant treatment to have nonzero assignment probability in the strata where it is evaluated (Acharki et al., 2023). Under these conditions, $v_a(C, Z) = \mathbb{E}[m_a(H) \mid C, Z]$.

For $\mathcal{F}$-measurable nuisance estimates, we use the standard augmented inverse-propensity-weighted (AIPW) score (Robins et al., 1994; Dudík et al., 2011):

$$\Phi_a(O) = \hat{m}_a(H) + \frac{\mathbf{1}\{D = a\}}{\hat{e}_a(H)}\big(Y - \hat{m}_a(H)\big). \tag{3}$$

Below, Φ denotes the *correction-level causal value source*: one AIPW score vector, or an average of independent score vectors, targeting the same conditional value and carrying both causal residual information and sampling noise. Its outcome nuisance $\hat{m}_a$ may use L and need not equal the preliminary predictor $\hat{v}_a(C, Z)$; no treatment is designated as a control arm. Estimated propensities are clipped below $\epsilon > 0$ only for numerical stability, not as a substitute for positivity.

Figure 2 places the two sources in the same contrast coordinates. Write $\hat{v}_c = U^\top \hat{v}$, $\Phi_c = U^\top \Phi$, and $d_c = \Phi_c - \hat{v}_c = U^\top(\Phi - \hat{v})$. Then d_c is the contrast-level information available for correction, while $\hat{v} - v$ and $\Phi - v$ are respectively the preliminary-prediction and causal-source errors.

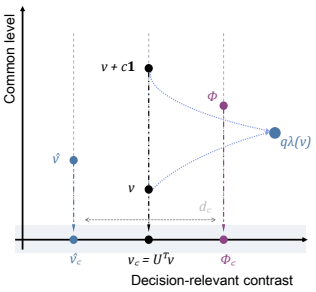

Figure 2: **Decision geometry and the two value sources.** Common shifts leave the normalized decision unchanged. The preliminary source $\hat{v}$ and causal source Φ therefore matter through their contrast coordinates, whose discrepancy supplies the information available for correction.

Lemma 1 (Conditional score bias)**.** *Under the preceding assumptions and integrability of the terms below,*

$$\mathbb{E}[\Phi_a \mid C, Z, \mathcal{F}] - v_a = \mathbb{E}\left[\left(1 - \frac{e_a}{\hat{e}_a}\right)(\hat{m}_a - m_a) \,\middle|\, C, Z, \mathcal{F}\right] =: b_a. \tag{4}$$

In particular, $b_a = 0$ if either $\hat{m}_a = m_a$ or $\hat{e}_a = e_a$ almost surely.

Lemma 1 concerns nuisance error inside Φ; preliminary error $\hat{v} - v$ is a separate downstream channel. Thus a doubly robust score can still be noisy and does not by itself determine the appropriate correction strength.

3.3 Information Conditions and Reliability Regimes

The assumptions enter at different stages. Sample separation makes fitted sources fixed for the conditional analysis. Positivity supports causal identification. Finite second moments are used for local target-risk calculations. Bounded score ranges are needed only for the finite-grid validation theorem, with uniform overlap imposed where stated. None of these assumptions requires the preliminary predictor to be correctly specified; Appendix C summarizes their roles.

The clean local frontier additionally assumes a zero *cross second moment* between preliminary and score errors. This removes the cross term in the quadratic local risk but does not require independence. Because B and S are uncentered second moments, independence alone need not imply a zero cross

second moment in the presence of persistent bias. Appendix D.8 treats the general nonzero-cross-moment case. Averaging m comparable independent score contributions reduces their variance at rate $1/m$ but does not remove persistent bias. We use m as a controlled reliability axis in the experiments, not as a universal model of how every causal score scales.

Section 4 introduces the operator that determines how much of d_c enters the target and whether correction differs across contrasts.

4 The Orthogonality Frontier

The discrepancy between the two sources is not yet a decision correction: it must be restricted to decision-relevant contrasts and propagated through normalization. Proposition 2 identifies the unique orthogonality endpoint, and Theorem 3 exposes its competing error channels. Theorem 4 and Corollaries 5–6 then characterize the reliability-optimal population frontier, after which Theorem 7 turns to validation-based deployment.

4.1 Target Orthogonality and Error Decomposition

For $A \in \mathbb{R}^{(K-1)\times(K-1)}$, define

$$T_A = q_\lambda(\hat{v}) + J_\lambda(\hat{v})UAU^\top(\Phi - \hat{v}). \tag{5}$$

A is the object left unspecified even when both value sources already exist. Writing $d_c = U^\top(\Phi - \hat{v})$, A acts on the contrast discrepancy before $J_\lambda(\hat{v})U$ maps it into target space. The cases $A = 0$, $A = aI$, general A, and $A = I$ give no, uniform partial, contrast-specific, and full first-order correction, respectively (Figure 3).

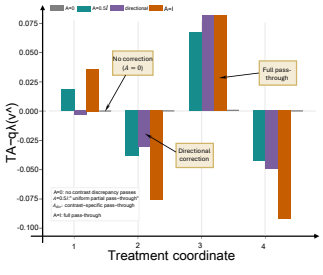

Figure 3: **How the correction operator changes the normalized target.** The contrast discrepancy is transformed by A before $J_\lambda(\hat{v})U$ propagates it to the target. The cases $A = 0$, $A = aI$, directional A, and $A = I$ represent no, uniform partial, contrast-specific, and full pass-through.

Proposition 2 (Uniqueness on the contrast quotient)**.** *Fix v and perturb the preliminary vector while holding the score mean at v. Then*

$$D_t\{q_\lambda(t) + J_\lambda(t)UAU^\top(v - t)\}_{t=v}[h] = J_\lambda(v)U(I - A)U^\top h. \tag{6}$$

The derivative vanishes for every ranking-relevant direction $h \in \mathbf{1}^\perp$ if and only if $A = I_{K-1}$.

Proposition 2 identifies the orthogonality endpoint, but that endpoint need not minimize target error. To expose the competing error channels, write $e_c = U^\top(\hat{v} - v)$, $b_c = U^\top b$, and $\Sigma_\Phi = \mathrm{Cov}(\Phi \mid C, Z, \mathcal{F})$. Let $\tilde{v} = v + UU^\top(\hat{v} - v)$; shift invariance gives $T_A(\hat{v}) = T_A(\tilde{v})$.

Theorem 3 (Conditional target decomposition)**.** *Suppose*

$$\sup_{r\in[v,\tilde{v}]} \|J_\lambda(r)U\|_{\mathrm{op}} \le M_\lambda,$$

and suppose J_λ is L_λ-Lipschitz on the segment joining v and $\tilde{v}$. Then, for every deterministic A conditional on $\mathcal{F}$,

$$\|\mathbb{E}[T_A \mid C, Z, \mathcal{F}] - q_\lambda(v)\| \le \frac{L_\lambda}{2}\|e_c\|^2 + M_\lambda \|I - A\|_{\mathrm{op}} \|e_c\| + M_\lambda \|A\|_{\mathrm{op}} \|b_c\|, \tag{7}$$

$$\mathrm{Cov}(T_A \mid C, Z, \mathcal{F}) = J_\lambda(\hat{v})UA(U^\top\Sigma_\Phi U)A^\top U^\top J_\lambda(\hat{v})^\top. \tag{8}$$

Theorem 3 makes the competing error channels explicit. Moving toward full correction suppresses first-order preliminary sensitivity, but it also transmits more of the causal source's bias and variability. Thus first-order orthogonality does not by itself identify the operator that minimizes local target error; that choice depends on the relative reliability of the two sources.

4.2 RELIABILITY-OPTIMAL FRONTIER

To place both error channels on the same scale, fix v and consider the local experiment $\hat{v}_t = v + tU\xi$ and $\Phi_t = v + tU\zeta$ with $\mathbb{E}[\xi\zeta^\top] = 0$. The scalar t is a perturbation index rather than a sample-size parameter. The theorem describes a fixed local reliability regime, with finite relative scales absorbed into ξ and ζ, and hence into B and S. Sequences whose relative error scales themselves diverge would require a separate triangular-array analysis. The local errors need not be centered: $B = \mathbb{E}[\xi\xi^\top]$ and $S = \mathbb{E}[\zeta\zeta^\top]$ are finite *total* second moments and therefore include persistent local bias. To first order,

$$t^{-1}\{T_{A,t} - q_\lambda(v)\} = J_\lambda(v)U\{(I - A)\xi + A\zeta\} + o_{L_2}(1). \tag{9}$$

Define $W = U^\top J_\lambda(v)^\top J_\lambda(v)U$ and $\mathcal{R}(A) = \lim_{t\downarrow 0} t^{-2}\mathbb{E}\,\|T_{A,t} - q_\lambda(v)\|_2^2$. Because softmax has strictly positive coordinates, $J_\lambda(v)$ has null space $\mathrm{span}(\mathbf{1})$; since U spans $\mathbf{1}^\perp$, $J_\lambda(v)U$ has full column rank and hence $W \succ 0$.

Theorem 4 (Reliability-optimal correction)**.** *Under the local experiment above, suppose*

$$\mathbb{E}\,\|\xi\|_2^2 + \mathbb{E}\,\|\zeta\|_2^2 < \infty, \qquad \mathbb{E}[\xi\zeta^\top] = 0, \qquad B + S \succ 0.$$

Then the limit defining $\mathcal{R}(A)$ exists and equals

$$\mathcal{R}(A) = \mathrm{tr}\big[W\{(I - A)B(I - A)^\top + ASA^\top\}\big]. \tag{10}$$

Its unique minimizer over all real $(K-1)\times(K-1)$ operators is

$$A^\star = B(B + S)^{-1}, \tag{11}$$

and every competitor satisfies the exact excess-risk identity

$$\mathcal{R}(A) - \mathcal{R}(A^\star) = \mathrm{tr}\{W(A - A^\star)(B + S)(A - A^\star)^\top\} \geq 0. \tag{12}$$

If B and S do not commute, $A^\star$ need not be symmetric; the invariant object is the induced action-space operator $UA^\star U^\top$.

Theorem 4 separates two mathematical criteria. The endpoint $A = I$ answers a first-order sensitivity question. The oracle $A^\star$ answers a local target-risk question under the stated moment condition. The matrix W cancels from the unrestricted oracle formula, but it still determines target risk, excess risk, and the optimal scalar restriction.

Corollary 5 (Scalar local-risk optimum and directional separation)**.** *Restrict to $A = aI$ and let $b_W = \mathrm{tr}(WB)$ and $s_W = \mathrm{tr}(WS)$. Then*

$$a^\star = \frac{b_W}{b_W + s_W}, \qquad \min_a \mathcal{R}(aI) = \frac{b_W s_W}{b_W + s_W}. \tag{13}$$

If $b_W, s_W > 0$, then $\min_a \mathcal{R}(aI) < \min\{\mathcal{R}(0), \mathcal{R}(I)\}$. Moreover, if $A^\star$ is not a scalar multiple of I, then $\mathcal{R}(A^\star) < \mathcal{R}(aI)$ for every scalar a.

Within the local criterion, an interior scalar correction can have lower risk than either endpoint. A scalar rule nevertheless applies the same fraction to every contrast, so it cannot represent contrast-specific reliability. The next corollary checks that the directional oracle is not an artifact of the arbitrary contrast basis.

Corollary 6 (Reliability spectrum and basis invariance)**.** *Assume $B, S \succeq 0$ and $B + S \succ 0$. Every eigenvalue of $A^\star = B(B + S)^{-1}$ is real and lies in $[0, 1]$. If the contrast basis is replaced by $U' = UQ$ for an orthogonal Q, then $B' = Q^\top BQ$, $S' = Q^\top SQ$, and $A^{\star\prime} = Q^\top A^\star Q$, so the induced action-space operator is unchanged: $U'A^{\star\prime}U'^\top = UA^\star U^\top$.*

Equation (11) is a population characterization: B and S are total errors relative to the unknown value vector, so routine deployment data do not generally identify $A^\star$. External calibration can support directional estimation, with estimation stability quantified by Proposition 9; without such an anchor, we turn to the validation-selected scalar family in Section 4.3.

4.3 Validation and Deployment

We therefore use $A = aI$ as a low-dimensional deployable decision class and select (a, λ) by independent validation. Importantly, the validation-selected $\hat{a}$ is not intended to estimate the local-risk coefficient $a^\star$ in Corollary 5: once the total reliability moments are unavailable, selection is performed directly for policy value over the prespecified scalar family. Each signed target is constructed on correction data, projected onto Δ_K, and scored only after the finite candidate family has been fixed.

Algorithm 1 Validation-selected RATO

Input: Independent-unit-separated $D_{\mathrm{pre}}, D_{\mathrm{corr}}, D_{\mathrm{val}}$; prespecified grids $\mathcal{A}, \Lambda$.

1: Fit and freeze $\hat{v}$ and score nuisances on D_{pre}.
2: Construct correction-level Φ on D_{corr} using the frozen nuisances.
3: **for** $(a, \lambda) \in \mathcal{A} \times \Lambda$ **do**
4: $\quad T_{a,\lambda} \leftarrow q_\lambda(\hat{v}) + J_\lambda(\hat{v})U(aI)d_c$.
5: $\quad \pi_{a,\lambda} \leftarrow \Pi_{\Delta_K}(T_{a,\lambda})$ and evaluate $\widehat{V}_{\mathrm{val}}(\pi_{a,\lambda})$ on D_{val}.
6: **end for**
7: Choose $(\hat{a}, \hat{\lambda}) \in \arg\max_{a,\lambda} \widehat{V}_{\mathrm{val}}(\pi_{a,\lambda})$ with fixed tie-breaking.
8: Freeze $\pi_{\hat{a},\hat{\lambda}}$ before any test evaluation or deployment; **return** $\pi_{\hat{a},\hat{\lambda}}$.

Because the candidate family is fixed before validation, a weighted Hoeffding bound followed by a finite union bound yields a direct selection guarantee (Hoeffding, 1963).

