FAIR AGENTS CAN FORM UNFAIR SYSTEMS: SHARP LIMITS OF FAIRNESS COMPOSITION

Anonymous authors
Paper under double-blind review

ABSTRACT

Demographic parity of individual agents does not imply demographic parity of their joint decision. We characterize the largest system disparity compatible with local rate audits and the utility cost of controlling it. By marginal duality, this disparity equals a regularized affine-approximation error, which yields sharp, attained bounds for every voting threshold: under exact local parity, every threshold rule over at least two agents admits a gap of at least $1/2$. Conversely, any monotone coordinator with worst-case gap at most ρ lies within $\rho/(1+\rho)$ of an agent–constant mixture that meets the same budget. This bounds the utility advantage of nonlinear coordination, and under an explicit tight-budget condition a mixture is exactly optimal. Matching all pairwise moments does not reduce odd majority's worst-case gap. Experiments separate fairness on the current distribution from robustness to dependence. In a three-LLM committee, the observed majority gap is 0.0417 and the largest observed individual gap is 0.0031. At larger budgets, observed accuracy differences between full coordination and fitted mixtures range from 0.74 to 1.87 percentage points in explicit-cue LLM comparisons. Public-data differences are about 0.1–0.2 percentage points, while observed and worst-case disparities differ substantially. Together, the results quantify what local fairness audits leave uncontrolled and how much predictive flexibility must be given up to control it.

1 INTRODUCTION

A decision system can certify each of its agents for demographic parity and still produce unequal outcomes across groups. A local audit records how often each agent approves members of each group. Majority, unanimity, and veto rules, however, also depend on *whether the agents approve the same subjects*. Equal approval rates therefore leave an essential part of the joint decision unspecified.

Figure 1 gives a minimal example. Three agents each approve one third of both groups: in group 0 they approve the same third, and in group 1 they approve disjoint thirds. An OR coordinator then approves one third of group 0 and all of group 1, a system gap of $2/3$ despite exact local parity. Selecting one agent uniformly at random instead preserves the common approval rate of $1/3$. No agent needs the protected attribute: group differences in otherwise observed evidence suffice.

Composition failures are well established in algorithmic fairness (Dwork & Ilvento, 2019; Wang et al., 2021; Gohar et al., 2023), and recent studies report bias amplification in LLM collectives (Li et al., 2026; Okawa, 2026). What has not been quantified is how large the system disparity can be when every local audit passes, and how much coordination capacity is lost when that disparity must be controlled for every compatible joint law. We answer both questions for a coordinator $h : \{0,1\}^m \to [0,1]$ that maps the agents' binary actions to the probability of a positive decision. Its worst-case gap $\mathcal{V}_\varepsilon(h)$ is the largest group-rate gap over all pairs of joint action laws whose per-agent rate differences are bounded by ε. Our contributions are as follows.

- **Exact characterization.** By marginal duality, $\mathcal{V}_\varepsilon(h)$ equals a regularized uniform approximation error: the distance from h to affine functions measures sensitivity to unaudited dependence, and a weighted slope penalty accounts for the permitted local disparities (Theorem 1). The same characterization yields a linear program for fitting robust coordinators.
- **Sharp limits.** We derive exact, attained disparity curves for every voting threshold (Theorem 2); a same-budget approximation of robust monotone coordinators by agent–constant mixtures, with an

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.

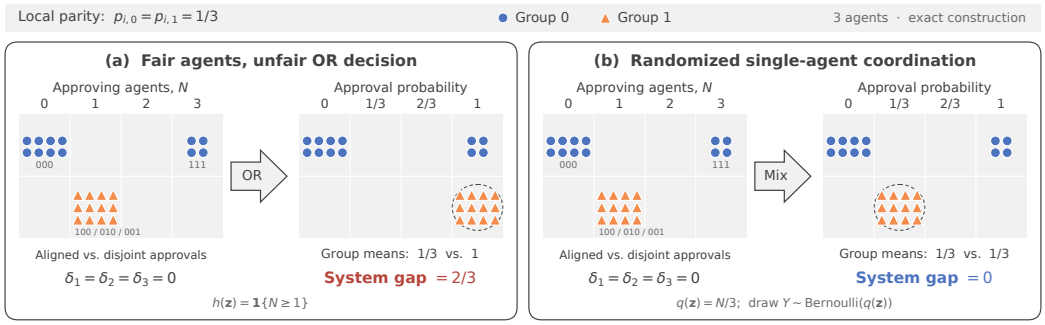

Figure 1: **Equal local approval rates can yield unequal system outcomes.** Each of three agents approves one third of both groups. Group 0 has coincident approvals, $P_0(000) = 2/3$ and $P_0(111) = 1/3$; group 1 has disjoint approvals, $P_1(100) = P_1(010) = P_1(001) = 1/3$. **(a)** OR yields approval rates $1/3$ and 1, attaining the sharp gap $2/3$. **(b)** Selecting one agent uniformly at random, i.e., using $q(\mathbf{z}) = (z_1 + z_2 + z_3)/3$ as the probability of a positive decision, yields approval rate $1/3$ in both groups. Each marker represents probability mass $1/12$.

exact tight-budget optimum (Corollary 2, Proposition 3); and an obstruction showing that pairwise audits do not improve odd majority's worst-case gap (Proposition 1).

- **Empirical evidence.** We compare fitted coordinators with fairness post-processors and examine communication and committee composition in LLM committees. The three-agent study exhibits a larger observed system gap than any individual gap; the reported accuracy differences describe fitted rules on the evaluation cases.

2 Related Work

Fairness under composition. Dwork & Ilvento (2019) established composition failures for individual and group fairness; later work studies pipelines, recommender systems, and ensembles (Bower et al., 2017; Dwork et al., 2020; Wang et al., 2021; Gohar et al., 2023; Schweighofer et al., 2025). We instead quantify the exact worst-case group-rate gap for a single subject's binary action vector under local rate constraints.

Marginal problems and approximation. Auditing at fixed rates is a two-sided Fréchet marginal problem; for OR, it recovers the classical union bounds. Marginal duality and moment-constrained optimization supply the underlying tools (Rüschendorf, 1991; Bertsimas & Popescu, 2005), and correlation-robust optimization compares worst-case costs under fixed marginals with those under independent sampling (Agrawal et al., 2010). In our setting, cross-group marginal differences constrain two laws, and exact parity links their disparity to degree-one uniform approximation (Nisan & Szegedy, 1994; O'Donnell, 2014). Beyond these tools, our contributions are the attained threshold/tolerance curves, the same-budget mixture restriction and its utility consequences, and the pairwise-audit obstruction for odd majority.

Agent collectives and fair prediction. Multi-agent debate can improve reasoning (Du et al., 2024; Liang et al., 2024) but can also amplify bias (Li et al., 2026; Okawa, 2026). Our paired binary tasks compare system disparity with the largest local disparity on common evidence. Recent post-processors rely on bias scores, discrepancy penalties, or linear decision rules (Chen et al., 2024; Tifrea et al., 2024; Xian & Zhao, 2024), and robust classification constrains fairness over uncertainty sets (Taskesen et al., 2020). Our uncertainty set contains every pair of joint action laws that satisfies the local-rate bounds, which directly exposes the cost of protection against dependence.

3 Setting and Sharp Composition Bounds

Let $A \in \{0,1\}$ denote the group, $\mathbf{Z} \in \{0,1\}^m$ the agents' actions, and $P_a = \mathcal{L}(\mathbf{Z} \mid A = a)$. Write $p_{i,a} = \mathbb{E}[Z_i \mid A = a]$ and $\delta_i = |p_{i,1} - p_{i,0}|$. A fixed group-blind coordinator $h : \{0,1\}^m \to [0,1]$

outputs $\widehat{Y} \sim \mathrm{Bernoulli}(h(\mathbf{Z}))$; its signed disparity and gap are

$$D_h(P_0, P_1) = \mathbb{E}_{P_1} h - \mathbb{E}_{P_0} h, \qquad \Delta_h = |D_h(P_0, P_1)|. \tag{1}$$

We study demographic parity for binary actions and single-subject coordination. Affine results apply when the affine score itself is the outcome probability; thresholding it generally yields a nonlinear coordinator.

Definition 1 (Worst-case composition gap)**.** For $\boldsymbol{\varepsilon} \in [0,1]^m$, let

$$\mathcal{V}_{\boldsymbol{\varepsilon}}(h) = \sup_{P_0, P_1 \colon |\mathbb{E}_{P_1} Z_i - \mathbb{E}_{P_0} Z_i| \le \varepsilon_i \ \forall i} |\mathbb{E}_{P_1} h - \mathbb{E}_{P_0} h|. \tag{2}$$

The joint laws are otherwise unrestricted; when $\boldsymbol{\varepsilon} = 0$, their common marginal vector is also optimized over.

Theorem 1 (Exact affine approximation characterization)**.** *For any $h : \{0,1\}^m \to [0,1]$ and $\boldsymbol{\varepsilon} \in [0,1]^m$,*

$$\begin{aligned} \mathcal{V}_{\boldsymbol{\varepsilon}}(h) &= \min_{b \in \mathbb{R}^m} \left\{ \mathrm{osc}_{\mathbf{z}}(h(\mathbf{z}) - b^\top \mathbf{z}) + \sum_i \varepsilon_i |b_i| \right\} \\ &= \min_{a,b} \left\{ 2\|h - a - b^\top \mathbf{z}\|_\infty + \sum_i \varepsilon_i |b_i| \right\}, \end{aligned} \tag{3}$$

where $\mathrm{osc}\, f = \max f - \min f$. Both extrema are attained. In particular,

$$\mathcal{V}_0(h) = 2 \inf_{\ell \in \mathrm{Aff}} \|h - \ell\|_\infty, \qquad \mathcal{V}_0(h) = 0 \iff h \in \mathrm{Aff}. \tag{4}$$

The approximating affine function need not take values in $[0,1]$, although the coordinator h must.

Proof sketch. Because the feasible set is invariant under exchanging the groups, the absolute value can be dropped. Dualizing the two probability normalizations and the $2m$ marginal inequalities yields an upper and a lower envelope of $h - b^\top \mathbf{z}$ together with the penalty $\sum_i \varepsilon_i |b_i|$. Strong LP duality gives equality, and centering the two envelopes gives twice the uniform approximation error. Appendix A gives the complete derivation.

Corollary 1 (An implementable robustness certificate)**.** *If $|g(\mathbf{z}) - a - b^\top \mathbf{z}| \le t$ for all $\mathbf{z}$ and $2t + \sum_i \varepsilon_i |b_i| \le \rho$, then $\Delta_g \le \rho$ for every pair of laws with $\delta_i \le \varepsilon_i$. Conversely, every g with $\mathcal{V}_{\boldsymbol{\varepsilon}}(g) \le \rho$ has such a certificate.*

The guarantee covers arbitrary changes in dependence and in common marginal rates, provided the local tolerances continue to hold; this larger uncertainty set can cost more utility than a constraint on one fitted pair of laws. For fixed marginal levels, the exact gap instead optimizes separately over the two Fréchet classes; Appendices A and L.4 give fixed-report and report-interval formulations.

Theorem 2 (Every threshold: an exact tolerance curve)**.** *For $h_k(\mathbf{z}) = \mathbf{1}\{\sum_i z_i \ge k\}$, $1 \le k \le m$, set $\beta_k = 1/\max\{k, m-k+1\}$. For every $\varepsilon \in [0,1]$,*

$$\boxed{\mathcal{V}_{\varepsilon \mathbf{1}}(h_k) = \min\{1,\ 1 - \beta_k + m\varepsilon\beta_k\}.} \tag{5}$$

In particular,

$$\mathcal{V}_0(h_k) = \max\left\{ \frac{k-1}{k}, \frac{m-k}{m-k+1} \right\}. \tag{6}$$

Every value is attained by a pair of laws realizable with deterministic group-blind agents. OR and AND have $\mathcal{V}_0 = 1 - 1/m$; odd-m majority has $\mathcal{V}_0 = (m-1)/(m+1)$. Every threshold reaches maximal vulnerability 1 when $\varepsilon \ge 1/m$.

Attainment. Permutation averaging reduces the dual to $b = \beta \mathbf{1}$. For $\beta \ge 0$, the oscillation of $\mathbf{1}\{N \ge k\} - \beta N$ equals $1 - \beta$ on $[0, \beta_k]$ and is nondecreasing thereafter; adding $m\varepsilon\beta$ and minimizing selects β_k or zero. Mixtures of uniform Hamming layers attain the bound at every tolerance (Appendix B). For OR at zero tolerance, $P_0 = m^{-1}\delta_{\mathbf{1}} + (1 - m^{-1})\delta_{\mathbf{0}}$ and the uniform law P_1 on one-hot vectors share the rates $p_i = 1/m$ but have system rates $1/m$ and 1. Group-blind agents realize these laws: draw $X \mid A = a \sim P_a$ and set $Z_i = X_i$. Removing the protected attribute from the inputs therefore leaves the obstruction intact.

Corollary 2 (Strong robustness restricts coordination). *Every threshold h_k with $m \geq 2$ has $\mathcal{V}_0(h_k) \geq 1/2$. For any monotone $g : \{0,1\}^m \to [0,1]$ with $\mathcal{V}_{\boldsymbol{\varepsilon}}(g) \leq \rho$, there is a convex mixture q of constants and single-agent rules satisfying the same budget $\mathcal{V}_{\boldsymbol{\varepsilon}}(q) \leq \rho$ and*

$$\|g - q\|_\infty \leq \frac{\rho}{1+\rho}. \tag{7}$$

At $\rho = 0$, such mixtures are exactly the monotone rules immune to arbitrary dependence under exact local parity.

For monotone g, the affine envelope $\ell = a + b^\top \mathbf{z}$ can be chosen with $b \geq 0$: replacing a negative slope by zero and recentering the intercept does not increase the uniform error. If $\|g - \ell\|_\infty \leq t$, then

$$q = \frac{\ell + t}{1+2t}, \qquad \|g - q\|_\infty \leq \frac{2t}{1+2t}, \qquad \mathcal{V}_{\boldsymbol{\varepsilon}}(q) = \frac{\sum_i \varepsilon_i b_i}{1+2t} \leq \rho. \tag{8}$$

Thus q is a valid agent–constant mixture, and $2t \leq \rho$ yields the stated distance within the same budget.

Additional joint-moment audits. Let $\phi_j(\mathbf{z})$ be audited joint features whose cross-group expectation differences are bounded by η_j. Replacing the coordinate features by ϕ gives the same exact dual:

$$\mathcal{V}_\eta^\phi(h) = \min_{a,\theta} \left\{ 2\|h - a - \theta^\top \phi\|_\infty + \sum_j \eta_j |\theta_j| \right\}. \tag{9}$$

Products of actions encode dependence constraints within an LP; with exact matching through degree d, the gap is twice the degree-d uniform approximation error. The value of additional moments depends on the coordinator: exactly matching first and second moments reduces five-agent OR's worst-case gap from 0.800 to 0.444, whereas the following obstruction holds for every odd majority.

Proposition 1 (Pairwise matching does not improve odd majority). *If $h(\mathbf{1} - \mathbf{z}) = 1 - h(\mathbf{z})$, the best degree-two and degree-one uniform approximation errors coincide. In particular, exact matching of local rates and all pairwise moments leaves every odd-majority worst-case gap at $(m-1)/(m+1)$.*

Complement symmetrization removes the quadratic part of an approximant without increasing its error. Appendix L gives the proof, the full moment hierarchy, and the distinction between raw-moment and covariance constraints.

4 Computing the Cost of Robust Coordination

With the agents fixed, repair selects an outcome probability $g : \{0,1\}^m \to [0,1]$. Let $\hat{P}$ be the pooled fitted action law and $\bar{u}_y(\mathbf{z})$ the estimated utility of outcome y. We maximize

$$\hat{U}(g) - \gamma \hat{C}(g) = \sum_{\mathbf{z}} \hat{P}(\mathbf{z}) \left[g(\mathbf{z}) \bar{u}_1(\mathbf{z}) + (1 - g(\mathbf{z})) \bar{u}_0(\mathbf{z}) - \gamma |g(\mathbf{z}) - h(\mathbf{z})| \right]. \tag{10}$$

For deterministic h, the objective is linear in g; randomized h requires absolute-value slack variables. The fitted table is an outcome probability, not a score to be thresholded: the deployed decision draws $\mathrm{Bernoulli}(g(\mathbf{z}))$.

Empirical repair. The standard finite-state fairness constraint

$$|\langle g, \hat{P}_1 - \hat{P}_0 \rangle| \leq \tau \tag{11}$$

controls disparity on the fitted law. We also evaluate its Lagrangian-penalty counterpart over the same function class.

Robust JointRepair. Choose a deployment tolerance vector $\boldsymbol{\varepsilon}$ and a system budget ρ. Introduce auxiliary variables a, b, t, v and require

$$\begin{aligned} &0 \leq g(\mathbf{z}) \leq 1, \quad -t \leq g(\mathbf{z}) - a - b^\top \mathbf{z} \leq t \quad \text{for every } \mathbf{z}, \\ &-v_i \leq b_i \leq v_i, \quad t, v_i \geq 0, \quad 2t + \sum_i \varepsilon_i v_i \leq \rho. \end{aligned} \tag{12}$$

Proposition 2 (Exact robust repair and its scope)**.** *Maximizing (10) subject to (12), and optionally (11), is an LP whose feasible coordinators are exactly those satisfying $\mathcal{V}_{\boldsymbol{\varepsilon}}(g) \le \rho$ (and, if imposed, the empirical constraint). Every returned g has $\Delta_g \le \rho$ on any deployment pair with $\delta_i \le \varepsilon_i$, whatever its joint dependence. The full table has $2^m + O(m)$ variables and $O(2^m + m)$ inequalities. Under uniform tolerance, count-only rules $g(\mathbf{z}) = q_{\sum_i z_i}$ admit an equivalent LP with $m+1$ rule variables and an affine envelope $a + b\sum_i z_i$.*

The experiments impose monotonicity through $g(\mathbf{z}) \le g(\mathbf{z} + e_i)$ whenever $z_i = 0$, adding $m2^{m-1}$ edge constraints; all action cells, including unseen ones, remain constrained. For count-only rules, permutation averaging reduces the envelope to $|q_j - a - bj| \le t$ for $j = 0, \dots, m$, with $2t + m\varepsilon|b| \le \rho$ and $q_j \le q_{j+1}$. This reduction is exact within the count-only class; matching the full optimum additionally requires a permutation-invariant objective. Constants are always feasible, but the structural constraint controls fairness only and does not preserve utility under distribution shift.

A matched mixture baseline. The baseline fits $m+3$ nonnegative weights with $\theta_h + \sum_i \theta_i + \theta_+ + \theta_- = 1$:

$$g_\theta(\mathbf{z}) = \theta_h h(\mathbf{z}) + \sum_{i=1}^{m} \theta_i z_i + \theta_+, \tag{13}$$

where θ_+ and θ_- weight the constants 1 and 0. Imposing Equation (12) on g_θ enforces the same budget as full repair, and both are fitted with the same objective on the same cases. For a fixed distribution and linear utility with $|u_1 - u_0| \le 1$, let U^*_{full} and U^*_{mix} denote the two class optima. Since the mixture class contains every monotone affine rule, including the feasible approximant q, Corollary 2 gives

$$0 \le U^*_{\text{full}} - U^*_{\text{mix}} \le \frac{\rho}{1+\rho}. \tag{14}$$

This bound compares optima on the same distribution under the structural constraint alone; it does not bound the test difference between independently fitted rules. Write $\mathcal{G}_{\boldsymbol{\varepsilon},\rho}$ for the monotone, probability-valued rules with $\mathcal{V}_{\boldsymbol{\varepsilon}}(g) \le \rho$. The next result makes the utility consequence exact under a tight budget.

Proposition 3 (An affine optimum under a tight budget)**.** *For ordinary accuracy on a fixed joint law of (Z, Y) with reference label Y, write $C = \max\{\Pr(Y=0), \Pr(Y=1)\}$ and $A_{\max} = \max_i \Pr(Z_i = Y)$. For uniform tolerance $\varepsilon > 0$, every monotone g with $\mathcal{V}_{\varepsilon\mathbf{1}}(g) \le \rho$ satisfies*

$$U(g) \le C + \rho \max\left\{C, \frac{A_{\max} - C}{\varepsilon}\right\}. \tag{15}$$

If $0 \le \rho \le \varepsilon$ and $A_{\max} \ge (1+\varepsilon)C$, the full monotone robust optimum equals

$$\max_{g \in \mathcal{G}_{\varepsilon\mathbf{1},\rho}} U(g) = C + \frac{\rho}{\varepsilon}(A_{\max} - C). \tag{16}$$

It is attained by mixing a most accurate agent, with weight ρ/ε, and a most accurate constant.

Proof sketch. A nonnegative-slope envelope with error t and total slope B gives $U(g) \le C + 2Ct + B(A_{\max} - C)$; maximizing under $2t + \varepsilon B \le \rho$ gives the bound, which the stated mixture attains. For $0 \le \rho \le \varepsilon$ and $A_{\max} > (1+\varepsilon)C$, every exact optimum is affine, and at $\rho = \varepsilon$ a unique best agent j forces $g(z) = z_j$ on the entire cube. Applied to a common fit law, this implies that the fitted rules are equal everywhere, including on test inputs, so train/test shift alone cannot separate them. Appendix L.2 proves these equality conditions.

Auditing observed dependence. Worst-case vulnerability does not identify the source of a particular observed gap. For paired case c and cue a, let $p_{c,a}$ be the vector of conditional agent rates, and let $\bar{h}(p) = \mathbb{E}_{\otimes_i \text{Bern}(p_i)} h$ be the independent-product reference. Writing $D_{\text{obs}} = D_h(P_0, P_1)$, the common case distribution gives

$$D_{\text{marg}} = \mathbb{E}_c[\bar{h}(p_{c,1}) - \bar{h}(p_{c,0})], \qquad D_{\text{dep}} = D_{\text{obs}} - D_{\text{marg}}. \tag{17}$$

The experiments condition on the case; conditioning at a different level reallocates the gap between the two channels (Appendix C). For $R \ge m$, let $\mathcal{I}_{R,m}$ be the ordered tuples of distinct run indices

and define

$$\hat{M}_{c,a} = \frac{1}{|\mathcal{I}_{R,m}|} \sum_{\mathbf{r} \in \mathcal{I}_{R,m}} h(Z_{1,c,a}^{(r_1)}, \ldots, Z_{m,c,a}^{(r_m)}), \tag{18}$$

$$\hat{O}_{c,a} = \frac{1}{R} \sum_{r} h(\mathbf{Z}_{c,a}^{(r)}), \qquad \widehat{D_{\text{marg}}} = \frac{1}{n} \sum_{c} (\hat{M}_{c,1} - \hat{M}_{c,0}),$$

$$\widehat{D_{\text{dep}}} = \frac{1}{n} \sum_{c} [(\hat{O}_{c,1} - \hat{M}_{c,1}) - (\hat{O}_{c,0} - \hat{M}_{c,0})].$$

With distinct run indices, the recombined coordinates are independent conditional on (c, a), so $\hat{M}_{c,a}$ is unbiased for $\bar{h}(p_{c,a})$ when runs are conditionally i.i.d. Reusing a run introduces a dependence bias that need not vanish as the number of cases grows at fixed R. Inference resamples whole cases, retaining both cues and all runs. The residual is a statistical comparison with a product law, not a causal effect of communication: the product law need not be achievable by intervening on messages.