Theorem 7 (Finite-grid value selection)**.** *Let $\Pi = \{\pi_1, \ldots, \pi_M\}$ be fixed independently of G independent validation units. For deterministic $w_g \geq 0$ with $\sum_g w_g = 1$, suppose every score $X_g(\pi)$ lies in an interval of deterministic length r_g, and define $\hat{V}(\pi) = \sum_g w_g X_g(\pi)$ and $V(\pi) = \sum_g w_g \mathbb{E}X_g(\pi)$. If $\hat{\pi} \in \arg\max_{\pi\in\Pi} \hat{V}(\pi)$, then with probability at least $1 - \delta$,*

$$V(\pi^*) - V(\hat{\pi}) \leq 2\sqrt{\frac{\log(2M/\delta)}{2} \sum_{g=1}^{G} w_g^2 r_g^2}, \qquad \pi^* \in \arg\max_{\pi\in\Pi} V(\pi). \tag{14}$$

For i.i.d. observation-level IPW scores with $Y \in [0, 1]$ and overlap $e_D(H) \geq \epsilon$, $w_g = 1/G$ and $r_g = 1/\epsilon$. Bounded AIPW scores and prespecified cluster aggregates follow by inserting their score ranges and weights.

For the equal-weight IPW specialization, Theorem 7 reduces to

$$V(\pi^\star) - V(\hat{\pi}) \leq \frac{\sqrt{2\log(2M/\delta)}}{\epsilon\sqrt{G}}.$$

Thus enlarging a prespecified correction–temperature grid incurs only a logarithmic selection penalty, while weak overlap directly enlarges the score range. Theorem 7 covers a fixed finite candidate family scored by deterministic-weight averages. The real-data selector in Section 5 instead uses context-standardized ratios with realized context denominators; its frozen-policy inference is therefore handled separately in Appendix E. Target error can additionally be transferred to stochastic and decoded deployment regret after simplex projection; the corresponding smoothing, projection, and margin calibration is given in Proposition 10 in Appendix D.1.

5 Experiments

The controlled experiments are organized around three theory-facing predictions. First, improving causal-score precision should increase the local-risk-optimal scalar correction toward its persistent-bias-limited value (Corollary 11). Second, contrast-specific reliability should separate directional from scalar correction (Theorem 4 and Corollary 5). Third, source correlation should alter both directional and scalar optima through the cross moment (Proposition 14). We separately test the local-to-exact approximation, calibration stability, and deployment margin effects in Propositions 12, 9, and 10.

Table 1: **Experimental settings.**

Quantity	Setting
Score averaging m	$\{1, 2, 4, 8, 16, 32, 64\}$
Synthetic seeds	96 main; 24 anisotropy
Scalar correction a	$\{0, .125, \ldots, 1\}$
Real-data λ	$\{.03, .06, .10, .20, .40\}$
Budget (%)	$\{5, 10, 20, 40, 60, 80, 100\}$
Budget repetitions	50
Participant folds	4 test / 2 pre / 1 corr. / 1 tune
Test bootstrap	10,000 cluster draws

The randomized reanalyses examine how the validation-selected scalar correction changes with randomized correction-and-tuning information, with selected correction as the primary endpoint and held-out value as a downstream outcome. The second study probes a weak-signal regime in which stronger correction need not translate into a value gain.

• **Controlled experiments:** We use Gaussian perturbations of treatment values to isolate the target-level theory rather than reproduce the bounded-outcome causal DGP of Section 3. We compare plug-in, full correction, scalar and directional local-risk oracles, calibration-estimated directionality, and score-only softmax across causal-score precision and contrast reliability. Population oracle moments are available only because simulation truth is known.

• **Real-world data:** Scaling-LLM contains 19,529 observations from 17,711 participant clusters, 10 contexts, and 72 actions per context; Roleplay-LLM contains 3,298 participants, 2 contexts, and 81 randomized bundles (Hackenburg et al., 2025; 2026). Participant-disjoint preliminary, correction, tuning, and test roles separate policy construction and selection from held-out evaluation. The preliminary predictor is fitted before correction and fixed thereafter. We compare uniform, plug-in, full correction, DR-DIRECT, and RATO under matched information access.

• **Ablations:** Only correction and tuning data are thinned as the randomized-information budget varies; preliminary fitting and test data remain fixed, jointly varying correction precision and selection noise. Across-thinning quantiles summarize repeated training/tuning perturbations; held-out uncertainty uses source-stratified participant-cluster bootstrap intervals. Complete grids, split-rotation diagnostics, and reconstruction details are in Appendices F–G.

The controlled studies manipulate reliability directly and measure target error, whereas the randomized reanalyses vary the information available for correction and selection. Together they probe the same reliability-adaptation mechanism from theory-facing and deployable perspectives.

5.1 Main Results: Reliability Response and Boundary Regimes

First, with preliminary reliability fixed, increasing causal-score precision makes stronger correction less costly in the controlled design, as suggested by the error decomposition. Second, the local-risk ordering is visible under the exact nonlinear target map, although finite perturbations leave a measurable gap from the quadratic approximation.

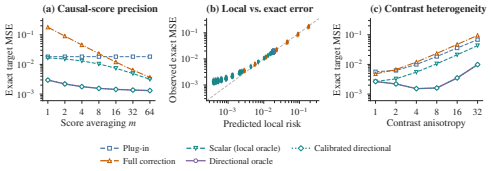

Figure 4: **Controlled checks of the local reliability frontier.** (a) Score precision with fixed preliminary reliability. (b) Predicted local risk versus observed exact MSE; the diagonal denotes equality. (c) Scalar and directional rules under increasing contrast heterogeneity. Lines connect recorded settings; bands are mean ± 1.96 Monte Carlo standard errors (96 seeds in (a), 24 in (c)). Oracles and gold-standard calibration are simulation controls.

Table 2 reports the corresponding exact-MSE grid across the full score-precision sweep. At $m = 8$, mean exact target MSE is 0.001558 for the directional local-risk oracle, 0.01037 for the scalar local-risk oracle, 0.01825 for plug-in, and 0.02312 for full correction. Under anisotropy 32, the mean seed-wise scalar-to-directional exact-MSE ratio is 4.47. These controlled finite-perturbation results are consistent with Corollary 5, whose all-scalar statement applies to the local quadratic risk rather than every finite-perturbation nonlinear MSE.

Table 2: **Controlled target error across causal-score precision.** Exact MSE ↓; oracle quantities use simulation truth.

m	Plug-in	Full	Scalar	**Directional**	Rel. red. (%)
1	.018251	.178958	.016606	**.002965**	82.1
2	.018251	.089897	.015254	**.002195**	85.6
4	.018251	.045376	.013146	**.001778**	86.5
8	.018251	.023123	.010367	**.001558**	85.0
16	.018251	.012002	.007410	**.001439**	80.6
32	.018251	.006445	.004905	**.001368**	72.1
64	.018251	.003669	.003162	**.001312**	58.5

The randomized-data reanalyses show how the validation-selected scalar rule responds to the available randomized budget in these two studies. In Scaling-LLM, mean selected correction is .59 at 5%, .80 at 40%, and about .94 at 60–80%; the full-budget split selects full correction. At 10% budget, mean held-out values are .5544 for RATO, .5465 for plug-in, .5562 for full correction, and .5581 for DR-DIRECT. We do not interpret these differences as a uniform superiority result.

Roleplay-LLM provides a boundary case. Selected correction becomes stronger as more randomized information is made available, while held-out value remains close to or below the plug-in and uniform references, separating correction strength from the strength of the underlying treatment signal.

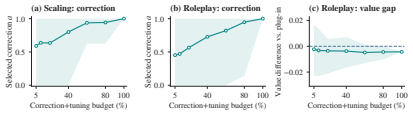

Figure 5: **Selected correction and held-out value across randomized-information budgets.** (a,b) Mean selected a. (c) Roleplay's same-thinning RATO-minus-plug-in value. Shading denotes 2.5–97.5% across-thinning quantiles. Full-budget repetitions are identical. Complete comparator curves appear in Appendix J.

Table 3 reports the fixed-test audit.

Table 3: **Real-data value and theory-facing robustness.** Panel A reports fixed-test stochastic value with 95% participant-cluster bootstrap intervals. Panel B reports exact target MSE under departures from diagonal, independent reliability.

Panel A: Randomized-data held-out value

Method	Scaling-LLM	Roleplay-LLM
Uniform	.5434 [.5361, .5503]	.5506 [.5444, .5570]
DR-DIRECT	.5524 [.5168, .5873]	.5498 [.5415, .5581]
Plug-in	.5462 [.5378, .5544]	.5499 [.5433, .5567]
Full correction	.5525 [.5081, .5955]	.5456 [.5337, .5575]
RATO	.5525 [.5081, .5955]	.5456 [.5337, .5575]

Panel B: Controlled robustness — exact MSE ↓

Stress	Directional	Scalar
Noncommuting B, S	**.00404** ($.00405_{\mathrm{cal}}$)	.01045
Correlated sources ($\rho = .5$)	**.00421**	.01261

cal denotes the calibration-estimated directional operator. At full budget, $\hat{a} = 1$ in both studies, so RATO and Full coincide in Panel A by construction. For the correlated row, the scalar rule uses the cross-moment optimum in Proposition 14.

In the noncommuting design, mean exact MSE is .00404 for the directional local-risk oracle versus .01045 for the scalar oracle. With correlated sources at $\rho = .5$, using the cross-moment scalar optimum from Proposition 14, the corresponding means are .00421 and .01261. These controlled finite-perturbation results are consistent with the local-risk theory, without implying global nonlinear optimality or real-world directional identifiability. Complete curves and persistent-bias stress tests appear in Appendix J.

5.2 Ablation and Sensitivity Analyses

Ablations vary correction structure ($A = 0, A = I, A = aI$), directional calibration, score averaging, temperature, calibration size, randomized budget, and participant splits. Complete grids and split rotations appear in Appendix J. Sensitivity varies $\lambda \in \{.05, .10, .20, .40\}$, calibration size $n_{\mathrm{cal}} \in \{128, 512, 2048\}$, score averaging m, reliability ratio, randomized budget, and participant split. Smaller calibration samples increase operator error. Scaling's plug-in/full/interior selection shares change from $20/44/36\%$ at 5% budget to $0/72/28\%$ at 80%. Complete grids and all 15 participant-split rotations are reported in Appendix J.

5.3 Efficiency and Qualitative/Error Analysis

Directional reliability requires $O(K^2)$ storage and an $O(K^3)$ solve. At $K = 256$, the implementation uses .99 MiB for storing B, S and records a median construction time of 2.85 ms. Full-budget Scaling provides 4,374 correction-and-tuning participants; complete timing details appear in Appendix J. A 1,000-context synthetic margin audit shows decoded errors of $36.8\%, 10.8\%, 2.0\%, 0\%$ across increasing true-reference margin quartiles, confirming that decisions stabilize away from near ties as predicted by Proposition 10. Additional overlap and implementation diagnostics are discussed in Appendix I.

6 Discussion, Limitations, and Conclusion

RATO identifies the target-level correction left unspecified by a preliminary predictor and a causal score. Within this family, $A = I$ removes first-order preliminary sensitivity, while local target risk selects a reliability-dependent operator on the treatment-contrast space. For $K \geq 3$, global normalization further couples treatment comparisons, making correction a joint contrast problem. Controlled experiments test the predicted frontier and directional separation, while randomized reanalyses show that selected correction strengthens with causal-information budget even when realized value responses differ across settings. The characterization assumes a fixed treatment library and predictor; margins, overlap, persistent preliminary error, and validation noise govern how target-level improvements translate into decoded decisions and policy value.

AI USE STATEMENT

Generative AI tools were used for language editing, structural refinement, formatting assistance, and exploratory discussion during manuscript preparation. The authors are responsible for all theoretical claims, experimental results, analyses, and final conclusions.

ETHICS STATEMENT

Generated persuasive treatments can be beneficial or manipulative. RATO ranks interventions for a specified outcome but does not establish that outcome's normative appropriateness. Deployment requires domain governance, informed consent where applicable, subgroup audits, and restrictions on high-stakes or coercive use.

REPRODUCIBILITY STATEMENT

Appendices D–G contain proofs, inference assumptions, protocols, and reconstruction commands. Appendix I distinguishes source inconsistencies from confirmed results. The artifact preserves original code and cached outputs, adds an independent panel renderer and a frozen-design reproduction wrapper, and records which checks were executed. Participant-level source data are not redistributed.

REFERENCES

Naoufal Acharki, Ramiro Lugo, Antoine Bertoncello, and Josselin Garnier. Comparison of meta-learners for estimating multi-valued treatment heterogeneous effects. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pp. 91–132, 2023.

Henri Arno, Dennis Frauen, Emil Javurek, Thomas Demeester, and Stefan Feuerriegel. Rank-learner: Orthogonal ranking of treatment effects. In *Proceedings of the 43rd International Conference on Machine Learning*, volume 306, 2026.

Susan Athey and Stefan Wager. Policy learning with observational data. *Econometrica*, 89(1): 133–161, 2021.

J. M. Bates and C. W. J. Granger. The combination of forecasts. *Journal of the Operational Research Society*, 20(4):451–468, 1969.

Marc Brooks, Gabriel Durham, Kihyuk Hong, and Ambuj Tewari. Generator-mediated bandits: Thompson sampling for GenAI-powered adaptive interventions. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68, 2018.

Miroslav Dudík, John Langford, and Lihong Li. Doubly robust policy evaluation and learning. In *Proceedings of the 28th International Conference on Machine Learning*, pp. 1097–1104, 2011.

Christian Fong and Justin Grimmer. Causal inference with latent treatments. *American Journal of Political Science*, 67(2):374–389, 2023.

Dylan J. Foster and Vasilis Syrgkanis. Orthogonal statistical learning. *The Annals of Statistics*, 51(3): 879–908, 2023.

Dennis Frauen, Valentyn Melnychuk, Jonas Schweisthal, Mihaela van der Schaar, and Stefan Feuerriegel. Treatment effect estimation for optimal decision-making. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Xingzhuo Guo, Yuchen Zhang, Jianmin Wang, and Mingsheng Long. Estimating heterogeneous treatment effects: Mutual information bounds and learning algorithms. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pp. 12108–12121, 2023.

Kobi Hackenburg, Ben M. Tappin, Paul Röttger, Scott A. Hale, Jonathan Bright, and Helen Margetts. Scaling language model size yields diminishing returns for single-message political persuasion. *Proceedings of the National Academy of Sciences*, 122(10):e2413443122, 2025.

Kobi Hackenburg, Lujain Ibrahim, Ben M. Tappin, and Manos Tsakiris. Comparing the persuasiveness of role-playing large language models and human experts on polarized U.S. political issues. *AI & Society*, 41(1):351–361, 2026.

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

Jean Kaddour, Yuchen Zhu, Qi Liu, Matt J. Kusner, and Ricardo Silva. Causal effect inference for structured treatments. In *Advances in Neural Information Processing Systems*, volume 34, pp. 24841–24854, 2021.

Nathan Kallus. Balanced policy evaluation and learning. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Rudolf E. Kalman. A new approach to linear filtering and prediction problems. *Journal of Basic Engineering*, 82(1):35–45, 1960.