5 Experiments

We first compare coordination methods and then examine the effects of communication and committee composition. Throughout, *accuracy* is ordinary task accuracy, *gap* is an absolute group-rate difference, and $\mathcal{V}_{.05}$ is the worst-case gap over laws with individual disparities at most 0.05. These quantities answer different questions: a small observed gap need not imply robustness to changes in dependence.

5.1 Experimental Setup

Tasks and splits. The three-agent study runs Qwen2.5-7B-Instruct, Llama-3.1-8B-Instruct, and Mistral-7B-Instruct-v0.3 on generated hiring, credit, admissions, and tenancy cases. Each agent sees one score field, and the label is positive when at least two of three criteria pass. Female/male variants share evidence and labels. There are 1,024 fit and 4,096 evaluation pairs, eight shared wording families per task, and eight runs per variant at temperature one. The five-agent study adds Gemma-2-9B-IT and Phi-3.5-mini-instruct (Google, 2024; Microsoft, 2024) and uses full narratives: 1,024 fit and 2,048 evaluation pairs per cue condition, with eight fit and sixteen disjoint evaluation wording families. On Adult and German Credit (Becker & Kohavi, 1996; Hofmann, 1994), three logistic agents use disjoint feature blocks, and group-blind local calibration fits monotone bin probabilities at DP tolerance 0.02. Group labels are used only for offline fitting and evaluation. Appendices H–I detail the protocols and record scope.

Comparators. All coordination methods use frozen agents. *Majority* is the original voting rule (Dietterich, 2000); *best single* and *constant* select an agent or a constant by fit accuracy; *random agent* selects an agent uniformly at random. The *fitted mixture* (Equation (13)) optimizes weights on majority, all single agents, and both constants. Under the matched protocol, both classes maximize pooled fit accuracy under the same (ε, ρ) constraint, with $\gamma = 0$ and no additional empirical-gap constraint; tight-budget analytic benchmarks are reported separately from fitted comparisons. Empirical LPs, adapted from Agarwal et al. (2018), instead constrain fitted disparity at $\tau \in \{0.02, 0.05\}$. On public data we also compare MBS, FRAPPÉ-KL, and LinearPost (Chen et al., 2024; Tifrea et al., 2024; Xian & Zhao, 2024), adapted to the same three-action input: MBS thresholds estimated bias scores, FRAPPÉ-KL penalizes soft disparity with a KL discrepancy before thresholding, and LinearPost solves a randomized LP using predicted-group scores. All three use shared smoothed fit-set label/group estimates and receive only the action vector at inference. Their current-distribution constraints differ from Equation (2), and predicted-group feasibility need not imply actual-group feasibility.

	Adult			German Credit		
Coordinator	Acc.↑	Gap↓	$\mathcal{V}_{.05}$ ↓	Acc.↑	Gap↓	$\mathcal{V}_{.05}$ ↓
Calibrated majority (Dietterich, 2000)	0.766	0.039	0.575	0.834	0.059	0.575
MBS (Chen et al., 2024)	0.847	0.009	1.000	0.827	0.011	0.550
FRAPPÉ-KL (Tifrea et al., 2024)	0.799	0.017	1.000	0.819	0.004	0.717
LinearPost (Xian & Zhao, 2024)	0.864	0.020	0.803	0.836	0.020	0.550
Fitted mixture (§ 5.1)	0.864	0.030	0.050	0.793	0.043	0.100
JOINTREPAIR (ours)	0.865	0.030	0.100	0.795	0.043	0.100

Table 1: **Accuracy, observed disparity, and structural robustness.** Acc.: predictive accuracy; Gap: observed demographic-parity gap; $\mathcal{V}_{.05}$: worst-case composition gap under local tolerances of 0.05. All coordinators use the same three-agent input. The post-processors use disparity tolerance $\tau = 0.02$; the mixture and full repair target $(\varepsilon, \rho) = (0.05, 0.10)$. Accuracy and observed disparity integrate response randomization; structural values come from the worst-case composition LP.

Reporting and uncertainty. Each rule is frozen before evaluation. Current comparisons report observed differences; uncertainty estimation is specified separately in Appendix I.1. Given case-level records, for an independent case-block difference $X_c \in [-1, 1]$ between cues or rules, with sample variance S_X^2 (denominator $n-1$), K candidates, and family error α_f, the empirical-Bernstein radius is (Maurer & Pontil, 2009)

$$r = \sqrt{\frac{2S_X^2 \log(4K/\alpha_f)}{n}} + \frac{14 \log(4K/\alpha_f)}{3(n-1)}. \tag{19}$$

The target is $n^{-1} \sum_c \mathbb{E}X_c$, including independent cases under a fixed template schedule. For an absolute gap, the corresponding interval is $[L, U] = [\max(0, |\bar{X}| - r), \min(1, |\bar{X}| + r)]$; for a signed accuracy loss, non-inferiority at margin η requires $\bar{X} + r \leq \eta$. Applying this construction requires the case-level variances and a declared comparison family. We report the current public-data and LLM comparisons descriptively, without confidence endpoints or a population separation certificate. Appendix H specifies the available records.

5.2 COMPARISON WITH FAIRNESS POST-PROCESSORS

Table 1 reports predictive accuracy, observed demographic-parity gap, and worst-case composition gap. On Adult, JOINTREPAIR reaches accuracy 0.865, versus 0.864 for both the fitted mixture and LinearPost, a gain of about 0.1 percentage points (pp). Its observed gap (0.030) exceeds those of the post-processors (0.009–0.020), but its worst-case gap is 0.100, compared with 0.803 for LinearPost and 1.000 for MBS and FRAPPÉ-KL. The fitted mixture attains an even smaller worst-case gap of 0.050; the small accuracy gain from full coordination thus uses more of the structural budget.

On German Credit, full repair raises accuracy from 0.793 to 0.795 relative to the fitted mixture (about 0.2 pp). Both attain the structural budget $\mathcal{V}_{.05} = 0.100$, and both observed gaps round to 0.043. LinearPost achieves the highest accuracy, 0.836, with an observed gap of 0.020 but a worst-case gap of 0.550; full repair thus trades 4.1 pp of accuracy for a much smaller worst-case gap.

Across both datasets, the observed accuracy differences between full coordination and fitted mixtures are about 0.1 and 0.2 pp, while post-processing attains smaller observed disparities. Observed and worst-case gaps diverge substantially: a small disparity on the evaluation distribution can coexist with high sensitivity to changes in the joint action law. The structural guarantee controls this sensitivity as long as the agents' local gaps remain within the audited tolerances.

5.3 LLM COMPARISONS ACROSS ROBUSTNESS BUDGETS

The three-agent comparison uses the mixed committee with two rounds of refreshed peer votes and rationales (V2; Section 5.4). Table 2 compares full coordination with the fitted mixture at $\rho = 0.10$ and 0.20, with $\varepsilon = 0.05$. Tighter budgets are examined through the analytic benchmarks below. At $\rho = 0.10$ and 0.20, the observed full-minus-mixture accuracy differences are 0.74 and 0.77 pp,

ρ	Mixture accuracy	Full accuracy	Observed difference (pp)
0.10	0.7985	0.8059	+0.74
0.20	0.8014	0.8091	+0.77

Table 2: **Three-agent full coordination versus the fitted mixture** ($\varepsilon = 0.05$). Differences are observed full-minus-mixture accuracy on the same evaluation cases. Reference accuracies: majority 0.8121, fit-selected agent 0.7971, selected constant 0.6182.

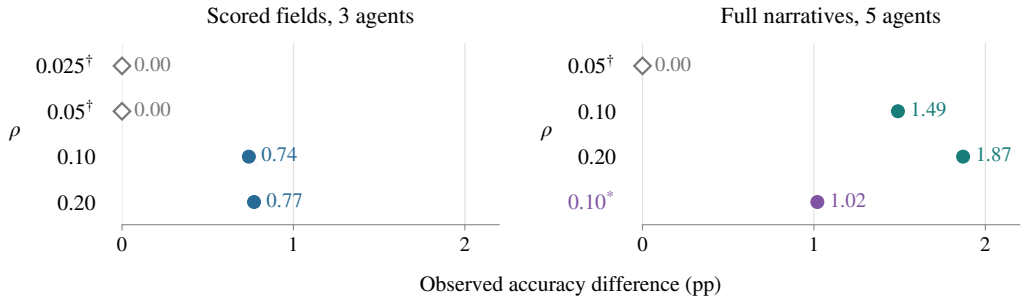

Figure 2: **Full-minus-mixture accuracy differences across LLM conditions.** Circles show observed accuracy differences on the evaluation cases. Open diamonds at zero mark analytic common-rule benchmarks, conditional on the fit-law assumptions of Proposition 3. The asterisk marks name/pronoun/proxy cues. Committee size, model composition, input evidence, and wording split change together between panels, so the comparison does not isolate the effect of committee size.

respectively. The resulting accuracies, 0.8059 and 0.8091, remain 0.62 and 0.30 pp below majority's 0.8121.

Tight-budget benchmarks. If $A_{\max,\mathrm{fit}} > 1.05\, C_{\mathrm{fit}}$ and the best agent and best constant are unique on the common fit law, Proposition 3 identifies a common optimum: the selected agent at $\rho = 0.05$, and its equal mixture with the selected constant at $\rho = 0.025$. Their evaluation accuracies are 0.7971 and $(0.7971 + 0.6182)/2 = 0.70765$, rounded to 0.7077; the selected-agent benchmark is 1.50 pp below majority. These entries evaluate the displayed analytic rules using the reported agent and constant accuracies. Their common-optimum interpretation is conditional on the fit-law assumptions above (Appendix I).

The five-agent comparison uses full narratives of 150–220 tokens and rationales capped at 96 tokens. Under explicit cues, the conditional benchmark at $\rho = 0.05$ is the selected agent (accuracy 0.7139); its common-optimum interpretation has the same fit-law requirements. At $\rho = 0.10$, full repair reaches accuracy 0.7311, compared with 0.7162 for the mixture and 0.7462 for majority, with observed gaps of 0.0179, 0.0158, and 0.0731, respectively. The observed full-minus-mixture differences are 1.49 and 1.87 pp at $\rho = 0.10$ and 0.20, and 1.02 pp at $\rho = 0.10$ under name/pronoun/proxy cues (Figure 2). These descriptive comparisons concern fitted rules on the evaluation cases. Equation (14) instead compares class optima on one distribution, so the observed differences do not measure the tightness of that bound.

5.4 Communication Ablations and Committee Controls

All three-agent conditions consist of an initial decision followed by two revisions on the same scored-field tasks. V0 permits self-revision without peer messages; V1 supplies the initial peer messages at both revisions; V2 refreshes the peer messages after each round. Each message contains a vote and at most 32 rationale tokens. Every condition evaluates 4,096 paired cases. The mixed committee uses eight runs per case and variant, and the homogeneous committees use four; cases remain the units of inference, and the run counts affect Monte Carlo precision.

Introducing fixed peer messages raises the observed majority gap from 0.0035 in V0 to 0.0231 in V1, and refreshing those messages in V2 raises it further to 0.0417. The V2–V1 difference is 1.86 pp and

Committee / condition	Accuracy	$\widehat{D}_{\mathrm{obs}}$
Mixed, V0	0.8086	0.0035
Mixed, V1	0.8104	0.0231
Mixed, V2	0.8121	0.0417
Qwen2.5-7B $\times 3$, V2	0.7938	0.0538
Llama-3.1-8B $\times 3$, V2	0.8026	0.0362
Mistral-7B $\times 3$, V2	0.7614	0.0294

Table 3: **Communication and committee controls.** Accuracy and observed group-rate difference of the majority coordinator. Entries are point estimates for the stated committee and communication condition.

the total V2–V0 difference is 3.82 pp, while accuracy increases by 0.35 pp from V0 to V2. In V2, the product-marginal reference contributes 0.0031 and the dependence residual 0.0386. This product-reference residual is a descriptive decomposition and does not identify the causal contributions of vote copying or rationale content.

For mixed V2, the observed individual gaps are 0.0024, 0.0031, and 0.0019, while majority's observed gap is 0.0417. Their point contrast is

$$\widehat{\Delta}_{\mathrm{majority}} - \max_i \widehat{\delta}_i = 0.0417 - 0.0031 = 0.0386. \tag{20}$$

This comparison describes the evaluated cases. The larger system gap is consistent with the composition mechanism studied by the theory; certifying population separation additionally requires uncertainty bounds.

The homogeneous controls exhibit observed majority gaps from 0.0294 to 0.0538. The five-agent explicit-cue majority gap is 0.0731, with 0.0288 under proxy cues. Committee size, models, input evidence, rationale length, and wording split change together, so differences across these designs cannot be attributed to committee size alone.

6 Discussion and Limitations

Local rate audits leave joint dependence unconstrained. Controlling the disparity for every compatible law restricts nonlinear coordination and, under an explicit tight-budget condition, makes an agent–constant mixture optimal. The reported comparisons exhibit a larger system gap than the individual gaps, positive observed accuracy differences at larger budgets, and a distinction between observed and worst-case disparity. Joint audits or valid marginal ranges (Appendix L.4) may retain useful interactions. Our studies concern historical data and generated cases; broader populations, templates, tool use, and long conversations remain untested. Deployment requires valid local tolerances, and the guarantee does not preserve utility under shift. Full-table optimization is exponential in the number of agents; count-only rules trade agent identity for scalability.

AI USE STATEMENT

In this work, we used generative AI tools (large language models) solely for language polishing, including improving grammar, clarity, and readability. These tools were not used to generate substantive content or to perform any research tasks, including developing methods or theoretical frameworks, formulating claims or hypotheses, constructing proofs, designing experiments, generating or processing data, analyzing or interpreting results, writing code, searching the literature, translating text, or preparing figures. All AI-assisted revisions were reviewed by the authors to ensure that they preserved the intended meaning and technical accuracy. We take full responsibility for the final content of this work, including all claims, results, and references.

REFERENCES

Alekh Agarwal, Alina Beygelzimer, Miroslav Dudík, John Langford, and Hanna Wallach. A reductions approach to fair classification. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 60–69. PMLR, 2018. URL https://proceedings.mlr.press/v80/agarwal18a.html.

Shipra Agrawal, Yichuan Ding, Amin Saberi, and Yinyu Ye. Correlation robust stochastic optimization. In *Proceedings of the Twenty-First Annual ACM-SIAM Symposium on Discrete Algorithms*, pp. 1087–1096, 2010. URL https://arxiv.org/abs/0902.1792.

R. R. Bahadur. A representation of the joint distribution of responses to n dichotomous items. In Herbert Solomon (ed.), *Studies in Item Analysis and Prediction*, pp. 158–168. Stanford University Press, Stanford, California, 1961.

Barry Becker and Ronny Kohavi. Adult [dataset]. UCI Machine Learning Repository, 1996. URL https://archive.ics.uci.edu/dataset/2/adult.

Dimitris Bertsimas and Ioana Popescu. Optimal inequalities in probability theory: A convex optimization approach. *SIAM Journal on Optimization*, 15(3):780–804, 2005. doi: 10.1137/S1052623401399903.

Peter J. Bickel and David A. Freedman. Some asymptotic theory for the bootstrap. *The Annals of Statistics*, 9(6):1196–1217, 1981. doi: 10.1214/aos/1176345637.

Amanda Bower, Sarah N. Kitchen, Laura Niss, Martin J. Strauss, Alexander Vargo, and Suresh Venkatasubramanian. Fair pipelines. In *Proceedings of the 4th Workshop on Fairness, Accountability, and Transparency in Machine Learning*, 2017. URL https://arxiv.org/abs/1707.00391.

Wenlong Chen, Yegor Klochkov, and Yang Liu. Post-hoc bias scoring is optimal for fair classification. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=FM5xfcaR2Y.

Thomas G. Dietterich. Ensemble methods in machine learning. In *Multiple Classifier Systems*, volume 1857 of *Lecture Notes in Computer Science*, pp. 1–15. Springer, 2000. doi: 10.1007/3-540-45014-9_1. URL https://doi.org/10.1007/3-540-45014-9_1.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. Improving factuality and reasoning in language models through multiagent debate. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 11733–11763. PMLR, 2024. URL https://proceedings.mlr.press/v235/du24e.html.

Cynthia Dwork and Christina Ilvento. Fairness under composition. In *10th Innovations in Theoretical Computer Science Conference (ITCS 2019)*, volume 124 of *Leibniz International Proceedings in Informatics*, pp. 33:1–33:20. Schloss Dagstuhl–Leibniz-Zentrum für Informatik, 2019. doi: 10.4230/LIPIcs.ITCS.2019.33.

Cynthia Dwork, Christina Ilvento, and Meena Jagadeesan. Individual fairness in pipelines. In *1st Symposium on Foundations of Responsible Computing (FORC 2020)*, volume 156 of *Leibniz International Proceedings in Informatics*, pp. 7:1–7:22. Schloss Dagstuhl–Leibniz-Zentrum für Informatik, 2020. doi: 10.4230/LIPIcs.FORC.2020.7.

Usman Gohar, Sumon Biswas, and Hridesh Rajan. Towards understanding fairness and its composition in ensemble machine learning. In *Proceedings of the 45th International Conference on Software Engineering*, pp. 1533–1545. IEEE, 2023. doi: 10.1109/ICSE48619.2023.00133.

Google. Gemma-2-9B-IT: Model card. Hugging Face, 2024. URL https://huggingface.co/google/gemma-2-9b-it.

Wassily Hoeffding. A class of statistics with asymptotically normal distribution. *The Annals of Mathematical Statistics*, 19(3):293–325, 1948. doi: 10.1214/aoms/1177730196.

Hans Hofmann. Statlog (german credit data) [dataset]. UCI Machine Learning Repository, 1994. URL https://archive.ics.uci.edu/dataset/144/statlog+german+credit+data.

Keyu Li, Jin Gao, and Dequan Wang. Aligned agents, biased swarm: Measuring bias amplification in multi-agent systems. In *International Conference on Learning Representations*, 2026. arXiv:2604.08963.

Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Shuming Shi, and Zhaopeng Tu. Encouraging divergent thinking in large language models through multi-agent debate. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 17889–17904. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.emnlp-main.992. URL https://aclanthology.org/2024.emnlp-main.992/.

Andreas Maurer and Massimiliano Pontil. Empirical Bernstein bounds and sample variance penalization. In *Proceedings of the 22nd Annual Conference on Learning Theory (COLT)*, 2009. arXiv:0907.3740.

Meta. Llama-3.1-8B-Instruct: Model card. Hugging Face, 2024. URL https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct.

Microsoft. Phi-3.5-mini-instruct: Model card. Hugging Face, 2024. URL https://huggingface.co/microsoft/Phi-3.5-mini-instruct.

Mistral AI. Mistral-7B-Instruct-v0.3: Model card. Hugging Face, 2024. URL https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.3.

Noam Nisan and Mario Szegedy. On the degree of boolean functions as real polynomials. *Computational Complexity*, 4(4):301–313, 1994. doi: 10.1007/BF01263419.

Ryan O'Donnell. *Analysis of Boolean Functions*. Cambridge University Press, Cambridge, 2014.

Maya Okawa. Emergence of biased consensus in multi-agent LLM debates. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. arXiv:2608.02827.

Guillermo Owen. Multilinear extensions of games. *Management Science*, 18(5):64–79, 1972. doi: 10.1287/mnsc.18.5.64.

Qwen Team. Qwen2.5-7B-Instruct: Model card. Hugging Face, 2024. URL https://huggingface.co/Qwen/Qwen2.5-7B-Instruct.

Ludger Rüschendorf. Fréchet-bounds and their applications. In Giorgio Dall'Aglio, Samuel Kotz, and Gabriella Salinetti (eds.), *Advances in Probability Distributions with Given Marginals*, pp. 151–187. Springer, 1991. doi: 10.1007/978-94-011-3466-8_9.

Kajetan Schweighofer, Adrian Arnaiz-Rodriguez, Sepp Hochreiter, and Nuria M. Oliver. The disparate benefits of deep ensembles. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 53663–53702. PMLR, 2025. URL https://proceedings.mlr.press/v267/schweighofer25a.html.

Bahar Taskesen, Viet Anh Nguyen, Daniel Kuhn, and Jose Blanchet. A distributionally robust approach to fair classification, 2020. URL https://arxiv.org/abs/2007.09530.

Alexandru Tifrea, Preethi Lahoti, Ben Packer, Yoni Halpern, Ahmad Beirami, and Flavien Prost. FRAPPÉ: A group fairness framework for post-processing everything. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 48321–48343. PMLR, 2024. URL https://proceedings.mlr.press/v235/tifrea24a.html.

Xuezhi Wang, Nithum Thain, Anu Sinha, Flavien Prost, Ed H. Chi, Jilin Chen, and Alex Beutel. Practical compositional fairness: Understanding fairness in multi-component recommender systems. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*, pp. 436–444. Association for Computing Machinery, 2021. doi: 10.1145/3437963.3441732.

Ruicheng Xian and Han Zhao. A unified post-processing framework for group fairness in classification. *arXiv preprint arXiv:2405.04025*, 2024. URL https://arxiv.org/abs/2405.04025.

A Proof of the Exact Vulnerability Characterization

Write $N = 2^m$, let $Z \in \mathbb{R}^{N\times m}$ have rows $\mathbf{z}^\top$, and identify h with its vector of values on the cube. The feasible set of pairs (P_0, P_1) is compact, nonempty, and invariant under swapping the groups. Hence the absolute-value supremum in Equation (2) equals the maximum of $h^\top(P_1 - P_0)$ over

$$\mathbf{1}^\top P_1 = \mathbf{1}^\top P_0 = 1, \quad Z^\top(P_1 - P_0) \le \boldsymbol{\varepsilon}, \quad -Z^\top(P_1 - P_0) \le \boldsymbol{\varepsilon}, \quad P_0, P_1 \ge 0.$$

Attach *free* multipliers $\lambda_1, \lambda_0 \in \mathbb{R}$ to the probability normalizations and $u, v \ge 0$ to the two marginal inequalities. The dual is

$$\begin{aligned} \min_{\substack{\lambda_0,\lambda_1\in\mathbb{R}\\ u,v\in\mathbb{R}^m_{\ge 0}}} \ & \lambda_1 + \lambda_0 + \boldsymbol{\varepsilon}^\top(u+v) \\ \text{subject to } & \lambda_1\mathbf{1} + Z(u-v) \ge h, \\ & \lambda_0\mathbf{1} - Z(u-v) \ge -h. \end{aligned} \tag{21}$$

For $b = u - v$, the least feasible envelope constants are

$$\lambda_1 = \max_{\mathbf{z}}\{h(\mathbf{z}) - b^\top\mathbf{z}\}, \qquad \lambda_0 = -\min_{\mathbf{z}}\{h(\mathbf{z}) - b^\top\mathbf{z}\}.$$

For fixed b, $u = b_+$ and $v = b_-$ minimize the nonnegative weighted sum $\boldsymbol{\varepsilon}^\top(u+v)$ and give $\sum_i \varepsilon_i|b_i|$. Thus the dual reduces to the first expression in Equation (3). The primal is feasible with a finite optimum; finite LP duality gives equality and attainment of the dual optimum. For any finite real-valued function f,

$$\operatorname{osc} f = 2\min_a \|f - a\|_\infty, \qquad a = \tfrac{1}{2}(\max f + \min f).$$

Centering the envelope proves the second expression. The affine functions form a closed finite-dimensional subspace, so zero distance is equivalent to affinity. This proves Theorem 1.