Toru Kitagawa and Aleksey Tetenov. Who should be treated? empirical welfare maximization methods for treatment choice. *Econometrica*, 86(2):591–616, 2018.

Sida Li and Nikolaos Ignatiadis. Prediction-powered adaptive shrinkage estimation. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267, pp. 34836–34875, 2025.

Min Qian and Susan A. Murphy. Performance guarantees for individualized treatment rules. *The Annals of Statistics*, 39(2):1180–1210, 2011.

James M. Robins, Andrea Rotnitzky, and Lue Ping Zhao. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89(427): 846–866, 1994.

Gabriel Sargent, Will Wei Sun, Zhengwu Zhang, and Yufeng Liu. Prediction-powered adaptive inference with pretrained AI models for contextual bandits. In *Proceedings of the 43rd International Conference on Machine Learning*, volume 306, 2026.

Lei Shi, David Arbour, Raghavendra Addanki, Ritwik Sinha, and Avi Feller. Leveraging semantic similarity for experimentation with AI-generated treatments. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Yi Su, Maria Dimakopoulou, Akshay Krishnamurthy, and Miroslav Dudík. Doubly robust off-policy evaluation with shrinkage. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119, pp. 9167–9176, 2020.

Jiaqi Tong and Fan Li. Orthogonal double residual learning for optimal individualized treatment rules. *arXiv preprint arXiv:2608.24085*, 2026.

Appendix

CONTENTS

A Notation and Information Model **14**

B Extended Related Work **14**

C Extended Positioning and Design Commitments **15**

D Complete Proofs **15**

D.1 Proof map and supporting statements 15

D.2 Proof of the conditional score-bias lemma 18

D.3 Proof of uniqueness on the contrast quotient 19

D.4 Proof of global coupling 19

D.5 Proof of the conditional target decomposition 19

D.6 Proof of the reliability frontier and scalar corollary 19

D.7 Proof of estimated-frontier stability 20

D.8 Correlated information sources 20

D.9 Proof of the minimax lower bound 21

E Source-Stratified Frozen-Policy Inference **22**

F Additional Experimental Results **23**

F.1 Synthetic data-generating process and oracle construction 23

F.2 Protocol, metrics, and baseline fairness 25

F.3 Main budget results 27

F.4 Ablations 28

F.5 Sensitivity 28

F.6 Efficiency 28

F.7 Qualitative error and failure analysis 28

F.8 Fixed-split real-data audit and decoded endpoint 29

F.9 Scope boundary 30

G Data Reconstruction and Reproducibility **30**

G.1 Raw-data reconstruction 30

G.2 Artifact map and commands 30

H Reproducibility and Responsibility Statements **31**

I Scope of Oracle and Deployment Claims **31**

J Complete Publication Panels and Diagnostics **32**

A Notation and Information Model

Table 4: Core notation. The main text defines every object at first use; this table is a lookup aid rather than a substitute for the formal setup.

Object	Meaning
C, L, Z	deployment context, learning-only adjustment variables, and the K-action candidate set
D, Y	randomized/logged action and bounded outcome
$v, \hat{v}$	true conditional arm-value vector and structured preliminary predictor
Φ	reference-free randomized or doubly robust arm score
U, P	orthonormal contrast basis and projector onto $\mathbf{1}^{\perp}$
q_λ, J_λ	softmax decision map and Jacobian
T_A	reliability-adaptive target indexed by contrast operator A
B, S	total second-moment reliability matrices of preliminary and score errors
W	local decision metric $U^\top J_\lambda(v)^\top J_\lambda(v) U$
$A^\star$	oracle directional correction $B(B+S)^{-1}$ under zero cross second moment
$V, R_{\mathrm{st}}, R_{\mathrm{dec}}$	stochastic value, stochastic regret, and decoded regret

Information model. The preliminary predictor is fixed with respect to the fold used to construct Φ; the selector is fixed with respect to the test fold; and uncertainty is aggregated at the randomization unit. External pretrained representations are permitted only if their access is declared and they do not use held-out outcomes. This contract makes explicit which source supplies prediction, which supplies causal identification, and which data are reserved for selection and evaluation.

B Extended Related Work

Orthogonality and doubly robust learning. Classical augmented inverse-propensity and one-step estimators combine an outcome model and propensity model so that first-order nuisance errors cancel under suitable product-rate conditions (Robins et al., 1994; Dudík et al., 2011; Chernozhukov et al., 2018). Orthogonal statistical learning generalizes this idea to machine-learning losses (Foster & Syrgkanis, 2023). These methods protect an estimand or loss from nuisance perturbations. RATO starts one layer later: it asks how much of an already constructed causal correction should enter a nonlinear multi-action decision target when correction variance is itself decision relevant.

Direct and decision-aware policy learning. Empirical-welfare maximization and balanced policy learning optimize policy value directly (Kitagawa & Tetenov, 2018; Kallus, 2018; Athey & Wager, 2021). Plug-in policy guarantees depend on how estimation error interacts with decision margins (Qian & Murphy, 2011); recent decision-aware CATE and ranking methods further emphasize that average effect error and treatment quality can diverge (Frauen et al., 2025; Arno et al., 2026; Tong & Li, 2026). Our target is complementary: it preserves a globally normalized multi-action interface and derives an explicit reliability operator on the coupled contrast geometry.

Shrinkage and prediction-powered methods. Forecast combination and Kalman filtering motivate inverse-covariance fusion (Bates & Granger, 1969; Kalman, 1960). Doubly robust off-policy evaluation can benefit from shrinkage (Su et al., 2020), and PAS adapts shrinkage between machine-learning predictions and limited labeled observations across compound mean-estimation tasks (Li & Ignatiadis, 2025). The closest conceptual overlap is reliability adaptation. The technical difference is the protected object: RATO derives the fusion problem from first-order risk of a causal softmax target, includes full orthogonalization as a named endpoint, works on a basis-invariant action-contrast quotient, and carries target error into stochastic and decoded deployment guarantees.

Many, structured, and generated treatments. Structured-treatment representation learning exploits low-dimensional or semantic structure (Kaddour et al., 2021; Fong & Grimmer, 2023; Shi et al., 2025), while multi-valued treatment learners address many-arm effect estimation without requiring semantic structure (Acharki et al., 2023; Guo et al., 2023). Randomized LLM-message studies

demonstrate concrete large libraries (Hackenburg et al., 2025; 2026); generator-mediated bandits study online allocation when a generative model proposes treatments (Brooks et al., 2025). Our setting is offline: the library and a structured predictor already exist, identified feedback is limited, and the unresolved statistical object is the reliability-weighted decision correction.

C Extended Positioning and Design Commitments

Table 5: Closest paradigms differ in the object they protect. RATO adapts correction on the globally normalized multi-action decision target.

Paradigm	Protected object	Geometry	Distinguishing mechanism
DML / orthogonal learning	causal estimand or loss	task dependent	nuisance orthogonality
Direct policy learning	policy value	global policy class	bypasses plug-in target
DR-OPE shrinkage / PAS	scalar value or mean	scalar	reliability shrinkage
Pairwise Rank-Learner	individual ordering by effects	pairwise ranking	decision-aware orthogonality
RATO	normalized decision target	contrast quotient	directional reliability

Table 6: Each assumption has an operational role, a diagnostic, and a fallback. Only the local-risk theorem uses the last row.

Commitment	Why it is needed	What can be checked	If it fails
Consistency and exchangeability given H	identifies m_a and v_a	randomization protocol; balance; timing of variables	sensitivity analysis or a narrower causal claim
Overlap $e_a(H) \geq \epsilon$ on deployed arms	controls score moments	arm counts; propensity tails; effective sample size	coarsen or prune actions; restrict deployment support
Cross-fitting / sample separation	fixes learned objects on evaluation folds	participant IDs and fold ledger	nested splitting or stability conditions
Smooth decision map and finite score moments	controls Taylor remainder and inference	temperature, target range, score tails	increase λ, clip scores, or use robust means
Zero cross second moment for the clean B, S formula	yields the closed-form frontier	calibration cross-moment; source design	use the formula in Section D.8

D Complete Proofs

D.1 Proof map and supporting statements

The main-text proof dependencies and the supporting results moved out of the nine-page narrative are summarized below. The moved statements sharpen coherence, estimability, and information limits; none is required to interpret the central reliability frontier.

Table 7: Proof map. Main-text results are listed first; supporting statements are formally recorded below before their proofs.

Result	Main dependency / proof idea
Lemma 1	Conditional AIPW identity and iterated expectation.
Proposition 2	Softmax Jacobian on the action-contrast quotient.
Theorem 3	Taylor expansion plus Lemma 1.
Theorem 4	Local expansion and quadratic completion in A.
Corollaries 5 and 6	Theorem 4 plus scalar restriction / similarity transform.
Corollary 11	Decomposition of total score error into persistent bias and $1/m$ sampling variance.
Proposition 12	Taylor remainder control linking local risk to exact nonlinear target MSE.
Proposition 14	Quadratic completion after retaining the source cross second moment.
Theorem 7	Hoeffding concentration and a finite-grid union bound.
Proposition 10	Entropy smoothing, simplex non-expansiveness, and margin geometry.
Proposition 8	Additivity of pairwise log-odds around triples.
Proposition 9	Inverse perturbation plus the exact excess-risk identity.
Theorem 13	Two-point Gaussian construction and Le Cam/Pinsker.

Proposition 8 (Global coupling begins at three actions)**.** *Let $r_{ab} \in (0,1)$ be pairwise choice probabilities with $r_{ba} = 1 - r_{ab}$. There exists a globally normalized softmax vector $q_\lambda(v)$ satisfying $r_{ab} = q_{\lambda,a}(v)/(q_{\lambda,a}(v) + q_{\lambda,b}(v))$ for every pair if and only if*

$$\operatorname{logit}(r_{ab}) + \operatorname{logit}(r_{bc}) + \operatorname{logit}(r_{ca}) = 0 \tag{15}$$

for every triple of distinct actions (a, b, c). The condition is vacuous at $K = 2$ and can fail under independent pairwise corrections when $K \geq 3$.

Proposition 9 (Stability of the estimated frontier)**.** *Let $M = B + S$, $\kappa = \left\|M^{-1}\right\|_{\mathrm{op}}$, $\widehat{M} = \widehat{B} + \widehat{S}$, $\varepsilon_B = \left\|\widehat{B} - B\right\|_{\mathrm{F}}$, and $\delta_M = \left\|\widehat{M} - M\right\|_{\mathrm{op}}$. If $\kappa\delta_M < 1$, then $\widehat{A} = \widehat{B}\widehat{M}^{-1}$ exists and*

$$\left\|\widehat{A} - A^\star\right\|_{\mathrm{F}} \leq \frac{\kappa}{1 - \kappa\delta_M}\{\varepsilon_B + \kappa \left\|B\right\|_{\mathrm{F}} \delta_M\}, \tag{16}$$

$$\mathcal{R}(\widehat{A}) - \mathcal{R}(A^\star) \leq \left\|W\right\|_{\mathrm{op}} \left\|M\right\|_{\mathrm{op}} \left\|\widehat{A} - A^\star\right\|_{\mathrm{F}}^2. \tag{17}$$

Thus, for fixed K, $\kappa = O(1)$, and total-moment estimation errors tending to zero, the estimated rule attains oracle local risk, including for noncommuting B and S.

Proposition 10 (Softmax smoothing, projection, and decoding calibration)**.** *Fix $v \in \mathbb{R}^K$ and $\lambda > 0$. Let*

$$a^\star \in \arg\max_{a \in [K]} v_a, \qquad v_\star = v_{a^\star}, \qquad \Delta = v_\star - \min_{a \in [K]} v_a,$$

and write $q = q_\lambda(v)$. For any signed target $T \in \mathbb{R}^K$, let

$$\pi = \Pi_{\Delta_K}(T).$$

Then the softmax smoothing error satisfies

$$v_\star - \langle q, v\rangle \leq \lambda \log K. \tag{18}$$

If $a^\star$ is unique and

$$\gamma = v_\star - \max_{a \neq a^\star} v_a > 0,$$

then the sharper margin-dependent bound

$$v_\star - \langle q, v\rangle \leq \Delta(K-1)e^{-\gamma/\lambda} \tag{19}$$

also holds. Consequently,

$$v_\star - \langle \pi, v\rangle \leq \min\left\{\lambda \log K,\ \Delta(K-1)e^{-\gamma/\lambda}\right\} + \frac{\Delta\sqrt{K}}{2} \left\|T - q\right\|_2, \tag{20}$$

where the margin-dependent branch is omitted when the maximizer is not unique.

If $a^\star$ is unique, define the softmax probability margin

$$m_q = q_{a^\star} - \max_{a \neq a^\star} q_a > 0.$$

For the decoded action $d_\pi = \arg\max_a \pi_a$ with fixed tie-breaking,

$$v_\star - v_{d_\pi} \leq \frac{2\Delta}{m_q^2} \|T - q\|_2^2 . \tag{21}$$

Moreover,

$$m_q \geq \frac{1 - e^{-\gamma/\lambda}}{1 + (K-1)e^{-\gamma/\lambda}}. \tag{22}$$

Thus target error controls stochastic regret linearly and decoded regret quadratically away from softmax near-ties.

Corollary 11 (Score averaging and the persistent-bias floor)**.** *Consider the local experiment of Theorem 4. Suppose the score error based on m comparable independent contributions has the form*

$$\zeta_m = \mu_\zeta + \frac{1}{m}\sum_{j=1}^{m} \varepsilon_j,$$

where

$$\mathbb{E}[\varepsilon_j] = 0, \qquad \mathbb{E}[\varepsilon_j \varepsilon_j^\top] = \Sigma_\zeta,$$

the ε_j are mutually independent, and the zero cross-second-moment condition $\mathbb{E}[\xi \zeta_m^\top] = 0$ holds. Then

$$S_m := \mathbb{E}[\zeta_m \zeta_m^\top] = \mu_\zeta \mu_\zeta^\top + \frac{1}{m}\Sigma_\zeta. \tag{23}$$

For the scalar family $A = aI$, write

$$b_W = \operatorname{tr}(WB), \qquad s_{\infty,W} = \mu_\zeta^\top W \mu_\zeta, \qquad v_W = \operatorname{tr}(W\Sigma_\zeta).$$

The local-risk optimal scalar correction is

$$a_m^\star = \frac{b_W}{b_W + s_{\infty,W} + v_W/m}. \tag{24}$$

Hence $a_m^\star$ is nondecreasing in m, and is strictly increasing when $b_W > 0$ and $v_W > 0$. Moreover,

$$\lim_{m\to\infty} a_m^\star = \frac{b_W}{b_W + s_{\infty,W}}. \tag{25}$$

Thus additional independent causal information removes the sampling variance component at rate $1/m$, but persistent score bias creates a nonvanishing reliability floor. In particular, $a_m^\star \to 1$ only when the persistent score-error component vanishes in the decision-relevant metric.