The envelope is not itself a coordinator. The affine approximation need not lie in $[0, 1]$. Imposing this additional restriction can increase the objective and destroy equality with $\mathcal{V}_{\boldsymbol{\varepsilon}}$; the resulting upper bound is still a valid, potentially conservative sufficient certificate. The probability-valued function constrained for deployment is h or g. Free normalization multipliers are essential: a constant rule $h \equiv 1$ must have vulnerability zero, attained in the dual with $b = 0, \lambda_1 = 1, \lambda_0 = -1$.

Fixed reports and the global envelope. For a prescribed marginal vector p, let $\mathcal{F}(p)$ be the laws with $\mathbb{E}Z = p$, and set

$$H^+(p) = \max_{P\in\mathcal{F}(p)} \mathbb{E}_P h, \qquad H^-(p) = \min_{P\in\mathcal{F}(p)} \mathbb{E}_P h. \tag{22}$$

These are finite LPs (the concave and convex closures on the cube). At a fixed report (p_0, p_1) the exact worst gap is

$$\max\{H^+(p_1) - H^-(p_0),\ H^+(p_0) - H^-(p_1)\}. \tag{23}$$

The extrema are separately attainable, giving $\mathcal{V}_0(h) = \max_p[H^+(p) - H^-(p)]$. For OR, $H^+(p) = \min(1, \sum_i p_i)$ and $H^-(p) = \max_i p_i$. Other thresholds generally require an LP for fixed heterogeneous reports; Theorem 2, proved next, also optimizes over the common report. When marginal levels can be monitored, Appendix L.4 gives an exact LP for report intervals; the global envelope protects against changes in both dependence and common rates.

B Exact Threshold Bounds and Attaining Laws

Proof of Theorem 2. The dual objective is convex in b and invariant under coordinate permutations when $h = h_k$ and $\boldsymbol{\varepsilon} = \varepsilon\mathbf{1}$. Averaging any minimizer over all permutations cannot increase the objective, so an optimum has $b = \beta\mathbf{1}$. If $\beta < 0$, the oscillation of $\mathbf{1}\{n \ge k\} - \beta n$ on $n = 0, \dots, m$ is $1 + m|\beta| \ge 1$, whereas $\beta = 0$ gives value one. We may restrict to $\beta \ge 0$. On that half-line,

$$\begin{aligned} \max_n\{\mathbf{1}\{n \ge k\} - \beta n\} &= \max(0, 1 - \beta k), \\ \min_n\{\mathbf{1}\{n \ge k\} - \beta n\} &= \min(-\beta(k-1), 1 - \beta m). \end{aligned}$$

Let $\beta_k = \min\{1/k, 1/(m-k+1)\}$. Before the first breakpoint, the oscillation is $1 - \beta$. After it, the slope is either $k - 1$ or $m - k$ until the second breakpoint, then m; all are nonnegative. The full objective on $[0, \beta_k]$ is $1 + (m\varepsilon - 1)\beta$, and it is nondecreasing thereafter. Its minimum is therefore $1 - \beta_k + m\varepsilon\beta_k$ when $m\varepsilon \le 1$, and one otherwise. Setting $\varepsilon = 0$ gives Equation (6). □

Explicit attainment at every tolerance. Let U_j be uniform on the vectors containing j ones, and let $d = \min(m\varepsilon, 1)$. All coordinates of U_j have mean j/m.

- If $k \geq m - k + 1$, take

$$P_0 = U_{k-1}, \qquad P_1 = \frac{k-1+d}{k} U_k + \frac{1-d}{k} U_0.$$

  Each marginal increases by $d/m \leq \varepsilon$; the system gap is $(k-1+d)/k = 1 - \beta_k + d\beta_k$.

- If $k \leq m - k + 1$, take

$$P_1 = U_k, \qquad P_0 = \frac{1-d}{m-k+1} U_m + \left(1 - \frac{1-d}{m-k+1}\right) U_{k-1}.$$

  The marginal increase is again d/m, and the gap is $1 - (1-d)/(m-k+1) = 1 - \beta_k + d\beta_k$.

At the tie either construction suffices. At $d = 1$, adjacent layers U_k, U_{k-1} attain gap one. The construction includes $m = k = 1$, where $\mathcal{V}_\varepsilon(h) = \varepsilon$. Realize any constructed law by $X \mid A = a \sim P_a$ and deterministic actions $Z_i = X_i$, which do not take A as a separate input.

An alternative zero-tolerance derivation. For a count $N \in \{0, \ldots, m\}$ of mean s,

$$\left[\frac{s-k+1}{m-k+1}\right]_+ \leq \mathbb{P}(N \geq k) \leq \min(1, s/k).$$

The upper inequality is Markov's bound. For the lower inequality, use $s \leq (k-1)\mathbb{P}(N < k) + m\mathbb{P}(N \geq k)$. The upper bound is attained by a mixture of U_0, U_k when $s \leq k$ and of U_k, U_m when $s \geq k$; a degenerate endpoint suffices if $k = m$. For the lower bound, when $0 \leq s \leq k-1$ mix U_0 and U_{k-1} to attain mean s with zero threshold probability (for $k = 1$, this interval consists only of $s = 0$). When $s \geq k - 1$, mix U_{k-1} and U_m with weight $(s-k+1)/(m-k+1)$ on U_m. The largest difference of the two envelopes occurs at $s = k-1$ or $s = k$, recovering the zero-tolerance formula.

C The Joint-Law Diagnostic and Its Interpretation



Figure 3: **Local parity leaves joint decisions undetermined.** Three agents have selection rate $1/3$ in each group. Aligning their approvals in one group and separating them in the other changes OR's selection rate from $1/3$ to 1, and AND's from $1/3$ to 0.

The case-conditioned diagnostic. For a nonlinear coordinator, the observed disparity also depends on interactions among agents. Let $\bar{h}(p)$ be its outcome probability under independent Bernoulli actions with rates p. For paired cases W drawn from a common evidence distribution, define

$$D_{\text{obs}} = D_{\text{marg}}(W) + D_{\text{dep}}(W), \qquad D_{\text{marg}}(W) = \mathbb{E}_W[\bar{h}(p_1(W)) - \bar{h}(p_0(W))]. \tag{24}$$

The multilinear expansion expresses the residual through derivatives of $\bar{h}$ and joint central moments (Bahadur, 1961; Owen, 1972). Its allocation depends on W and has no intervention interpretation without further assumptions. With R stochastic runs per case, products of sample rates retain $O(1/R)$ bias. Cross-run U-statistics use different runs for different agents to remove this bias (Hoeffding, 1948). Appendix D gives the estimator and its assumptions.

The multilinear extension of a cube function satisfies the pointwise expansion

$$h(\mathbf{z}) = \sum_{S \subseteq [m]} \partial_S \bar{h}(p) \prod_{i \in S} (z_i - p_i).$$

Taking expectation eliminates the singleton terms and proves Equation (25). This is the classical Bernoulli/multilinear representation (Bahadur, 1961; Owen, 1972). It uses central joint moments; those moments are not independent causal mechanisms. For an affine coordinator every term of order at least two vanishes.

Conditioning can move the entire gap between channels. With $m = 2$ and $h = z_1 z_2$, consider two equally weighted cases with conditionally independent actions. The rates in group 1 are $(0.8, 0.8)$ and $(0.2, 0.2)$; in group 0 they are $(0.6, 0.6)$ and $(0.4, 0.4)$. The observed signed gap is 0.08. At case-level conditioning, $D_{\text{marg}} = 0.08$ and $D_{\text{dep}} = 0$ because there is no *within-case* dependence. At constant W, pooled rates are $1/2$ in both groups, giving $D_{\text{marg}} = 0$ and $D_{\text{dep}} = 0.08$. Pooling heterogeneous cases induces dependence; the system has not changed. Conditioning must therefore be stated with each estimate.

Average local parity is weaker than case-level parity. For one agent, case-specific rates $(p_{1,c}, p_{0,c}) = (0.8, 0.6)$ and $(0.2, 0.4)$ give zero average local gap but opposite nonzero cue responses. Reweighting cases changes its average marginal response. In this one-agent example the dependence residual remains identically zero, since every binary one-agent coordinator is affine. For nonlinear multi-agent rules, changing marginals can also change the derivative weights in Equation (25). Hence a residual change need not be attributable solely to a change in central dependence moments.

A product law is a reference, not necessarily an intervention. Two agents that both output $\mathbf{1}\{X > c\}$ stay perfectly correlated without communication. The product law places probability on disagreements unavailable to interventions on messages alone. It is nevertheless evaluable for a known replayable coordinator. Conversely, unchanged D_{dep} need not mean unchanged joint dependence: it is a scalar weighted contrast, and different joint moments can cancel or receive zero weight. Finally, a coordinator with large $\mathcal{V}_0$ can have zero observed disparity whenever $P_1 = P_0$. These distinctions motivate separate reporting of observed gap, reference, residual, and structural vulnerability.

D Cross-Run Audit: Estimator and Inference

D.1 Diagnostic identity and estimation

Worst-case vulnerability does not measure the disparity of a particular deployment. For that purpose let $\bar{h}(p) = \mathbb{E}_{\otimes_i \text{Bern}(p_i)} h(\mathbf{Z})$ be the multilinear extension, and $\mu_{S,a} = \mathbb{E}_{P_a} \prod_{i \in S} (Z_i - p_{i,a})$. The classical multilinear Taylor identity gives

$$\mathbb{E}_{P_a} h = \bar{h}(p_a) + \sum_{|S| \geq 2} \partial_S \bar{h}(p_a) \mu_{S,a}. \tag{25}$$

The signed decomposition is $D_{\text{obs}} = D_{\text{marg}} + D_{\text{dep}}$, with $D_{\text{marg}} = \bar{h}(p_1) - \bar{h}(p_0)$ and $D_{\text{dep}} = D_{\text{obs}} - D_{\text{marg}}$. The derivative weights identify which dependence terms the coordinator reads. The decomposition is exact, but individual terms may cancel and a large $\mathcal{V}_0(h)$ need not imply a large observed gap.

Conditioning defines the diagnostic. Choose legitimate information W before evaluation and apply (25) conditionally on (A, W). Then

$$D_{\text{marg}}(W) = \mathbb{E}[\bar{h}(p_1(W)) \mid A = 1] - \mathbb{E}[\bar{h}(p_0(W)) \mid A = 0], \qquad D_{\text{dep}}(W) = D_{\text{obs}} - D_{\text{marg}}(W). \tag{26}$$

Changing W changes the allocation between channels, not D_{obs}. At W constant, local rates are sufficient for the reference term. At finer W, conditional rates are needed. Conditional parity plus conditional independence implies parity *within each stratum*; unconditional parity additionally requires an appropriate common distribution of strata. For equalized odds, evaluate each label stratum separately instead of averaging it away.

Limits of attribution. The independent product law is a well-defined statistical reference, including when it puts mass on unobserved action patterns for a known replayable coordinator. It need not be realizable by changing only communication in the original system. Even exact *average* local parity does not imply equality of case-specific marginal responses. When those responses differ, a change in $D_{\mathrm{dep}}(W)$ can reflect marginal changes as well as dependence changes. Therefore we report signed channels and conditional-sensitivity diagnostics, and do not infer a causal mediator or the optimal repair layer from channel magnitudes. Appendix C gives explicit counterexamples.

Repeated-run logs. Equation (18) defines the complete injective estimator for paired cases. The next result states its assumptions and inference guarantees; the fixed-run plug-in bias and a subset algorithm are derived below.

Proposition 4 (Unbiased case-level audit)**.** *Suppose cases are i.i.d., and for each group the R runs are i.i.d. conditional on the case, with $R \geq m$. Cross-group dependence within a case is allowed. Then $\hat{M}_{c,a}$ is unbiased for $\bar{h}(p_{c,a})$, and the two channel averages are unbiased for (26) under the paired case distribution. Sampling a fixed number B of uniform injective tuples independently across cases preserves unbiasedness. For fixed R, m, B, the case averages satisfy a multivariate central limit theorem, and a case-block bootstrap is consistent for their centered law.*

Each recombined coordinate comes from a different independent run; hence its expectation is the product-law expectation. Bounded i.i.d. case summaries give the inference result (Hoeffding, 1948; Bickel & Freedman, 1981); Section D.2 gives the proof, and Appendix K an $O(nRm2^m)$ subset algorithm for complete recombination. Independent incomplete-replay randomness is part of the case block. Affine coordinators are evaluated analytically so their residual is zero up to numerical roundoff, not merely zero in expectation. The number of independent cases, rather than the number of runs, determines population uncertainty. Given sampled tuples, coordinate assembly costs $O(nBm)$ plus coordinator evaluations; a degree-d polynomial can instead use order-d U-statistics when $R \geq d$.

D.2 Proof and computation

Proof of Proposition 4. Conditional on case c, the vectors from distinct runs are independent within each group. Every injective tuple in Equation (18) therefore draws its m coordinates independently with the required case/group marginal rates, so its conditional expected coordinator output is $\bar{h}(p_{c,a})$. Averaging preserves this expectation. The original-run average has expectation $\mathbb{E}[h(\mathbf{Z}) \mid c, a]$, giving unbiased marginal and residual differences after averaging cases. Arbitrary coupling between the two group variants does not affect these expectations.

For incomplete recombination, sample B injective tuples independently of the data and use independent tuple randomness across cases. Conditional expectation over that randomness recovers the complete statistic, so the tower property proves unbiasedness. Treat this extra randomness as part of the case block.

For fixed R, m, B, the vector of case-level channel summaries is bounded and i.i.d. Its centered average obeys the multivariate CLT, including a possibly degenerate limiting covariance. The ordinary case-block bootstrap consistently estimates this centered sampling law (Bickel & Freedman, 1981). This is asymptotic consistency, not a finite-sample coverage guarantee. No additional limit in R is required, since the independent units for this claim are cases. The injective average is a symmetrized U-statistic within each case (Hoeffding, 1948). □

Fixed-run plug-in bias. For $h = z_1 z_2$, expanding the empirical product into equal-run and distinct-run terms gives

$$\mathbb{E}[\hat{p}_1 \hat{p}_2 \mid c, a] = p_1 p_2 + \mathrm{Cov}(Z_1, Z_2 \mid c, a)/R.$$

The extra term survives any increase in the number of cases at fixed R. For synchronized group 0 and independent group 1 with both marginal rates $1/2$, the true residual is $-1/4$ and its plug-in expectation is $-(1 - 1/R)/4$.

Exact evaluation without enumerating tuples. Appendix K derives the subset recurrence used by the implementation, costing $O(nRm2^m)$ for complete recombination. This is practical for modest m and avoids the combinatorial $(R)_m$ enumeration. For larger panels, B sampled tuples cost $O(nBm)$ plus coordinator evaluations. If the polynomial degree is $d < m$, monomial-specific

order-d recombination suffices for $R \geq d$. Affine references are computed analytically to make their residual exactly zero up to numerical roundoff. Resample whole cases, including both groups and all conditions; resampling individual runs as if they were new cases understates case-population uncertainty.

E Robust Repair and Matched Baselines

Proof of Proposition 2. For each table g, Theorem 1 attains its minimum. Thus $\mathcal{V}_{\boldsymbol{\varepsilon}}(g) \leq \rho$ if and only if there exist a, b, t with $|g(\mathbf{z}) - a - b^\top \mathbf{z}| \leq t$ for all cells and $2t + \sum_i \varepsilon_i |b_i| \leq \rho$. Slack variables $v_i \geq \pm b_i$ give Equation (12); they can be chosen as $|b_i|$. With a linear utility and linearized proximity cost the full optimization is an LP. Adding the empirical disparity constraint intersects this feasible set without weakening the structural guarantee. Constants are always feasible. The rule variables number 2^m, with $O(m)$ auxiliary variables and $O(2^m + m)$ inequalities before any optional shape constraints.

If $g(\mathbf{z}) = q_{\sum_i z_i}$ and $\boldsymbol{\varepsilon} = \varepsilon \mathbf{1}$, permutation averaging makes the dual slope $b\mathbf{1}$ without increasing the objective. One envelope constraint pair per count $j = 0, \ldots, m$ suffices, with budget $2t + m\varepsilon v \leq \rho$. Utility coefficients aggregate over count layers. This establishes an exact reduced LP *within the count-only class*; it need not match the best identity-sensitive rule. □

Monotonicity and other optional constraints. Full-table monotonicity is enforced on every cube edge $g(\mathbf{z}) \leq g(\mathbf{z} + e_i)$ for $z_i = 0$, adding $m2^{m-1}$ inequalities. Count monotonicity uses $q_j \leq q_{j+1}$. These constraints are distinct from the $O(2^m + m)$ base robust formulation. The experiments set proximity weight $\gamma = 0$ and maximize ordinary expected accuracy, or majority fidelity in the synthetic stress test. A fairness-penalty baseline uses an epigraph for the absolute fitted gap, with the same monotone full-table class and utility objective.

Why constants and single agents are necessary controls. For a monotone affine coordinator $g(\mathbf{z}) = a + \sum_i b_i z_i$, probability-valuedness and monotonicity imply $a \geq 0$, $b_i \geq 0$, and $a + \sum_i b_i \leq 1$. Hence

$$g = a \cdot 1 + \sum_i b_i z_i + (1 - a - \sum_i b_i) \cdot 0$$

is a convex mixture of the two constants and the single-agent rules. Without an additional restrictive budget, maximizing a linear utility over this class chooses a best atom. With a budget, mixtures can be necessary. The stronger baseline in Equation (13) also includes the original rule and uses the full optimizer's affine-envelope constraints. All coefficients are fitted on the coordinator-fit split. This contains arbitrary convex shrinkage of majority toward a monotone affine coordinator, not only a fixed random-agent mixture. The full monotone optimizer contains this class; its fit objective cannot be worse at matched budgets, while held-out utility can be worse.

Scope of the guarantee. The uncertainty set permits arbitrary common-rate and dependence changes subject to the specified local gaps. It constrains fairness, not shifted utility or individual fairness. A score subsequently thresholded by another module is a different coordinator and must be audited again. A coordinator, agents, or decoding rule changed after certification also defines a new system. Full-table methods remain exponential in agent count; the count-only reduction obtains tractability by discarding agent identity.

F Finite-Sample Certificates

Section 5.1 gives the paired-case radius, gap bounds, and non-inferiority criterion. Here we derive the range adjustment and unpaired-group construction. These formulas describe how to certify a result from case-level records; they do not assign confidence endpoints to the current aggregate comparisons.

For independent variables in an interval of width w, let S^2 be the usual sample variance with denominator $n - 1$. The target is $n^{-1} \sum_c \mathbb{E} X_c$; the variables need not be identically distributed (Maurer & Pontil, 2009, Theorem 11). Applying the empirical-Bernstein inequality to each tail and

taking a union bound across K fixed candidates gives the radius

$$r_w(n, S^2, K, \alpha_f) = \sqrt{\frac{2S^2 \log(4K/\alpha_f)}{n}} + \frac{7w \log(4K/\alpha_f)}{3(n-1)}, \qquad n > 1.$$

The variance uses independent sampling units. Conditioning on a fixed task/template schedule preserves the bound when the case blocks are independent; its target is the average expectation for that schedule. Under i.i.d. sampling this reduces to the ordinary population mean. This extension concerns the finite-sample certificate, not a new bootstrap-coverage claim. For paired group variants or utility losses on the same subject, $w = 2$, which gives Equation (19). With repeated runs, first average within case. For the residual itself, the per-case signed contrast lies in $[-2, 2]$, so use $w = 4$ rather than the system-gap radius.

Unpaired groups. For expected outputs $s_j \in [0, 1]$, compute separate group means, variances S_a^2, and sizes $n_a > 1$. Divide the family budget equally between the two groups. Then

$$r_a = \sqrt{\frac{2S_a^2 \log(8K/\alpha_f)}{n_a}} + \frac{7 \log(8K/\alpha_f)}{3(n_a - 1)}, \qquad |D - (\bar{s}_1 - \bar{s}_0)| \le r_0 + r_1$$

simultaneously across the family. No claim about random cross-group matching is needed. In the archived tabular split, Adult uses $(n_0, n_1) = (5421, 10860)$ and German $(62, 138)$. The small German group drives the large generic radius.

Error allocation and interpretation. For the archived tabular primary split, set $\alpha = 0.05$ and assign $\alpha/3$ each to three local-agent gaps, 46 coordinator gaps, and 46 paired accuracy losses relative to calibrated majority. A union bound controls these families jointly under the sampling assumptions, conditional on fitted candidates. The five mixture candidates are an exploratory follow-up on an already inspected benchmark; the simultaneous construction is computed over the full expanded roster, but no fresh confirmatory evaluation is claimed. Unreported adaptive use of the same evaluation sample would require additional control or new data.

A system current-gap upper bound below 0.05 and a signed loss upper bound below 0.01 constitute the stated certification targets. A point estimate inside either margin is insufficient. To invoke the structural tolerance $\varepsilon = 0.05$, also check each local upper bound against 0.05. This does not validate the premise under an arbitrary future shift. Historical tabular data and label-stratified benchmark splits need not be literal i.i.d. superpopulation samples; population interpretations are conditional on that model. Paired model cases drawn from a finite template pool have an additional generalization limitation. Exact finite-benchmark expected outcomes remain well defined without a population model.

Sample-size floor. For a paired local-gap analysis with $K = 3$, $\alpha_f = 0.05/3$, even zero observed variance leaves the term $14 \log(720)/[3(n-1)]$. It is below 0.02 only at $n \ge 1537$. Larger K or nonzero variance raises this requirement. Increasing runs per case does not eliminate this case-level remainder. This explains why a small pilot cannot support a tight generic local certificate even when its point gap looks favorable. A structurally constant coordinator is fair analytically; the generic statistical bound need not exploit this fact.

G Recent Fairness Post-Processors

Our implementations adapt the post-processing objectives to the eight action vectors of three calibrated agents. The following specifications document these finite-state adaptations. The implementation and evaluation records are distinguished in Appendix H.