Proof of Proposition 12. Write

$$\hat{v}_t = v + tU\xi, \qquad \Phi_t = v + tU\zeta.$$

The definition of $T_{A,t}$ gives

$$\begin{aligned} T_{A,t} - q_\lambda(v) = q_\lambda(v + tU\xi) - q_\lambda(v) \\ + tJ_\lambda(v + tU\xi)UA(\zeta - \xi). \end{aligned}$$

By the integral Taylor formula and the L_λ-Lipschitz property of J_λ,

$$q_\lambda(v + tU\xi) - q_\lambda(v) = tJ_\lambda(v)U\xi + r_{1,t},$$

with

$$\|r_{1,t}\|_2 \leq \frac{L_\lambda}{2} t^2 \|\xi\|_2^2 .$$

Likewise,

$$tJ_\lambda(v + tU\xi)UA(\zeta - \xi) = tJ_\lambda(v)UA(\zeta - \xi) + r_{2,t},$$

where

$$\|r_{2,t}\|_2 \le L_\lambda t^2 \|A\|_{\mathrm{op}} \|\xi\|_2 \|\zeta - \xi\|_2 .$$

Therefore

$$T_{A,t} - q_\lambda(v) = tG_A + r_{A,t}, \qquad r_{A,t} = r_{1,t} + r_{2,t},$$

and Equation (26) follows.

Under the zero cross-second-moment condition,

$$\mathbb{E}\,\|G_A\|_2^2 = \mathcal{R}(A)$$

by Theorem 4. Hence

$$t^{-2}\mathbb{E}\,\|T_{A,t} - q_\lambda(v)\|_2^2 - \mathcal{R}(A) = 2t^{-1}\mathbb{E}\langle G_A, r_{A,t}\rangle + t^{-2}\mathbb{E}\,\|r_{A,t}\|_2^2 .$$

By Cauchy–Schwarz and Equation (26),

$$2t^{-1}\,|\mathbb{E}\langle G_A, r_{A,t}\rangle| \le 2L_\lambda t\sqrt{\mathbb{E}\,\|G_A\|_2^2\ \mathbb{E}H_A^2},$$

and

$$t^{-2}\mathbb{E}\,\|r_{A,t}\|_2^2 \le L_\lambda^2 t^2 \mathbb{E}H_A^2.$$

Substituting $\mathbb{E}\,\|G_A\|_2^2 = \mathcal{R}(A)$ proves Equation (27). □

Proposition 12 (Finite-perturbation accuracy of the local frontier)**.** *Under the local experiment of Theorem 4, fix a deterministic operator A. Suppose J_λ is L_λ-Lipschitz in operator norm on every segment joining v and $v + tU\xi$ for $0 \le t \le t_0$, and suppose ξ and ζ have finite fourth moments. Define*

$$G_A = J_\lambda(v)U\{(I - A)\xi + A\zeta\}$$

and

$$H_A = \frac{1}{2}\,\|\xi\|_2^2 + \|A\|_{\mathrm{op}}\,\|\xi\|_2\,\|\zeta - \xi\|_2 .$$

Then for $0 \le t \le t_0$,

$$T_{A,t} - q_\lambda(v) = tG_A + r_{A,t}, \qquad \|r_{A,t}\|_2 \le L_\lambda t^2 H_A. \tag{26}$$

Under the zero cross-second-moment condition of Theorem 4,

$$\left|t^{-2}\mathbb{E}\,\|T_{A,t} - q_\lambda(v)\|_2^2 - \mathcal{R}(A)\right| \le 2L_\lambda t\sqrt{\mathcal{R}(A)\,\mathbb{E}H_A^2} + L_\lambda^2 t^2 \mathbb{E}H_A^2. \tag{27}$$

Hence the rescaled exact nonlinear target MSE differs from the quadratic local risk by $O(t)$ for every fixed A satisfying the stated moment conditions.

Theorem 13 (Overlap-sensitive minimax lower bound)**.** *Let $K \ge 2$, $e_{\min} > 0$, and $\sigma > 0$. Consider a separate Gaussian observation model with n independent draws, a fixed logging distribution, and an arm whose assignment probability is exactly $e_{\min}$. Conditional on $D = a$, let $Y \sim N(v_a, \sigma^2)$, with mean range at most $\Delta_{\max}$. For every decision rule $\hat{d}$ based on these observations, there exists a mean vector in this model such that*

$$\mathbb{E}[v_* - v_{\hat{d}}] \ge \frac{1}{8}\min\left\{\Delta_{\max}, \frac{\sigma}{\sqrt{ne_{\min}}}\right\}. \tag{28}$$

Under uniform logging, this is $\Omega\{\min(\Delta_{\max}, \sigma\sqrt{K/n})\}$. In this unrestricted Gaussian submodel, auxiliary semantic prediction that imposes no valid restriction on arm means cannot improve the stated lower bound. Structural restrictions can evade the bound by excluding the two-point submodel.

D.2 Proof of the conditional score-bias lemma

Condition on $(H, \mathcal{F})$. Independence of the evaluation unit from the training data, consistency, and exchangeability give

$$\mathbb{E}[\Phi_a \mid H, \mathcal{F}] = \hat{m}_a + \frac{e_a}{\hat{e}_a}\mathbb{E}[Y - \hat{m}_a \mid D = a, H, \mathcal{F}] = \hat{m}_a + \frac{e_a}{\hat{e}_a}(m_a - \hat{m}_a).$$

Subtracting m_a yields $(1 - e_a/\hat{e}_a)(\hat{m}_a - m_a)$. Taking the conditional expectation over L given $(C, Z, \mathcal{F})$ and using evaluation–training independence, so $\mathbb{E}[m_a \mid C, Z, \mathcal{F}] = v_a$, proves Equation (4). If either $\hat{e}_a = e_a$ or $\hat{m}_a = m_a$, the product is zero. □

D.3 Proof of uniqueness on the contrast quotient

Define $F_A(t) = q_\lambda(t) + J_\lambda(t)UAU^\top(v-t)$. The product rule gives

$$DF_A(t)[h] = J_\lambda(t)h + DJ_\lambda(t)[h]UAU^\top(v-t) - J_\lambda(t)UAU^\top h.$$

At $t = v$, the middle term vanishes. Since $J_\lambda(v)\mathbf{1} = 0$ and $P = UU^\top$, $J_\lambda(v)h = J_\lambda(v)UU^\top h$, which proves Equation (6). Softmax has strictly positive coordinates, so $J_\lambda(v)$ has null space $\mathrm{span}(\mathbf{1})$ and rank $K-1$. Consequently $J_\lambda(v)U$ has full column rank. The derivative is zero for every $h \in \mathbf{1}^\perp$ exactly when $(I-A)U^\top h = 0$ for every such h. Because $U^\top$ maps $\mathbf{1}^\perp$ onto $\mathbb{R}^{K-1}$, this is equivalent to $A = I$. □

D.4 Proof of global coupling

If the pairwise probabilities arise from $q_\lambda(v)$, then

$$\mathrm{logit}(r_{ab}) = (v_a - v_b)/\lambda.$$

Summing around a triple proves necessity of Equation (15). Conversely, fix a reference action K and set $v_K = 0$ and $v_a = \lambda\,\mathrm{logit}(r_{aK})$. Applying the cycle identity to (a, b, K) gives $\mathrm{logit}(r_{ab}) = (v_a - v_b)/\lambda$, so the softmax vector built from v has every stated pairwise marginal. For $K = 2$ there is no triple. For $K \geq 3$, independently changing one pairwise logit while holding the other two fixed violates the identity, proving the final claim. □

D.5 Proof of the conditional target decomposition

By shift invariance, replace $\hat{v}$ by $\tilde{v} = v + UU^\top(\hat{v} - v)$ without changing the target; hence $\tilde{v} - v = Ue_c$. Let $\bar{\Phi} = \mathbb{E}[\Phi \mid C, Z, \mathcal{F}] = v + b$. Conditional expectation of the target is

$$q_\lambda(\tilde{v}) + J_\lambda(\tilde{v})UAU^\top(v + b - \tilde{v}).$$

Because $J_\lambda(\tilde{v})P = J_\lambda(\tilde{v})$, add and subtract the full correction to obtain

$$\begin{aligned}\mathbb{E}[T_A \mid C, Z, \mathcal{F}] - q_\lambda(v) = q_\lambda(\tilde{v}) - q_\lambda(v) - J_\lambda(\tilde{v})Ue_c \\ + J_\lambda(\tilde{v})U(I-A)e_c + J_\lambda(\tilde{v})UAb_c.\end{aligned}$$

The integral Taylor formula on $[v, \tilde{v}]$ bounds the first line by $L_\lambda \|e_c\|^2/2$. Submultiplicativity and the derivative bound give the remaining terms in Equation (7). Conditional on $(C, Z, \mathcal{F})$, only Φ is random. Applying $\mathrm{Cov}(MX) = M\,\mathrm{Cov}(X)M^\top$ with $M = J_\lambda(\hat{v})UAU^\top$ and $U^\top U = I$ proves Equation (8). □

D.6 Proof of the reliability frontier and scalar corollary

A first-order expansion in t gives $t^{-1}\{T_{A,t} - q_\lambda(v)\} \to J_\lambda(v)U\{(I-A)\xi + A\zeta\}$ in L_2 under the stated second-moment and smoothness conditions. The assumed zero cross second moment removes the cross term; no centering of either source is required because B and S are total second moments. Taking squared Euclidean norm and using $\mathbb{E}[x^\top Mx] = \mathrm{tr}(M\mathbb{E}[xx^\top])$ gives the limit in Equation (10).

Set $C = B + S$. Direct expansion shows

$$\begin{aligned}(I-A)B(I-A)^\top + ASA^\top \\ = B - AB - BA^\top + ACA^\top.\end{aligned}$$

Since $A^\star C = B$, expansion of $(A - A^\star)C(A - A^\star)^\top$ gives exactly the difference between the preceding expression at A and at $A^\star$. Multiplication by W, trace, and $W \succ 0$, $C \succ 0$ establish Equation (12) and uniqueness.

For $A = aI$, $\mathcal{R}(aI) = (1-a)^2 b_W + a^2 s_W$. Differentiation gives $a^\star = b_W/(b_W + s_W)$ and substitution gives $b_W s_W/(b_W + s_W)$, which is strictly below both b_W and s_W when both are positive. Finally, if $A^\star$ is not scalar, then $aI - A^\star \neq 0$ for every a; positive definiteness makes Equation (12) strictly positive. □

Proof of Corollary 6. Let $M = B + S \succ 0$. The operator $A^\star = BM^{-1}$ is similar to the symmetric matrix $M^{-1/2}BM^{-1/2}$ because $M^{-1/2}A^\star M^{1/2} = M^{-1/2}BM^{-1/2}$. Since $0 \preceq B \preceq M$, the latter has spectrum in $[0, 1]$, proving the eigenvalue claim. Under $U' = UQ$, orthogonality of Q gives $B' = Q^\top BQ$ and $S' = Q^\top SQ$, hence $A^{\star\prime} = Q^\top B(B + S)^{-1}Q = Q^\top A^\star Q$ and therefore $U'A^{\star\prime}U'^\top = UA^\star U^\top$. □

Proof of Corollary 11. By independence and centering of the ε_j,

$$\begin{aligned} S_m &= \mathbb{E}\left[\left(\mu_\zeta + \frac{1}{m}\sum_{j=1}^m \varepsilon_j\right)\left(\mu_\zeta + \frac{1}{m}\sum_{j=1}^m \varepsilon_j\right)^\top\right] \\ &= \mu_\zeta\mu_\zeta^\top + \frac{1}{m^2}\sum_{j=1}^m \mathbb{E}[\varepsilon_j\varepsilon_j^\top] \\ &= \mu_\zeta\mu_\zeta^\top + \frac{1}{m}\Sigma_\zeta, \end{aligned}$$

which proves Equation (23). Therefore

$$\operatorname{tr}(WS_m) = s_{\infty,W} + \frac{v_W}{m}.$$

Applying the scalar formula in Corollary 5 gives

$$a_m^\star = \frac{b_W}{b_W + s_{\infty,W} + v_W/m},$$

proving Equation (24). The denominator is nonincreasing in m, which proves monotonicity and strict monotonicity under $b_W, v_W > 0$. Taking $m \to \infty$ gives Equation (25). □

D.7 Proof of estimated-frontier stability

Write $\widehat{M} = M + E$. The perturbation condition and the Banach lemma give $\left\|\widehat{M}^{-1}\right\|_{\mathrm{op}} \le \kappa/(1 - \kappa\delta_M)$. The resolvent identity and $\widehat{A} - A^\star = (\widehat{B} - B)\widehat{M}^{-1} + B(\widehat{M}^{-1} - M^{-1})$ yield Equation (16) after applying Frobenius–operator submultiplicativity. The excess identity Equation (12) and $\operatorname{tr}(WXMX^\top) \le \|W\|_{\mathrm{op}}\|M\|_{\mathrm{op}}\|X\|_{\mathrm{F}}^2$ give Equation (17). □

D.8 Correlated information sources

The zero-cross-second-moment assumption is convenient but not structurally necessary. Let

$$C_{\xi\zeta} = \mathbb{E}[\xi\zeta^\top]$$

and define

$$M_c = B + S - C_{\xi\zeta} - C_{\xi\zeta}^\top = \mathbb{E}[(\zeta - \xi)(\zeta - \xi)^\top].$$

Proposition 14 (Reliability frontier with correlated sources)**.** *Without imposing $\mathbb{E}[\xi\zeta^\top] = 0$, the local target risk is*

$$\begin{aligned} \mathcal{R}_c(A) = \operatorname{tr} W\{&(I - A)B(I - A)^\top + ASA^\top \\ &+ (I - A)C_{\xi\zeta}A^\top + AC_{\xi\zeta}^\top(I - A)^\top\}. \end{aligned} \tag{29}$$

If $M_c \succ 0$, the unique unrestricted minimizer is

$$A_c^\star = (B - C_{\xi\zeta})M_c^{-1} = (B - C_{\xi\zeta})\{B + S - C_{\xi\zeta} - C_{\xi\zeta}^\top\}^{-1}. \tag{30}$$

Moreover,

$$\mathcal{R}_c(A) - \mathcal{R}_c(A_c^\star) = \operatorname{tr}\big[W(A - A_c^\star)M_c(A - A_c^\star)^\top\big] \ge 0. \tag{31}$$

For the scalar family $A = aI$, define

$$b_W = \operatorname{tr}(WB), \qquad s_W = \operatorname{tr}(WS), \qquad c_W = \operatorname{tr}(WC_{\xi\zeta}).$$

Then

$$\mathcal{R}_c(aI) = (1-a)^2 b_W + a^2 s_W + 2a(1-a)c_W, \tag{32}$$

whose unrestricted minimizer is

$$a_c^\star = \frac{b_W - c_W}{b_W + s_W - 2c_W}. \tag{33}$$

If deployment restricts $a \in [0,1]$, the optimal scalar rule is the projection of this value onto $[0,1]$.