Shared scores and data access. For fitted subject i, let $J_i(z)$ be the probability of action vector z, obtained by integrating the cached independent Bernoulli agent responses. Put $c_z = \sum_i J_i(z)$ and $w_z = c_z/n$. The label and group scores are

$$\hat{\eta}_z = \frac{\sum_i J_i(z) y_i + 1/2}{c_z + 1}, \qquad \hat{a}_z = \frac{\sum_i J_i(z) a_i + 1/2}{c_z + 1}.$$

These scores and the post-processing parameters use the coordinator-fit split only. Group labels are available during fitting; inference uses only z. The group-law difference $d_z = \hat{P}_1(z) - \hat{P}_0(z)$ defines the fitted DP gap $|d^\top g|$. Test utility and disparity integrate response randomization exactly. Recent competitors need not be monotone; JointRepair retains its monotonicity constraint.

MBS. Following Chen et al. (2024), let $b_z = \mathbf{1}\{\hat{\eta}_z > 1/2\}$ and let π be fitted group prevalence. The DP bias score is

$$s_z = \frac{(2b_z - 1)\{(1 - \hat{a}_z)/(1 - \pi) - \hat{a}_z/\pi\}}{|2\hat{\eta}_z - 1|}.$$

A 10^{-12} denominator floor handles ties. We use the authors' threshold grid: 5,000 equally spaced values on $[0, 50]$, plus infinity, traversed in decreasing order. Predictions above the threshold are flipped in the direction of the fitted disparity; traversal stops before that disparity changes sign. Selection maximizes fit accuracy among retained rules with gap at most τ. If none qualifies, the least-gap rule is retained and marked as failing the target.

FRAPPÉ-KL. We instantiate Equation (7) of Tifrea et al. (2024) with a cell-specific logit correction, KL discrepancy, and a squared soft-DP penalty:

$$\min_{q \in [0,1]^8} \sum_z w_z \, \mathrm{KL}(\mathrm{Ber}(q_z) \,\|\, \mathrm{Ber}(\hat{\eta}_z)) + \lambda (d^\top q)^2.$$

For $w_z > 0$, the convex optimum satisfies $q_z = \sigma(\mathrm{logit}\, \hat{\eta}_z - t d_z / w_z)$ with $t = 2\lambda d^\top q$; a scalar monotone root solve determines t. Empty cells retain their base score. The final classifier is $g_z = \mathbf{1}\{q_z \geq 1/2\}$, so the soft penalty alone does not guarantee hard-decision DP. We evaluate $\lambda \in \{0, 0.1, 0.5, 1, 2, 5, 10, 25, 50, 100, 250, 1000, 10000\}$ and apply the same fit-accuracy/actual-gap selection rule as MBS. This is one declared instance of the FRAPPÉ framework.

LinearPost. For Xian & Zhao (2024), we solve the weighted finite-support primal with predicted group prior $\hat{\pi} = \sum_z w_z \hat{a}_z$:

$$\max_{0 \leq g \leq 1} \sum_z w_z (2\hat{\eta}_z - 1) g_z \quad \text{s.t.} \quad \left| \sum_z w_z \left\{ \frac{\hat{a}_z}{\hat{\pi}} - \frac{1 - \hat{a}_z}{1 - \hat{\pi}} \right\} g_z \right| \leq \tau.$$

We retain randomized primal ties on the atomic input space rather than add perturbations to the dual decision rule. Empty cells retain their base scores. The predicted-group constraint need not equal the actual fitted group gap after score estimation and smoothing; both are saved. All 40 primal optima agree with an independently formulated dual within 10^{-8}.

H Public-Data Protocol and Result Records

Materials and validation scope. The current public-data and LLM comparisons are reported as aggregate point estimates. Matching case-level outputs, complete prompts, fit accuracies, fitted rules, and solver residuals are not included for independent reconstruction. Accordingly, the current tables and figure present descriptive differences without confidence endpoints. Archived tabular per-subject outputs and seed-level summaries belong to a separate evaluation. Tight-budget entries evaluate analytic benchmarks whose optimality interpretation is conditional on the stated fit-law assumptions.

The protocol below documents the tabular implementation and its archived split design. The current aggregate record does not establish whether Table 1 reused the identical split identifiers, seed, and candidate roster; those archived settings are therefore identified explicitly below.

Data and targets. We use the original UCI Adult and Statlog German Credit archives, retaining all 48,842 and 1,000 rows, respectively. The archives, SHA-256 hashes, source URLs, and dataset attribution are provided. For Adult, $Y = 1$ denotes income above \$50K; for German, $Y = 1$ denotes the original "good" credit label. In both experiments $A = 1$ denotes the source dataset's male category and $A = 0$ its female category. These historical binary encodings do not describe the full range of gender identities. German sex is derived from personal-status codes A91/A93/A94 versus A92/A95, and the entire personal-status variable is removed from prediction. Adult excludes sex,

race, and sampling weight, and uses education number in place of the duplicate education category. Remaining features can proxy protected information; excluding a field does not establish statistical or causal independence from it.

Agent blocks and preprocessing. Adult blocks are (1) education number and occupation; (2) capital gain, capital loss, and weekly hours; (3) age, work class, marital status, relationship, and country. German blocks are (1) duration, purpose, and amount; (2) checking status, credit history, savings, employment, and installment rate; (3) guarantor, residence, property, age, other installment plans, housing, existing credits, job, dependents, telephone, and foreign-worker status. Every feature belongs to at most one block. Each agent is a logistic regression with $C = 1$, LBFGS, and at most 1,000 iterations. Numeric median imputation and standardization, categorical modal imputation, and one-hot encoding are fitted only on the agent-training split. Unknown test categories map to zero indicators. No hyperparameter search is performed. This intentionally simple, constrained ensemble is not a state-of-the-art tabular predictor.

Reference split design. The archived primary seed is 20260913. Adult preserves its official test split; its original training pool is divided into $19{,}536$ training, $6{,}512$ calibration, and $6{,}513$ coordinator-fit rows. German reserves a fixed 200-row test set and divides its remaining rows into 480/160/160. Splits are stratified by the joint (A, Y) category. Ten consecutive seeds vary training/calibration/fit allocation while retaining the same held-out test set. Every source row ID is recorded, and the code asserts disjointness of all four index sets. Repeated feature vectors and source sampling dependencies are not automatically removed. A source record, rather than a Bernoulli response draw, is the unit of analysis.

Local calibration and action laws. For each agent, calibration-set score quantiles define up to ten bins. A probability $q_b \in [0, 1]$ for each bin is fitted to maximize calibration accuracy subject to $q_b \leq q_{b+1}$ and an empirical demographic-parity gap at most 0.02. Both groups use the same bins and probabilities at inference. This is a group-blind score-bin post-processor, not a claim to implement the full reductions algorithm of Agarwal et al. (2018). On a subject x, independent random coins with these response probabilities define an exact eight-cell law. All expectations integrate those coins analytically. Dependence in the population arises from variation in shared subject evidence, even though the action coins are conditionally independent. The coordinator has no access to A.

Reference fitting configuration. The archived experiment fits 41 calibrated-agent coordinator candidates: majority; best single agent chosen by fit accuracy; random-agent voting; fit-selected constant; full/count-only empirical LPs with $\tau = 0.05$; full/count-only robust LPs for all combinations of $\varepsilon \in \{0.02, 0.05, 0.10\}$ and $\rho \in \{0.025, 0.05, 0.10, 0.20, 0.40\}$; and five identical-class penalty LPs at $\lambda \in \{0.1, 0.5, 1, 2, 5\}$. Penalty objectives maximize fit accuracy minus $\lambda|D|$. Every optimized rule is monotone and uses $\gamma = 0$. Utility is pooled ordinary accuracy at the empirical group prevalence. Unseen cells have zero utility weight but remain subject to the structural constraints. The primary robust configuration is full-table $(\varepsilon, \rho) = (0.05, 0.10)$; it is not selected by test performance. Its follow-up adds five optimally fitted mixtures of majority, all single agents, and both constants at the same five budgets and $\varepsilon = 0.05$. These use the identical fit split and structural constraints, giving 46 calibrated-coordinator candidates in total. The raw majority, before local calibration, is reported separately. All fitted response tables and grids are retained, including inferior candidates.

Scoring and structural checks. Table 1 reports accuracy and observed disparity under probability integration. Thresholding output probabilities would define a different coordinator. The majority value agrees with the analytic $\mathcal{V}_{.05}(h_2) = 0.575$, and the mixture and full-repair structural values lie within the budget 0.10. These are consistency checks on the reported summaries, with validation scope specified above.

I LLM Evaluation: Protocols, Complete Comparisons, and Inference

Reported comparisons. The LLM tables and comparison figure report aggregate point estimates, with material scope given in Appendix H. Larger-budget entries compare fitted methods. Tight-budget

entries are conditional analytic benchmarks, marked † and defined below. Uncertainty estimation is specified separately in Appendix I.1.

I.1 THREE-AGENT PROTOCOL AND METHOD COMPARISONS

Models and task distribution. The three-agent committee contains three agents, one each from Qwen2.5-7B-Instruct, Llama-3.1-8B-Instruct, and Mistral-7B-Instruct-v0.3 (Qwen Team, 2024; Meta, 2024; Mistral AI, 2024). Three homogeneous controls use the same number of roles with one model family throughout. The mixed committee evaluates coordination across families; it does not disentangle a model-family effect from its assigned role.

Four tasks cover hiring, credit, graduate admissions, and tenancy screening. The generator seed is 20260914. Each task has 1,280 newly generated cases, split into 256 fit and 1,024 evaluation cases: 1,024 fit and 4,096 evaluation cases in total. Eight wording families per task are assigned round-robin and balanced across the splits. Thus the evaluation concerns new cases under the study's template mixture, rather than an unseen-template-family test. Cases carry three task-specific scores in $\{0, \ldots, 10\}$, each agent sees its assigned field, and the label is positive when at least two task criteria pass. Labels are computed before inference. Paired female/male prompts use identical evidence and labels and differ only in the audited cue; both state that sex is irrelevant. These are controlled decision tasks, not outcomes for actual applicants.

Communication and sampling. Each condition has an initial decision and two revisions. V0 permits self-revision without peers; V1 supplies initial peer messages and retains them at the second revision; V2 refreshes peer messages after each round. Messages contain a binary vote and a rationale of at most 32 tokens. Decisions use a constrained one-token `no`/`yes` interface at temperature one. With rationale histories, the effective state space is too large for practical exact integration, even though bounded token sequences have finite support.

There are $R = 8$ runs per case/group/condition for the mixed committee and $R = 4$ for each homogeneous V2 control. Both satisfy $R \geq m = 3$. Runs are independent conditional on a case; the case block contains both cue variants and all compared conditions. Without peers, conditional independence of agents additionally requires independent agent random streams. Across fit and evaluation splits, the recorded terminal outputs comprise 737,280 per-agent actions for the mixed committee and 122,880 for each control, totaling 1,105,920. These counts refer to final agent actions; they do not count every intermediate revision or rationale token as a separate forward pass.

Communication and committee controls are in Table 3. Table 4 reports the pooled-accuracy comparison, followed by two analytic benchmarks for tighter budgets.