Proof. Write

$$(I-A)\xi + A\zeta = \xi + A(\zeta - \xi).$$

Since

$$\mathbb{E}[(\zeta-\xi)(\zeta-\xi)^\top] = M_c,$$
$$\mathbb{E}[\xi(\zeta-\xi)^\top] = C_{\xi\zeta} - B,$$

and

$$\mathbb{E}[(\zeta-\xi)\xi^\top] = C_{\xi\zeta}^\top - B,$$

the second-moment matrix of the local target error is

$$B + A(C_{\xi\zeta}^\top - B) + (C_{\xi\zeta} - B)A^\top + AM_cA^\top,$$

which is equivalent to Equation (29).

Because

$$A_c^\star M_c = B - C_{\xi\zeta},$$

direct quadratic completion gives

$$\begin{aligned} &B + A(C_{\xi\zeta}^\top - B) + (C_{\xi\zeta} - B)A^\top + AM_cA^\top \\ &\quad = B + A_c^\star(C_{\xi\zeta}^\top - B) + (C_{\xi\zeta} - B)A_c^{\star\top} + A_c^\star M_c A_c^{\star\top} \\ &\qquad + (A - A_c^\star)M_c(A - A_c^\star)^\top. \end{aligned}$$

Multiplication by W, taking traces, and $W \succ 0$, $M_c \succ 0$ prove Equation (30)–Equation (31) and uniqueness.

For $A = aI$, symmetry of W implies

$$\operatorname{tr}(WC_{\xi\zeta}^\top) = \operatorname{tr}(WC_{\xi\zeta}),$$

so Equation (29) reduces to Equation (32). Differentiating gives

$$\frac{d}{da}\mathcal{R}_c(aI) = -2(b_W - c_W) + 2a(b_W + s_W - 2c_W),$$

and therefore Equation (33). □ □

D.9 Proof of the minimax lower bound

Choose an action with logging probability $e_{\min}$ and another available action. Construct two experiments P_+ and P_- that agree except that the first action has mean $+\delta$ under P_+ and $-\delta$ under P_-; set the means of every other arm to zero in both experiments. The best decision is the first action under P_+ and any other action under P_-. Their n-sample Kullback–Leibler divergence is

$$\mathrm{KL}(P_+^n \| P_-^n) = ne_{\min}\frac{(2\delta)^2}{2\sigma^2} = \frac{2ne_{\min}\delta^2}{\sigma^2}.$$

Take $\delta = \min\{\Delta_{\max}, \sigma/(2\sqrt{ne_{\min}})\}$. Then the divergence is at most $1/2$, so Pinsker gives total variation at most $1/2$. Le Cam's two-point inequality implies that the average probability of choosing the wrong side is at least $(1 - \mathrm{TV})/2 \geq 1/4$. The average regret is therefore at least $\delta/4$, which is no smaller than the right side of Equation (28). □

E SOURCE-STRATIFIED FROZEN-POLICY INFERENCE

Let the evaluation sample contain a fixed finite set of released sampling sources $s \in \mathcal{S}$, with participant clusters $g = 1, \dots, G_s$, total $G = \sum_s G_s$, and $p_{s,G} = G_s/G$. A policy $\hat{\pi}$ and nuisances are trained without these clusters and are therefore fixed conditional on $\mathcal{F}$. For a finite context set $\mathcal{C}$, define the score R_{sgi} and cluster cell totals

$$R_{sgi} = \frac{\hat{\pi}_{D_{sgi}}(C_{sgi}, Z_{sgi})}{\hat{e}_{D_{sgi}}(H_{sgi})}\{Y_{sgi} - \hat{m}_{D_{sgi}}(H_{sgi})\} + \sum_a \hat{\pi}_a(C_{sgi}, Z_{sgi})\hat{m}_a(H_{sgi}),$$

$$N_{sgc} = \sum_{i \in (s,g)} \mathbf{1}(C_{sgi} = c), \qquad Q_{sgc} = \sum_{i \in (s,g)} \mathbf{1}(C_{sgi} = c) R_{sgi},$$

$$\rho_{sc} = \mathbb{E}[N_{sgc} \mid \mathcal{F}, s], \qquad \nu_{sc} = \mathbb{E}[Q_{sgc} \mid \mathcal{F}, s],$$

$$\bar{\rho}_{c,G} = \sum_s p_{s,G}\rho_{sc}, \qquad \theta_{c,G} = \frac{\sum_s p_{s,G}\nu_{sc}}{\bar{\rho}_{c,G}},$$

$$H_{sg} = \frac{1}{|\mathcal{C}|}\sum_{c \in \mathcal{C}} \frac{Q_{sgc} - \theta_{c,G}N_{sgc}}{\bar{\rho}_{c,G}}, \qquad h_{s,G} = \mathbb{E}[H_{sg} \mid \mathcal{F}, s], \qquad S_{sg,G} = H_{sg} -$$

By construction, $\sum_s p_{s,G}h_{s,G} = 0$. Put $V_{\mathcal{F},G} = |\mathcal{C}|^{-1}\sum_c \theta_{c,G}$ and $\widehat{V}_{\mathrm{eq}} = |\mathcal{C}|^{-1}\sum_c(\sum_{s,g} Q_{sgc})/(\sum_{s,g} N_{sgc})$.

Theorem 15 (Conditional stratified-cluster CLT and sandwich inference)**.** *Assume, conditional on $\mathcal{F}$, that clusters are independent across sources and i.i.d. within each source; $\mathcal{S}$ and $\mathcal{C}$ are fixed; $p_{s,G} \to p_s > 0$; and $\min_c \bar{\rho}_{c,G} \geq \rho_0 > 0$ with probability tending to one. For some $\eta > 0$, assume*

$$\max_{s \in \mathcal{S}} \mathbb{E}\left[\sum_c (|N_{sgc}|^{2+\eta} + |Q_{sgc}|^{2+\eta}) \middle| \mathcal{F}, s\right] = O_p(1),$$

$$\sigma_G^2 = G^{-1}\sum_s G_s \operatorname{Var}(S_{sg,G} \mid \mathcal{F}, s) \to_p \sigma^2 \in (0, \infty).$$

These conditions imply the conditional Lyapunov condition for the finite-source triangular array. Then

$$\sqrt{G}(\widehat{V}_{\mathrm{eq}} - V_{\mathcal{F},G}) = G^{-1/2}\sum_{s \in \mathcal{S}}\sum_{g=1}^{G_s} S_{sg,G} + o_p(1) \rightsquigarrow N(0, \sigma^2) \tag{34}$$

conditionally in probability. Define $\hat{\rho}_c = G^{-1}\sum_{s,g} N_{sgc}$, $\hat{\theta}_c = (\sum_{s,g} Q_{sgc})/(\sum_{s,g} N_{sgc})$,

$$\hat{H}_{sg} = \frac{1}{|\mathcal{C}|}\sum_c \frac{Q_{sgc} - \hat{\theta}_c N_{sgc}}{\hat{\rho}_c}, \qquad \bar{H}_s = G_s^{-1}\sum_g \hat{H}_{sg}, \qquad \hat{\sigma}^2 = G^{-1}\sum_s\sum_g (\hat{H}_{sg} - \bar{H}_s)^2.$$

Then $\hat{\sigma}^2 \to_p \sigma^2$ and $\sqrt{G}(\widehat{V}_{\mathrm{eq}} - V_{\mathcal{F},G})/\hat{\sigma} \rightsquigarrow N(0,1)$. If additionally $\sqrt{G}\{V_{\mathcal{F},G} - V_{\mathrm{eq,G}}(\hat{\pi})\} \to_p 0$, where $V_{\mathrm{eq,G}}$ is the context-standardized causal value for the fixed source mixture, as under exact randomization propensities or a suitable product-rate consequence of Lemma 1, the same statements hold with $V_{\mathrm{eq,G}}(\hat{\pi})$ in place of $V_{\mathcal{F},G}$.

Proof. Within-source conditional laws of large numbers and $p_{s,G} \to p_s$ give $G^{-1}\sum_{s,g} N_{sgc} \to_p \bar{\rho}_{c,G}$ uniformly over the finite context set, so empirical denominators are positive with probability tending to one. The multivariate delta method applied to the context numerator and denominator means yields Equation (34); the deterministic source means cancel because $\sum_s p_{s,G}h_{s,G} = 0$. The displayed moment bound, finite $\mathcal{S}$, and nondegenerate σ_G^2 imply Lyapunov, so the conditional independent-array CLT gives the limit. Within-source centering makes $\hat{H}_{sg} - \bar{H}_s$ converge in mean square to $S_{sg,G}$ on average. The stratified law of large numbers therefore gives $\hat{\sigma}^2 \to_p \sigma^2$. Studentization and the causal-centering replacement follow by Slutsky. □

Proposition 16 (Source-stratified participant-cluster bootstrap)**.** *Under the conditions of Theorem 15, draw independently across sources*

$$(W_{s1}, \dots, W_{sG_s}) \sim \mathrm{Multinomial}\big(G_s; G_s^{-1}, \dots, G_s^{-1}\big),$$

and recompute every cell ratio as $\hat{\theta}_c^* = (\sum_{s,g} W_{sg} Q_{sgc})/(\sum_{s,g} W_{sg} N_{sgc})$ *and* $\widehat{V}_{\mathrm{eq}}^* = |\mathcal{C}|^{-1} \sum_c \hat{\theta}_c^*$. *Then, conditionally on the evaluation data and on the event of positive weighted cell denominators,*

$$\frac{\sqrt{G}(\widehat{V}_{\mathrm{eq}}^* - \widehat{V}_{\mathrm{eq}})}{\hat{\sigma}} \rightsquigarrow N(0,1) \quad \textit{in probability.}$$

The probability of a nonpositive weighted denominator tends to zero. Consequently percentile or studentized intervals formed from the weighted ratios consistently approximate the frozen-policy evaluation law.

Proof. Because $\sum_g (W_{sg} - 1) = 0$ within every source, the bootstrap linear term automatically centers each empirical source influence:

$$\sum_{s,g} (W_{sg} - 1)\hat{H}_{sg} = \sum_{s,g} (W_{sg} - 1)(\hat{H}_{sg} - \bar{H}_s).$$

The $2 + \eta$ moment condition makes the largest centered influence negligible relative to their quadratic sum. The finite-stratum multinomial bootstrap CLT and the bootstrap delta method therefore give

$$\sqrt{G}(\widehat{V}_{\mathrm{eq}}^* - \widehat{V}_{\mathrm{eq}}) = G^{-1/2} \sum_{s,g} (W_{sg} - 1)(\hat{H}_{sg} - \bar{H}_s) + o_p^*(1),$$

with conditional variance converging to σ^2. Finally, the weighted context denominators converge conditionally to their positive limits and $\hat{\sigma}^2 \to_p \sigma^2$, proving all claims. □

F Additional Experimental Results

F.1 Synthetic data-generating process and oracle construction

The controlled experiments instantiate the same two-source local experiment used in Section 4. Their purpose is not to reproduce the bounded-outcome causal data-generating process of Section 3, but to isolate the target-level reliability mechanism under known population error moments.

For each synthetic context, let $v \in \mathbb{R}^K$ denote the true treatment-value vector and let $U \in \mathbb{R}^{K \times (K-1)}$ be an orthonormal basis for $\mathbf{1}^\perp$. The preliminary and causal value sources are generated as

$$\hat{v}_t = v + tU\xi, \qquad \Phi_t = v + tU\zeta, \tag{35}$$

where $t > 0$ controls the finite perturbation size and $\xi, \zeta \in \mathbb{R}^{K-1}$ are jointly Gaussian contrast errors.

We parameterize their joint law by

$$\begin{pmatrix} \xi \\ \zeta \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mu_\xi \\ \mu_\zeta \end{pmatrix}, \begin{pmatrix} \Sigma_\xi & \Sigma_{\xi\zeta} \\ \Sigma_{\xi\zeta}^\top & \Sigma_\zeta \end{pmatrix} \right). \tag{36}$$

The corresponding total second moments are

$$B = \mathbb{E}[\xi\xi^\top] = \Sigma_\xi + \mu_\xi \mu_\xi^\top, \qquad S = \mathbb{E}[\zeta\zeta^\top] = \Sigma_\zeta + \mu_\zeta \mu_\zeta^\top, \tag{37}$$

and the cross second moment is

$$C_{\xi\zeta} = \mathbb{E}[\xi\zeta^\top] = \Sigma_{\xi\zeta} + \mu_\xi \mu_\zeta^\top. \tag{38}$$

Thus the simulation uses the same uncentered reliability objects as the theory and retains persistent bias whenever the corresponding error mean is nonzero.

For every simulated draw and every correction operator A, we evaluate the same normalized target as in the main text:

$$T_{A,t} = q_\lambda(\hat{v}_t) + J_\lambda(\hat{v}_t) U A U^\top (\Phi_t - \hat{v}_t). \tag{39}$$

The reference target is $q_\lambda(v)$. Reported exact nonlinear error is

$$R_{\mathrm{exact}}(A) = \mathbb{E}\left[\|T_{A,t} - q_\lambda(v)\|_2^2 \right], \tag{40}$$

whereas the theory-facing quadratic criterion is the local risk $\mathcal{R}(A)$ from Theorem 4. The two quantities are computed separately; Proposition 12 describes their local relationship but does not assert global finite-t optimality.

Clean zero-cross-moment design. In the baseline reliability experiments, the joint Gaussian parameters are chosen so that

$$C_{\xi\zeta} = \mathbb{E}[\xi\zeta^\top] = 0.$$

The directional local-risk oracle is therefore

$$A^\star = B(B+S)^{-1}, \tag{41}$$

and the scalar oracle restricted to $A = aI$ is

$$a^\star = \frac{\operatorname{tr}(WB)}{\operatorname{tr}(WB) + \operatorname{tr}(WS)}, \qquad W = U^\top J_\lambda(v)^\top J_\lambda(v) U. \tag{42}$$

The plug-in and full-correction endpoints correspond to $A = 0$ and $A = I$, respectively.

Causal-score precision. To vary causal-source precision while holding preliminary reliability fixed, the score error is generated from

$$\zeta_m = \mu_\zeta + \frac{1}{m}\sum_{j=1}^{m} \varepsilon_j, \qquad \mathbb{E}[\varepsilon_j] = 0, \qquad \mathbb{E}[\varepsilon_j \varepsilon_j^\top] = \Sigma_\zeta, \tag{43}$$

with mutually independent contributions. Hence

$$S_m = \mu_\zeta \mu_\zeta^\top + \frac{1}{m}\Sigma_\zeta, \tag{44}$$

matching Corollary 11. The reported sweep uses

$$m \in \{1, 2, 4, 8, 16, 32, 64\}.$$

Directional-reliability stress. The anisotropy experiments vary the spread of the contrast-specific reliability eigenvalues while retaining the same target construction. The reported anisotropy grid is

$$\{1, 2, 4, 8, 16, 32\}.$$

As anisotropy increases, B and S assign increasingly different reliability to different contrast directions, creating the setting in which an unrestricted directional operator can differ materially from the best scalar restriction.

Noncommuting and correlated-source stress tests. The noncommuting design rotates the contrast reliability structure so that B and S are not simultaneously diagonalizable. The directional oracle remains

$$A^\star = B(B+S)^{-1},$$

which need not be symmetric.

The correlated-source design instead allows

$$C_{\xi\zeta} \neq 0.$$

Writing

$$M_c = B + S - C_{\xi\zeta} - C_{\xi\zeta}^\top,$$

the corresponding directional oracle is

$$A_c^\star = (B - C_{\xi\zeta}) M_c^{-1}, \tag{45}$$

as in Proposition 14. For the scalar family,

$$a_c^\star = \frac{b_W - c_W}{b_W + s_W - 2c_W}, \qquad b_W = \operatorname{tr}(WB), \quad s_W = \operatorname{tr}(WS), \quad c_W = \operatorname{tr}(WC_{\xi\zeta}), \tag{46}$$

with projection onto $[0, 1]$ when that deployment restriction is imposed. The reported correlated-source stress uses $\rho = .5$ and recomputes both directional and scalar oracles using the nonzero cross moment.

Persistent-bias stress. Persistent error is introduced through nonzero μ_ξ and/or μ_ζ. Because the oracle quantities use total rather than centered second moments, these bias components enter B, S, and, where applicable, $C_{\xi\zeta}$ directly. We therefore recompute the corresponding oracle from the total moments rather than treating repeated-fit covariance as a substitute for total error.

Calibration-estimated directionality. Because simulation truth is observed, an independent calibration sample can estimate the total contrast-error moments directly. For calibration draws $\{(\xi_i, \zeta_i)\}_{i=1}^{n_{\mathrm{cal}}}$, define

$$\widehat{B} = \frac{1}{n_{\mathrm{cal}}} \sum_{i=1}^{n_{\mathrm{cal}}} \xi_i \xi_i^\top, \qquad \widehat{S} = \frac{1}{n_{\mathrm{cal}}} \sum_{i=1}^{n_{\mathrm{cal}}} \zeta_i \zeta_i^\top. \tag{47}$$

For the zero-cross-moment design, the calibrated directional operator is

$$\widehat{A} = \widehat{B}(\widehat{B} + \widehat{S})^{-1}. \tag{48}$$

When the cross moment is estimated, we additionally use

$$\widehat{C} = \frac{1}{n_{\mathrm{cal}}} \sum_{i=1}^{n_{\mathrm{cal}}} \xi_i \zeta_i^\top \tag{49}$$

and replace the clean oracle by the corresponding sample analogue of Proposition 14. The calibration-size sensitivity uses

$$n_{\mathrm{cal}} \in \{128, 512, 2048\}.$$

All methods within a synthetic replicate share the same true value vector, preliminary draw, causal-source draw, perturbation size, temperature, and evaluation target. Oracle methods differ only in the restriction placed on the correction operator and in whether population reliability moments are available. Population oracles and gold-standard calibration controls are used only as theory-facing simulation benchmarks and are not presented as directly deployable estimators. The main frontier uses 96 independent seeds and 256 contexts per seed; the anisotropy experiments use 24 seeds. Monte Carlo intervals are formed across seed-level means rather than across contexts within a seed.

F.2 Protocol, metrics, and baseline fairness

This appendix expands the evidence supporting Section 5. It supplies complete protocol, budget-thinning results, fixed-split checks, calibration diagnostics, robustness stress tests, efficiency measurements, and failure cases; none of these sections introduces a stronger central claim than the main text.

The synthetic main frontier uses 96 independent seeds and 256 contexts per seed; confidence intervals are $\bar{r} \pm 1.95996 s_r/\sqrt{96}$ across seed-level means, never across contexts within a seed. Exact nonlinear target MSE and local theory risk are computed independently before plotting. All synthetic methods share the simulated draw, $\hat{v}$, Φ, λ, and evaluation truth; oracle and gold-standard calibration controls additionally access true error moments. The correction operators differ, and the score-only control uses its separate softmax transformation. The public-data analyses use participant-disjoint preliminary, correction, tuning, and test roles, known assignment propensities, and a frozen policy on test.

Executable baseline ledger. Table 8 records the role, construction, and matched information budget of every comparator used in the real-data analysis.

Table 8: Real-data baseline roles and information budgets. Every learned comparator is frozen before test evaluation.

Method	Reviewer role	Construction	Tuning / information budget
Uniform	Classical control	$\pi_a = 1/K$	No tuning; ignores correction data.
Plug-in	Target-family endpoint	$q_\lambda(\hat{v})$	λ selected on the same thinned tuning role; fixed preliminary predictor.
Full correction	Orthogonal endpoint	T_I from thinned correction AIPW scores	Same thinned correction+tuning participants as RATO; λ selected on tuning only.
DR-DIRECT	Finite-family direct control	$q_\lambda(\Phi_{\mathrm{corr}})$ from correction-split AIPW/DR arm scores	Same thinned correction+tuning participants; λ selected on tuning only.
RATO	Proposed adaptive target	T_{aI} with (a, λ) selected by value	Same thinned correction+tuning participants; validation-only selection.
Oracle scalar/directional	Theory-facing controls	Use ground-truth reliability in simulation	Synthetic only; never presented as deployable baselines.

DR-DIRECT is a same-information finite-family direct comparator using correction-split AIPW values and a temperature grid. It is not a general policy-class optimizer or a reproduction of every cited policy-learning algorithm; its score nuisance still uses the preliminary predictor. We do not manufacture separate PAS, DR-OPE-shrinkage, or pairwise-ranking numbers because those methods protect different statistical objects; the main text states this non-comparability rather than treating the omission as evidence of superiority.

Hyperparameter and resampling ledger. Table 9 collects every claim-relevant grid, resampling count, and selection rule.

Table 9: Claim-relevant grids and selection rules.

Quantity	Grid / value	Selection or role
Scalar correction a	$\{0, .125, .25, .375, .5, .625, .75, .875, 1\}$	Selected on tuning value for RATO.
Real-data temperature λ	$\{.03, .06, .10, .20, .40\}$	Selected on the same tuning role for plug-in, full, DR-DIRECT, and RATO.
Ridge penalty	$\{.01, .1, 1, 10, 100\}$	Inner participant-disjoint validation inside the preliminary role only.
Budget fraction	$\{.05, .10, .20, .40, .60, .80, 1\}$	Prespecified causal-information stress axis.
Budget repetitions	50 per nontrivial fraction	Training/tuning perturbation quantiles; not prospective-trial CIs.
Fixed-test bootstrap	10,000 draws	Source-stratified participant-cluster inference for the frozen policy.
Split rotations	15	Descriptive analyst/split sensitivity only; excluded from fixed-split intervals.
Synthetic λ stress	$\{.05, .10, .20, .40\}$	Sensitivity only.
Calibration size	$\{128, 512, 2048\}$	Directional-estimation sensitivity only.

Budget-thinning protocol. We first construct the same participant folds used by the frozen-policy analysis. The preliminary predictor is fitted once and the test set is never thinned. For each repetition and each budget fraction $f \in \{.05, .10, .20, .40, .60, .80, 1\}$, we subsample participants only from the correction and tuning roles within source/context strata, recompute their assignment-aware arm scores, tune the scalar (a, λ) grid on the thinned tuning role, freeze the policy, and evaluate it on the unchanged test role. Nontrivial fractions use 50 independent thinnings. The 2.5–97.5% bands in the budget figures are quantiles across those repeated training/tuning perturbations, not confidence intervals for a prospective randomized trial.

Table 10: Complete controlled score-precision grid. Entries are exact nonlinear target MSE, averaged across the 96 stored synthetic seeds. The scalar and directional rules optimize the quadratic local risk; the reported MSE is evaluated under the exact nonlinear target map.

m	Plug-in	Full	Scalar oracle	Directional oracle	Calibrated directional
1	.018251	.178958	.016606	.002965	.002972
2	.018251	.089897	.015254	.002195	.002199
4	.018251	.045376	.013146	.001778	.001780
8	.018251	.023123	.010367	.001558	.001559
16	.018251	.012002	.007410	.001439	.001440
32	.018251	.006445	.004905	.001368	.001368
64	.018251	.003669	.003162	.001312	.001312

F.3 Main budget results

Table 11: Scaling-LLM randomized-budget diagnostic. Values are means across 50 thinnings (except the deterministic full-budget row); n_{c+t} is the average number of correction+tuning participants exposed to the learner. Quantile ranges are in the released CSV.

Budget	n_{c+t}	selected a	Plug-in	Full	RATO
5%	248	.590	.54686	.55048	.54942
10%	484	.638	.54651	.55622	.55441
20%	947	.635	.54647	.55022	.55094
40%	1,855	.800	.54594	.55543	.55410
60%	2,731	.938	.54559	.55516	.55545
80%	3,574	.943	.54597	.55498	.55485
100%	4,374	1.000	.54623	.55255	.55255

At 10%, the mean selected policy uses roughly 484 randomized correction+tuning participants and attains mean held-out value 0.55441 versus 0.54651 for plug-in, 0.55622 for full correction, and 0.55812 for DR-Direct. The RATO 2.5–97.5% across-thinning range is $[0.51250, 0.59356]$, so this comparison is deliberately interpreted as mechanism evidence rather than a confirmatory superiority test. At 60%, RATO has mean 0.55545 versus 0.54559 for plug-in, 0.55516 for full correction, and 0.55310 for DR-Direct.

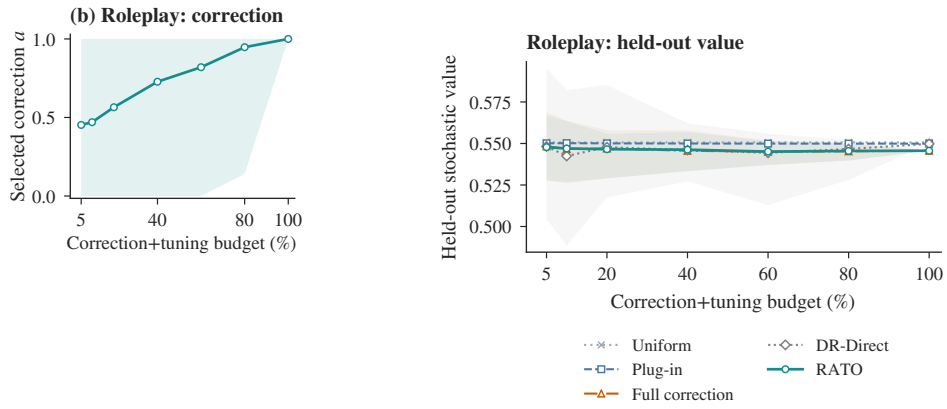

Figure 6: **Roleplay-LLM boundary stress case.** The selected correction weight increases with randomized budget, but held-out stochastic value does not improve over plug-in or uniform. This is a boundary result: reliability adaptation changes how much causal evidence is used, not whether that additional correction produces a realized value gain.

Table 12: Roleplay-LLM budget stress. Mean selected correction grows with budget, while held-out value remains lower than the plug-in mean throughout the budget curve.

Budget	selected a	Plug-in	Full	RATO
5%	.453	.55011	.54826	.54780
10%	.470	.55018	.54670	.54698
20%	.565	.55007	.54697	.54655
40%	.728	.54998	.54560	.54627
60%	.820	.54992	.54501	.54512
80%	.948	.54991	.54528	.54551
100%	1.000	.54991	.54562	.54562

F.4 Ablations

The synthetic design isolates the method components without importing neural-network-style module language. $A = 0$ removes causal correction, $A = I$ removes reliability adaptation, $A = aI$ removes directional adaptation, calibration-estimated $\hat{A}$ replaces the oracle operator, and score-only softmax removes structured preliminary information. Across the anisotropy grid $\{1, 2, 4, 8, 16, 32\}$, the directional advantage grows as reliabilities diverge. At anisotropy 32, the scalar/directional exact-MSE ratio is 4.47 with a 95% Monte Carlo half-width of 0.24 in the executable artifact.

F.5 Sensitivity

Beyond randomized budget, the artifact varies $\lambda \in \{.05, .10, .20, .40\}$, calibration size $n_{\mathrm{cal}} \in \{128, 512, 2048\}$, causal-score averaging $m \in \{1, 2, 4, 8, 16, 32, 64\}$, and the score-to-preliminary reliability ratio. Smaller calibration samples widen $\|\hat{A} - A^\star\|_F$ but do not change the oracle frontier. For the fixed-split real studies, 15 participant-fold rotations are reported only as descriptive split-sensitivity diagnostics; they are distinct from the 50 within-split budget thinnings.

A rotated design makes B and S genuinely noncommuting. Over 24 seeds, mean exact MSE is 0.00404 for the directional oracle and 0.01045 for the local-risk scalar rule; the estimated directional operator is 0.00405. With correlated information sources at $\rho = .5$, mean exact MSE is 0.00421 for the correlated-source directional oracle and 0.01261 for the scalar cross-moment optimum in Proposition 14; the calibration-estimated directional operator is 0.00422.

F.6 Efficiency

The main efficiency object is randomized information. In Scaling-LLM the correction+tuning roles contain about 4,374 participants at full budget; the 10% and 60% rows expose about 484 and 2,731, respectively. These counts are the causal-information cost underlying Figure 5. Computationally, the benchmark times only operator construction: median directional times (ms) for $K = (8, 16, 32, 64, 128, 256)$ are approximately $(0.056, 0.030, 0.047, 0.169, 0.679, 2.848)$ on the reference CPU run, compared with scalar times $(0.014, 0.014, 0.018, 0.040, 0.145, 0.769)$. Storing B and S at $K = 256$ requires about 0.99 MiB in double precision. No GPU/NPU claim is made.

Reference compute environment. The source artifact reports its reference execution in a Debian GNU/Linux 13 container with five allocated CPU cores on an AMD EPYC 9V74 processor and 5.8 GiB RAM. The recorded software stack is Python 3.13.5, NumPy 2.3.5, SciPy 1.17.0, pandas 2.2.3, scikit-learn 1.8.0, and Matplotlib 3.10.8. The central cost claim concerns randomized participants rather than accelerator throughput, so no GPU or NPU is required by the method or reported as an efficiency advantage.

F.7 Qualitative error and failure analysis

A 1,000-context audit bins examples by the true reference softmax top-two probability margin, not an estimated deployable margin. Decoded error falls from 36.8% in the lowest-margin quartile (mean margin 0.0257) to 10.8%, 2.0%, and 0% in progressively larger-margin quartiles. This is the empirical counterpart of Proposition 10: decoded decisions are non-smooth near ties even when

stochastic value varies smoothly. Additional failure modes are weak overlap, persistent preliminary misspecification invisible to replicated-fit variance, and noisy tuning under small randomized budgets.

F.8 FIXED-SPLIT REAL-DATA AUDIT AND DECODED ENDPOINT

The fixed split assigns participant folds 0–3 to test, 4–5 to preliminary fitting, 6 to correction, and 7 to tuning. Scaling-LLM uses multistage propensities $1/168$, $1/72$, and $1/24$ among AI-message actions; Roleplay-LLM uses $1/81$ for the participant-level 81-bundle design. Tuning and test use the same assignment-aware value identity. The selected stochastic policy is frozen before test evaluation. The fixed split selects $(a, \lambda) = (1, .03)$ in both studies. At the full-budget fixed split, RATO selects $a = 1$ in both studies, so its frozen test policy coincides with the full-correction endpoint; the identical entries in Panel A are therefore by construction rather than duplicated reporting.

Table 13: Supplied fixed-test context-standardized stochastic value (10,000-draw source-stratified participant-cluster bootstrap; 95% intervals). This table is a reproducibility audit, not the primary application result.

Method	Scaling-LLM ($K = 72$)	Roleplay-LLM ($K = 81$)
Uniform	.5434 [.5361, .5503]	.5506 [.5444, .5570]
DR-DIRECT	.5524 [.5168, .5873]	.5498 [.5415, .5581]
Plug-in	.5462 [.5378, .5544]	.5499 [.5433, .5567]
Full correction	.5525 [.5081, .5955]	.5456 [.5337, .5575]
RATO	.5525 [.5081, .5955]	.5456 [.5337, .5575]

Relative to plug-in, the fixed-split RATO difference is 0.00632 for Scaling-LLM (paired 95% CI $[-0.03488, 0.04681]$) and -0.00429 for Roleplay-LLM ($[-0.01446, 0.00586]$). Thus neither fixed-split study supports a superiority claim.

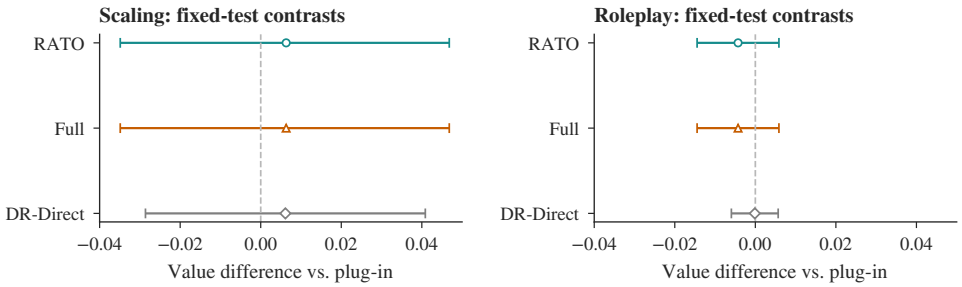

Figure 7: **Fixed-split paired diagnostics.** Scaling (left) and Roleplay (right) show stochastic-value differences from plug-in with the supplied 10,000-draw source-stratified participant-cluster bootstrap intervals. All 15 descriptive participant-split rotations are displayed separately in Appendix J; they do not enter these intervals.

Table 14: Supplied decoded value on the fixed participant-disjoint test split (10,000-draw source-stratified participant-cluster bootstrap; 95% intervals). Decoded value is a stress endpoint and was not used for tuning.

Method	Scaling-LLM ($K = 72$)	Roleplay-LLM ($K = 81$)
DR-DIRECT	.5491 [.4904, .6056]	.5591 [.5068, .6107]
Plug-in	.5744 [.4974, .6492]	.5564 [.5195, .5899]
Full correction	.5399 [.4605, .6163]	.5422 [.4851, .5927]
RATO	.5399 [.4605, .6163]	.5422 [.4851, .5927]

F.9 Scope boundary

The directional oracle is a theorem about total error second moments. In the synthetic calibration design those quantities are observable by construction. In a real deployment, repeated preliminary fits generally identify instability rather than persistent error relative to the unknown truth. We therefore do not claim that repeated fits alone estimate B, and we do not present directional real-data superiority. The budget curves test only the deployable scalar selector. When a separate calibration design does not identify total directional reliability, the operational method is validation-selected scalar RATO.

G Data Reconstruction and Reproducibility

G.1 Raw-data reconstruction

For Scaling-LLM, we read the released analysis table, retain the randomized AI-message condition, and align anti-issue outcomes by $1-Y$. Participant identity is handled explicitly: 19,525 observations use the released `PROLIFIC_PID`; one observation uses a valid released cleaned `prolific_id` that links to an existing participant; and three placeholder/test-ID observations are conservatively grouped into a single unknown-ID cluster so uncertain identity cannot cross sample roles. This yields 17,711 participant clusters. The published multistage protocol samples an issue uniformly, an AI model-size bin uniformly, a model uniformly within that bin (14 small, 6 medium, and 2 each large and extra-large), and one of three message variants uniformly. Conditional on issue and AI-message exposure, the known action propensities are therefore $1/168$, $1/72$, and $1/24$ by size bin; they sum to one within every context. We verify exactly ten contexts with 72 unique message IDs each. For Roleplay-LLM, we concatenate three released survey sources (1,634 Prolific-Democrat, 835 Prolific-Republican, and 829 Lucid-Republican complete responders), recover the released LLM treatment and stance, reconstruct the Qualtrics message assignments, and retain the participant-level randomized bundle consisting of one common author type and three independently randomized message indices. We verify 3,298 complete-response LLM-arm participants, two stance contexts, all 81 bundles in each context, and propensity $1/81$. Outcomes are the mean of the three released issue-specific response means, divided by 100 and sign-aligned so larger values favor the assigned stance.

The source manuscript records the public archives for its original rerun with the following SHA-256 hashes; those participant-level archives are not included in the present upload:

```
f2f9786e38dab4d545a6fb431f669582b3cf64086a2ede75d494fdd37c792999
01ff633ca75ba87350975ae348e265c656ae60414632bcbc86c37a8f599e89c4
```

The first line is Scaling-LLM and the second is Roleplay-LLM.

G.2 Artifact map and commands

The accompanying package preserves the original experiment code and outputs. The original quick-summary entry point below recomputes secondary tables but requires an existing main-results CSV. The added `reproduce_frozen_design.py` wrapper, run from the new artifact root, replays all reported synthetic designs without overwriting source outputs. The original commands are

```
cd experiments
pip install -r requirements.txt
PYTHONPATH=. python quick_summaries.py
# optional full grid:
PYTHONPATH=. python run_synthetic_suite.py
```

The scripts write manuscript-facing CSVs under `experiments/outputs/` and vector figures under `figures/`. Unit tests verify the contrast basis and endpoint-dominance identities.

The real-data pipeline reconstructs the canonical tables directly from the released repositories, then runs the frozen-policy analysis:

```
python real_cases/reconstruct_public_data.py \
```

```
  --scaling-root /path/to/scaling-LLM-persuasion-main \
  --roleplay-root /path/to/llm-roleplaying-experts-main \
  --out data \
  --scaling-archive /path/to/scaling-LLM-persuasion-main.zip \
  --roleplay-archive /path/to/llm-roleplaying-experts-main.zip
python real_cases/run_validation.py \
  --scaling-csv data/scaling_canonical.csv \
  --roleplay-csv data/roleplay_canonical.csv \
  --out outputs/real_results \
  --bootstrap 10000 --rotations 15 --jobs 4
python real_cases/make_real_artifacts.py
# randomized-budget diagnostic (50 repeated thinnings):
PYTHONPATH=.:real_cases python real_cases/run_budget_thinning.py \
  --scaling-csv data/scaling_canonical.csv \
  --roleplay-csv data/roleplay_canonical.csv \
  --out outputs/budget_thinning --repetitions 50
PYTHONPATH=. python real_cases/make_budget_artifacts.py
```

The artifact writes reconstruction audits, tuning grids, fixed-test values, paired intervals, rotation ledgers, budget-thinning repetitions and summaries, and the manuscript-facing real-data figures. The package README records the sample-separation contract and the boundary that repeated fits do not identify persistent preliminary bias.

H Reproducibility and Responsibility Statements

Reproducibility. The submission includes an executable synthetic pipeline, study-specific public-data reconstruction, a canonical frozen-policy evaluator, 50-repetition randomized-budget thinning, seed-level, thinning-level, and rotation-level CSV outputs, 10,000-draw bootstrap summaries, vector figures, unit tests, package requirements, and verified SHA-256 hashes for both supplied public source archives. The distributable artifact omits participant-level raw and canonical data; those files are regenerated locally from the cited public repositories by the included reconstruction script.

Statistical scope. The executable analysis enforces participant-disjoint preliminary, correction, tuning, and test roles; the resulting policy is frozen before the fixed test split is evaluated. The source-stratified participant-cluster bootstrap preserves each released sample's realized size and estimates evaluation-sample uncertainty for this frozen policy, matching Section E. Complementary 15-rotation analyses map the training, tuning, split, and analyst-choice sensitivity of the selected correction. Because the public outcomes had already been explored through these rotations, we interpret the fixed-split intervals as sample-separated evidence and the rotation display as the corresponding stability map.

Ethics. Generated persuasive treatments can be beneficial or manipulative. The method ranks interventions for a specified outcome; it does not determine whether that outcome is normatively appropriate. Deployment requires domain governance, informed consent where applicable, subgroup audits, and restrictions on high-stakes or coercive use. The public studies analyzed here report de-identified randomized data; our artifact stores no participant-level source data.

AI use. Generative AI tools were used for language editing, structural refinement, formatting assistance, and exploratory discussion during manuscript preparation. The authors are responsible for all theoretical claims, mathematical derivations, experimental results, analyses, and final conclusions.

I Scope of Oracle and Deployment Claims

This section clarifies the information required by the oracle results and the scope of the empirical comparisons. It is intended to separate structural characterizations from directly selectable procedures.

Oracle versus directly selectable correction. The directional rules in Theorem 4 and Proposition 14 depend on total contrast-error moments relative to the unknown true value vector. These moments include persistent error, not only repeated-fit variability. They are therefore not generically identified from routine deployment data. A directional implementation requires an external calibration design or another valid anchor for the relevant total moments. By contrast, Algorithm 1 selects among prespecified scalar corrections and temperatures by independent validation and does not estimate B, S, or the cross moment.

Local risk versus exact nonlinear error. The scalar and directional oracle rules minimize the quadratic local target risk. The synthetic experiments evaluate those rules under the exact nonlinear target map at finite perturbation size. Proposition 12 controls the local-to-exact approximation under its stated smoothness and moment conditions, but it does not make a local-risk oracle globally optimal for finite perturbations. Accordingly, experimental references to a scalar or directional "oracle" mean the corresponding local-risk rule. In the correlated-source stress, the scalar comparator uses the cross-moment optimum from Proposition 14 rather than the zero-cross-moment formula.

Validation and held-out inference. Theorem 7 concerns a fixed finite candidate family evaluated on independent validation units with deterministic weights and bounded score ranges. The real-data implementation instead uses context-standardized ratios with realized context denominators. We therefore do not use Theorem 7 as a finite-sample certificate for that implemented ratio selector. The source-stratified participant-cluster bootstrap in Appendix E addresses held-out uncertainty for a frozen selected policy and is a separate inferential statement.

Experimental scope. Randomized-budget thinning changes both correction precision and tuning noise, so the selected scalar weight need not vary monotonically at every budget. Fixed-test intervals condition on the learned policy and do not include retraining or analyst-selection uncertainty. The Gaussian synthetic experiments diagnose the normalized target in value space and are not intended to reproduce the bounded-outcome causal data-generating process. DR-DIRECT removes direct dependence on the preliminary softmax target but still uses preliminary predictions inside its DR nuisance. The margin audit is indexed by the true reference margin and should be read as a mechanism diagnostic rather than as a deployable margin estimator.

Comparator and outcome-scale boundaries. DR-DIRECT uses the same correction-split DR values, whose outcome nuisance uses preliminary predictions. It removes direct reliance on the preliminary softmax target, not all preliminary information. The Gaussian synthetic value experiment can produce values outside $[0, 1]$; it diagnoses the smooth target map in value space, rather than simulating the bounded-outcome identification model. The minimax lower bound already explicitly uses a separate Gaussian experiment. The error audit bins by the true softmax top-two margin, not an estimated deployable margin. Roleplay's negative held-out result is consistent with weak signal but does not identify all true treatment contrasts.

J Complete Publication Panels and Diagnostics

The following figures route supporting evidence out of the nine-page main text. Every empirical mark is computed from the supplied result tables; panel-level sources, seeds, and uncertainty definitions are recorded in the artifact's `panel_provenance.json`. Each panel is rendered independently before vector composition. The Figure 3 illustration uses a Helmert U, $\hat{v} = (.4, .5, .6, .45)^\top$, $d_c = (.08, -.06, .10)^\top$, $\Phi = \hat{v} + Ud_c$, and $\lambda = .2$, with $A = 0, .5I, \mathrm{diag}(.2, .8, .5), I$. Analytic figures in the main text are explicitly distinguished from these experiments.

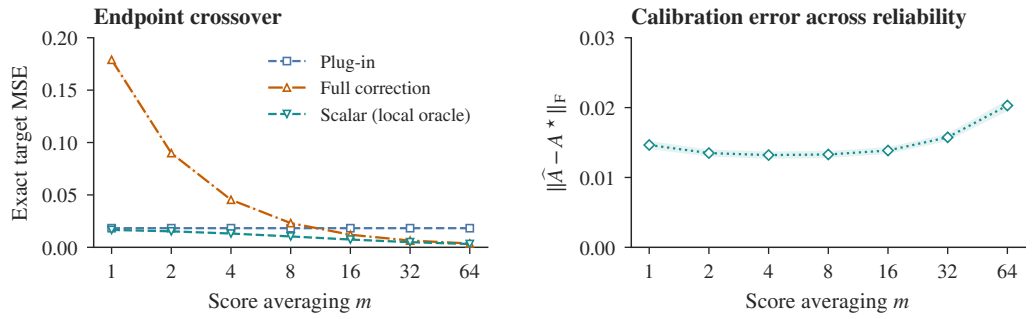

Figure 8: Endpoint crossover and calibration error across score precision. The left panel compares the two endpoints and the local-risk scalar rule. The right panel varies score averaging, *not* calibration size; its nonmonotone operator error is retained. Bands are mean ± 1.96 Monte Carlo standard errors across 96 seeds.

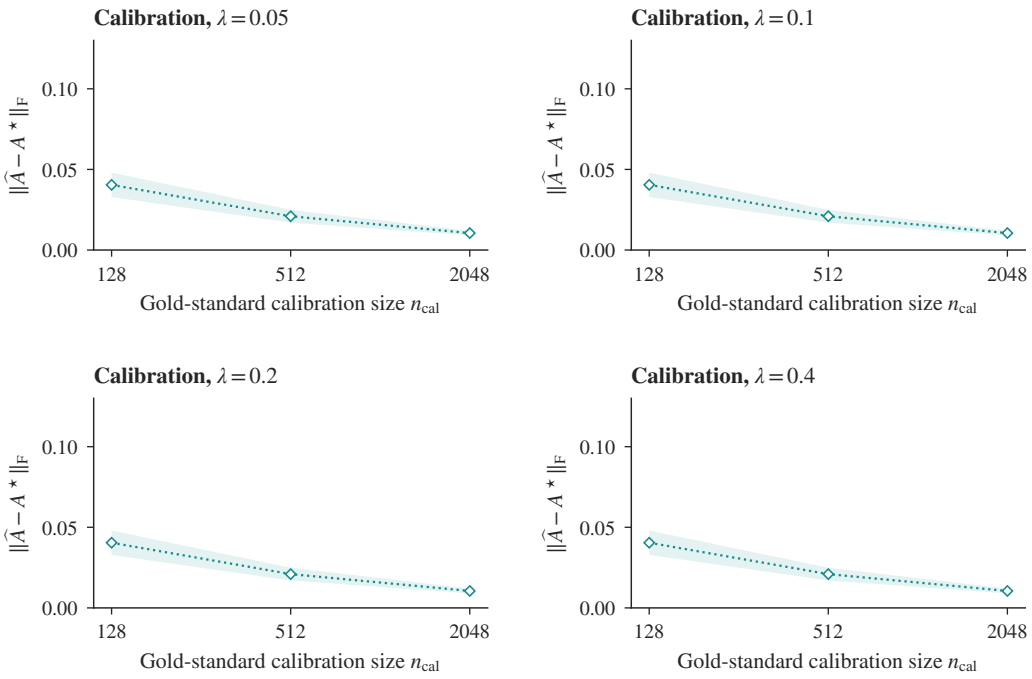

Figure 9: All supplied calibration-size curves, at all four temperatures. The calibration source observes total errors relative to known truth. Bands are descriptive normal-approximation mean ± 1.96 standard errors across eight seeds. They are not a statement of real-world total-error identifiability.

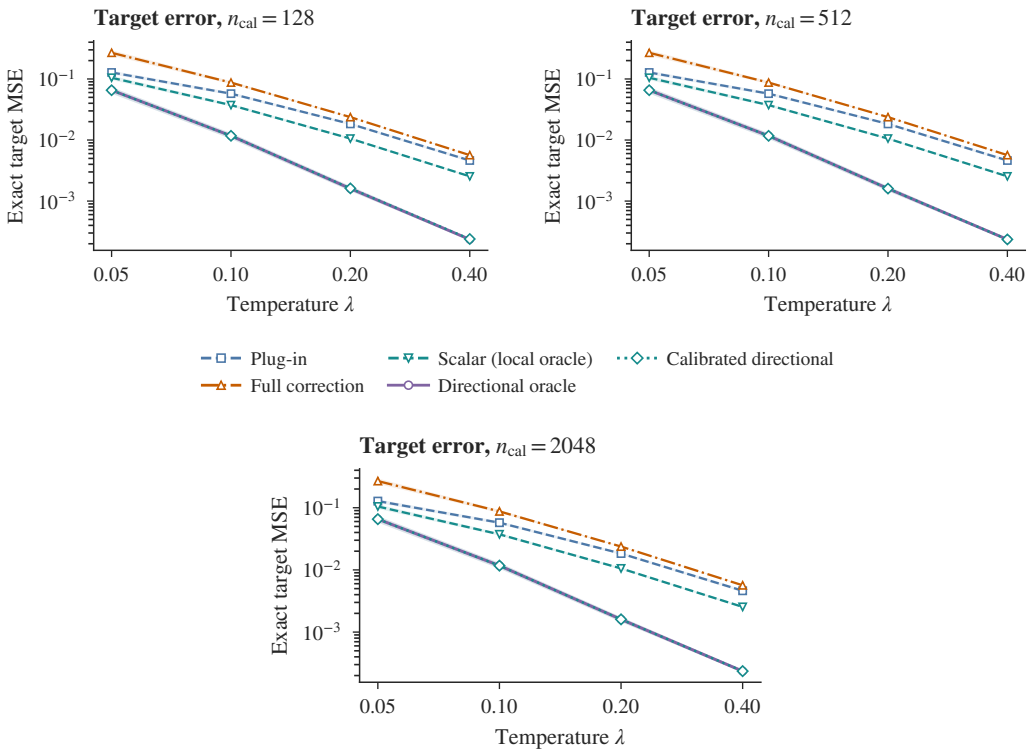

Figure 10: The complete stored temperature-by-calibration grid: eight seeds, four temperatures, and three calibration sizes. Each method is the same mathematical restriction as in Figure 4. Lines connect only the supplied grid values; bands are mean ± 1.96 Monte Carlo standard errors.

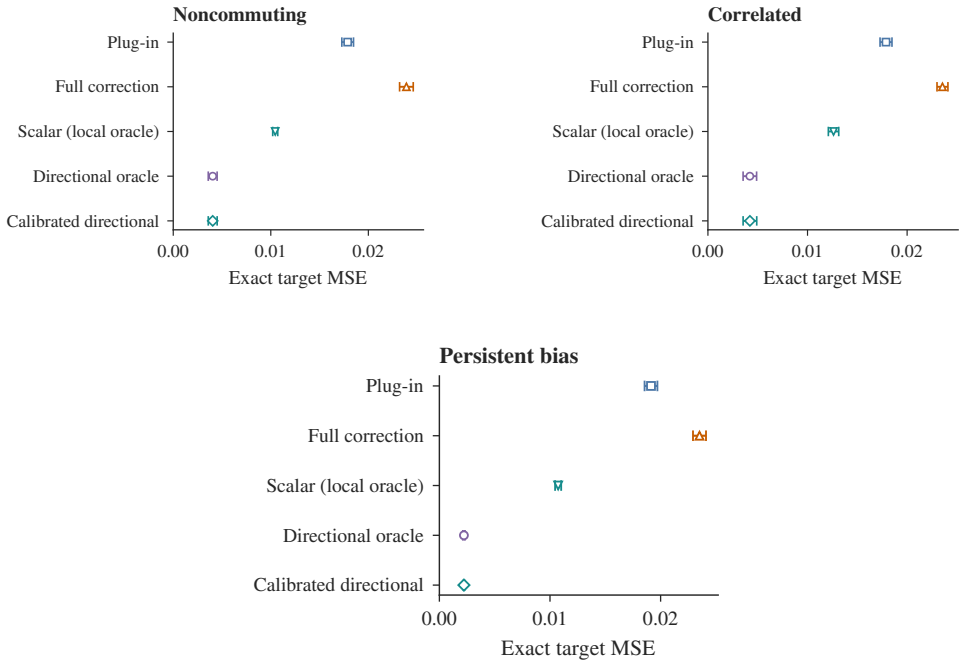

Figure 11: All supplied targeted robustness regimes, 24 seeds each. Error bars are mean ± 1.96 Monte Carlo standard errors. For the correlated-source setting, the scalar comparator uses the corresponding cross-moment optimum from Proposition 14, rather than the zero-cross-moment scalar rule.

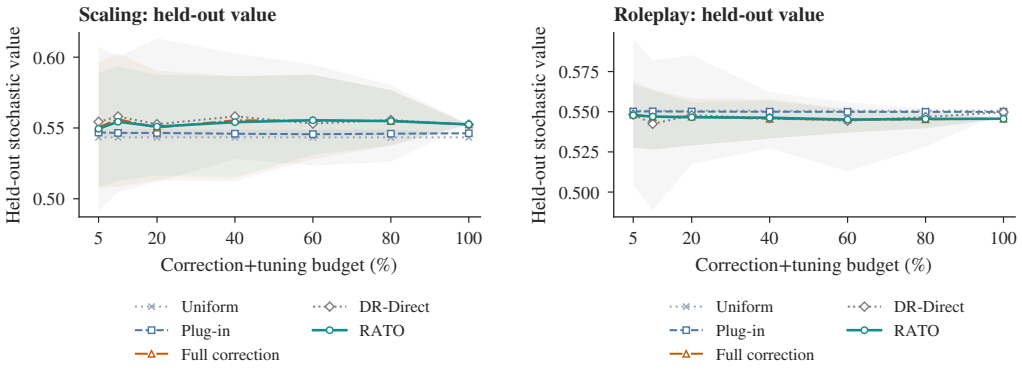

Figure 12: Complete matched-information stochastic-value curves for both randomized studies. Every method uses the same available split information. Shading denotes 2.5–97.5% across-thinning quantiles, not fixed-test confidence intervals. The preliminary predictor is fixed, but its temperature can change with tuning data. Full-budget repetitions are identical.

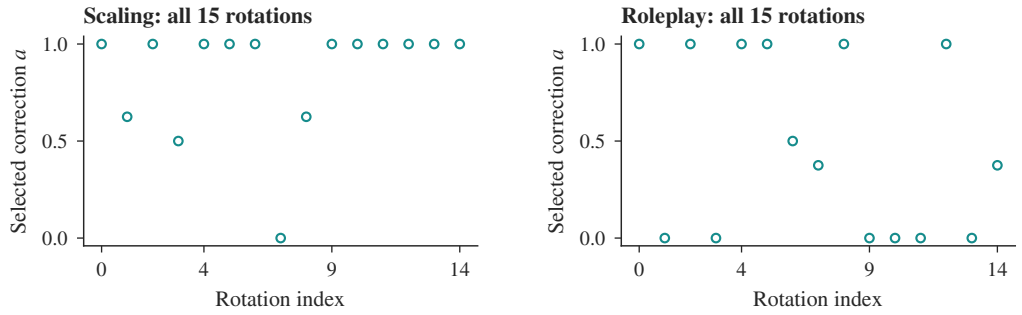

Figure 13: All 15 participant-fold rotations in each study. These are descriptive sensitivity points, not independent replications or confidence intervals. They map instability of the selected correction without changing the main fixed-test values.

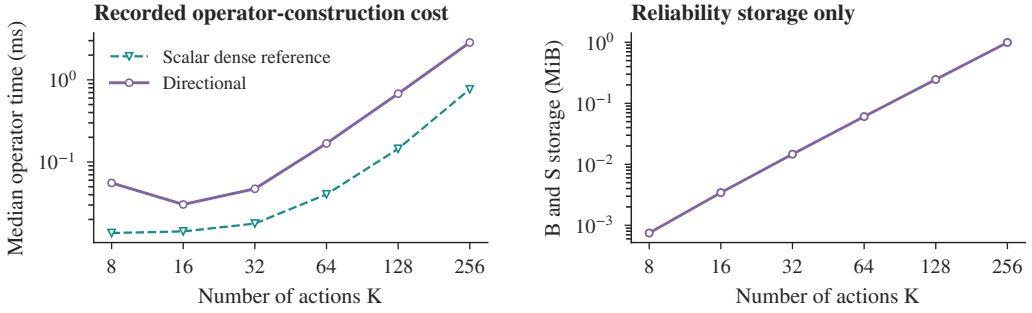

Figure 14: Recorded dense operator-construction medians and deterministic reliability-matrix storage. The original timing values are not remeasured on this machine. The scalar reference uses dense matrix operations; the storage panel counts only B, S, not peak process memory.

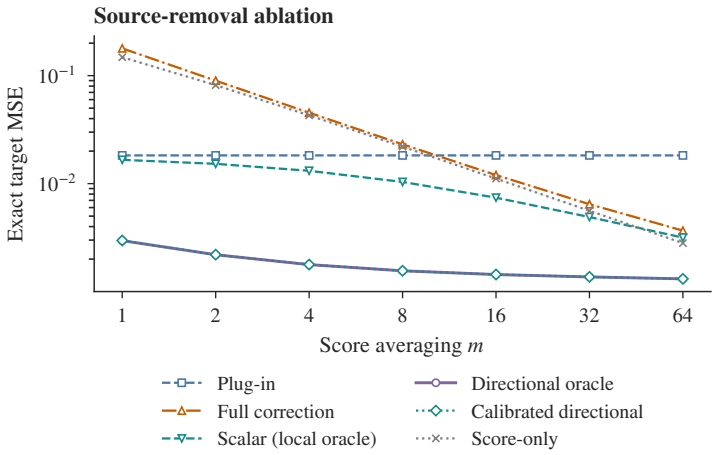

Figure 15: Source-removal ablation on the full stored reliability grid. Score-only softmax uses only the simulated causal value source; unlike the real-data DR-DIRECT nuisance construction, this synthetic control removes preliminary information. Other curves are the same target restrictions as in Figure 4. Bands use 96 seed-level means and the same Monte Carlo convention.