Coordinator	ρ	Accuracy	Gap	$\mathcal{V}_{.05}$
Majority (Dietterich, 2000)	–	0.8121	0.0417	0.575
Random agent (§ 5.1)	–	0.7893	$\leq$ 0.0026	0.050
Best single (§ 5.1)	–	0.7971	0.0031	0.050
Constant (§ 5.1)	–	0.6182	0.0000	0.000
Empirical ($\tau = 0.05$) (Agarwal et al., 2018, adapt.)	–	0.8121	0.0417	0.575
Empirical ($\tau = 0.02$) (Agarwal et al., 2018, adapt.)	–	0.8088	0.0193	0.488
Fitted mixture (§ 5.1)	0.10	0.7985	0.0068	0.100
JOINTREPAIR (ours)	0.10	0.8059	0.0091	0.100
Fitted mixture (§ 5.1)	0.20	0.8014	0.0141	0.200
JOINTREPAIR (ours)	0.20	0.8091	0.0168	0.200
Analytic benchmarks				
Agent–constant benchmark†	0.025	0.7077	0.0016	0.025
Selected-agent benchmark†	0.05	0.7971	0.0031	0.050

Table 4: **Complete three-agent V2 comparison.** $\varepsilon = 0.05$. †: conditional common-rule benchmarks, using the reported selected-agent and constant results. The random-agent gap is an upper bound inherited from the local point gaps. Fitted-rule structural values follow the current aggregate record (Appendix H).

System/local contrast. Equation (20) reports the system-minus-largest-local point difference. Its value, 0.0386, happens to equal the dependence residual, although the two quantities are different estimands. For an agent–constant mixture with nonnegative agent weights w_i, the observed gap is bounded by $\sum_i w_i \widehat{\delta}_i$. Uniform random-agent voting therefore has an observed gap at most 0.0026, using the displayed three-agent rates and outward rounding.

The case distribution is defined by a fixed schedule of 32 task–wording cells. Statistical inference conditional on that schedule concerns its average expectation; generalization to new wording families requires a separate sampling design. The observed contrast is not assigned a population-separation certificate.

Dependence diagnostic. Channel estimates use complete injective recombination at case-level conditioning, as defined in Appendix D. Original panels estimate outcomes and utility; different run indices for different agents estimate the product-marginal reference. At V2 the reference is 0.0031 and the residual is 0.0386, or 92.6% of the signed observed contrast. The latter ratio is descriptive. Small pooled local gaps alone do not bound the reference by their sum: the valid bound is

$$|D_{\mathrm{marg}}(W)| \le \sum_i \mathbb{E}_W |p_{i,1}(W) - p_{i,0}(W)|. \tag{27}$$

For majority, $|\partial_i \bar{h}| \le 1$; apply the mean-value bound within each case and then average. The expectation of an absolute difference cannot be replaced by the absolute pooled difference.

Analytic benchmarks and their optimality conditions. Proposition 3 identifies a common fitted optimum under the following conditions: both methods must use the same pooled fit distribution, ordinary accuracy with $\gamma = 0$, the same monotonicity and structural constraints, and no additional empirical-gap constraint. The fit law must satisfy $A_{\max,\mathrm{fit}} > (1+\varepsilon) C_{\mathrm{fit}}$. A unique best agent determines the rule at $\rho = \varepsilon$; at $0 < \rho < \varepsilon$, uniqueness of the best constant is also required. The mixture includes majority, every single agent, and both constants. Under these assumptions, both exact optima equal

$$g_\rho(z) = \frac{\rho}{\varepsilon} z_{j^*} + \left(1 - \frac{\rho}{\varepsilon}\right) c^*.$$

The tight-budget entries evaluate g_ρ from the reported selected-agent and constant results. The aggregate record does not establish the fit-law conditions, so these entries are analytic references rather than observations of independently fitted optimizer agreement. Verifying such agreement would require comparing all 2^m output probabilities and checking fit accuracies, mixture weights, and feasibility and optimality residuals.

Approximate solver agreement. If $d = (A_{\max,\mathrm{fit}} - C_{\mathrm{fit}})/\varepsilon > C_{\mathrm{fit}}$ and a feasible rule has fit-accuracy suboptimality at most ξ, the proof gives $t \le \xi/[2(d - C_{\mathrm{fit}})]$ for its optimal affine certificate. For two returned probability tables, $\max_z |g_{\mathrm{full}}(z) - g_{\mathrm{mix}}(z)| \le \kappa$ directly bounds their absolute expected-accuracy difference by κ on any evaluation law. The accompanying rule-checking utility implements this comparison for supplied fitted-rule records, allowing agreement to be stated at the solver's numerical precision.

Case-block uncertainty protocol. For subsequent verification, we specify a paired percentile bootstrap with 10,000 resamples. The resampling unit is the case: both cue variants, all repeated runs, and all compared methods remain together. For each resample, recompute the signed accuracy difference between the fixed fitted rules; its 2.5th and 97.5th bootstrap percentiles define a nominal 95% interval. For a fixed task/template schedule, resample cases within its task–wording cells to preserve the design weights. Rules are not refitted, so the target is evaluation uncertainty conditional on the fitted rules and the schedule. This protocol has not been executed on matching case-level records for the current comparisons; no interval endpoints are inferred from its specification. Simultaneous claims would additionally require a declared comparison family and error allocation.

I.2 Five-agent follow-up with held-out wording families

Protocol. The original three models are joined by Gemma-2-9B-IT and Phi-3.5-mini-instruct (Google, 2024; Microsoft, 2024). Every agent receives the full 150–220-token narrative covering

employment history, financial conduct, and references; no integer score fields are exposed. A fixed rubric assigns the label from the generating latent state before inference. Communication follows V2 with rationales capped at 96 tokens. The generator seed is 20260921, with 1,024 fit and 2,048 evaluation paired cases per cue condition and $R = 8$ runs per case/variant. Eight fit wording families and sixteen evaluation families are disjoint. The explicit condition states sex directly. The implicit condition conveys it through first name, pronouns, and an affiliation proxy, without a separate sex statement; this is a lexical/proxy-cue intervention, not removal of all gender information. Theorem 2 gives majority's structural vulnerability $\mathcal{V}_{.05}(h_3) = 0.750$ for $m = 5$.

Coordinator	ρ	Accuracy	Gap	$\mathcal{V}_{.05}$
Majority (Dietterich, 2000)	–	0.7462	0.0731	0.750
Best single (§ 5.1)	–	0.7139	–	0.050
Random agent (§ 5.1)	–	0.7016	–	0.050
Constant (§ 5.1)	–	0.5943	0.0000	0.000
Empirical ($\tau = 0.02$) (Agarwal et al., 2018, adapt.)	–	0.7431	0.0402	0.606
Fitted mixture (§ 5.1)	0.10	0.7162	0.0158	0.100
JOINTREPAIR (ours)	0.10	0.7311	0.0179	0.100
Fitted mixture (§ 5.1)	0.20	0.7208	0.0246	0.200
JOINTREPAIR (ours)	0.20	0.7395	0.0281	0.200
Analytic benchmark				
Selected-agent benchmark†	0.05	0.7139	–	0.050

Table 5: **Five-agent explicit-cue comparison.** $\varepsilon = 0.05$; 2,048 evaluation pairs. †: the conditional selected-agent benchmark at a tight budget. Dashes denote omitted local-gap entries. Fitted structural values follow the aggregate record; all empirical comparisons are descriptive.

Agent–constant gap bounds. For any agent–constant mixture $q(\mathbf{z}) = w_0 + \sum_i w_i z_i$ with $w_i \geq 0$, its observed gap obeys

$$\widehat{\Delta}_q = \left| \sum_i w_i (\widehat{p}_{i,1} - \widehat{p}_{i,0}) \right| \leq \sum_i w_i \widehat{\delta}_i.$$

Uniform random-agent voting gives the mean local-gap bound, and the tight-budget benchmark gives $(\rho/\varepsilon)\widehat{\delta}_{j^*}$. These identities describe the required calculation without assigning numerical bounds from unverified local-rate entries.

Complete paired comparisons. Table 6 reports the five larger-budget paired contrasts. The proxy-cue condition is reported at the contrast level. Evaluation on the held-out wording families concerns those selected families; the proposed case resampling conditions on that selection. The five-agent design also changes the models, input evidence, and rationale length, so differences from the three-agent study cannot identify the contribution of any one change.

Condition	ρ	Observed difference (pp)
Scored, $m = 3$, explicit	0.10	+0.74
Scored, $m = 3$, explicit	0.20	+0.77
Open, $m = 5$, explicit	0.10	+1.49
Open, $m = 5$, explicit	0.20	+1.87
Open, $m = 5$, proxy cues	0.10	+1.02

Table 6: **Observed full-minus-mixture accuracy differences.** Entries compare fixed fitted rules on the same evaluation cases and do not estimate population class-optimum differences.

J EXACT-LAW AND IMPLEMENTATION VERIFICATION

Verification role. These finite-law checks assess sharpness, implementation, and a specified dependence shift; they are separate from evidence about real or generated task accuracy. Independent primal and dual solvers match 251 threshold/tolerance settings with $m = 1, \ldots, 8$ within 10^{-7},

alongside 56 random tables with heterogeneous tolerances. The 66 additional moment LPs and their full hierarchy appear in Appendix L.

A constructed dependence shift. For five agents, take both fitted groups to be uniform on the cube, so empirical fairness retains majority. Shift to $P_1 = \frac{5}{6}U_3 + \frac{1}{6}U_0$ and $P_0 = \frac{5}{6}U_2 + \frac{1}{6}U_5$, where U_j is uniform on the j-ones layer. Each agent retains rate $1/2$, yet majority's gap becomes $2/3$. At $\varepsilon = 0.02, \rho = 0.10$, count-only repair retains fidelity 0.737, versus 0.724 for the matched mixture; at $\rho = 0.20$, these values are 0.798 and 0.770. This is an exact-law fidelity comparison, not an additional application-accuracy result.

Coordinator	ρ	Fit fidelity	Shift gap	$\mathcal{V}_{.02}$
Majority / empirical (Dietterich, 2000; Agarwal et al., 2018)	–	1.000	0.667	0.700
Best affine (§ 5.1)	–	0.688	0.000	0.020
Constant (§ 5.1)	–	0.500	0.000	0.000
Optimal mixture (§ 5.1)	0.05	0.701	0.029	0.050
JointRepair (ours)	0.05	0.706	0.029	0.050
Optimal mixture (§ 5.1)	0.10	0.724	0.078	0.100
JointRepair (ours)	0.10	0.737	0.078	0.100
Optimal mixture (§ 5.1)	0.20	0.770	0.176	0.200
JointRepair (ours)	0.20	0.798	0.176	0.200
Optimal mixture (§ 5.1)	0.40	0.862	0.373	0.400
JointRepair (ours)	0.40	0.890	0.373	0.400

Table 7: **Exact-law dependence shift, $m = 5$.** Fit fidelity is agreement with the original majority, not labeled accuracy. Both groups have agent rates $1/2$ before and after the shift. The mixture and robust rows use the same $\varepsilon = 0.02$ and ρ. All entries are exact-law/LP quantities.

The executable `run_verified.py` checks all 35 thresholds for $m = 2, \ldots, 8$ at zero tolerance and 56 random tables with alternating zero/heterogeneous tolerances. The follow-up `revision_analysis.py` checks all thresholds for $m = 1, \ldots, 8$ over the deduplicated tolerance set $\{0, 0.01, 0.05, 0.1, 1/m, 0.4, 1\}$, giving 251 cases. Independent primal and dual values match the closed form within 10^{-7}. Constant tables test the normalization-multiplier boundary. These checks complement the proofs; they do not replace them.

For the archived tabular evaluation, verification independently solves the primal for all 600 robust fitted rules (two datasets, ten seeds, three local tolerances, five budgets, and two function classes), checks disjoint split IDs, and reconstructs metrics from per-subject response probabilities. The follow-up adds 100 fitted mixture rules under the same five budgets and verifies their primal feasibility. Complete U-statistic values are compared with brute-force tuples. Earlier decision-only pilot caches are retained in the code archive but are not part of the larger-model evaluation reported here. The larger-model tables use the aggregate result record; no exact transition-law integration is claimed for the rationale study.

K Statistical Experiments and Computational Scaling

Across 2,400 synthetic datasets, complete cross-run U-statistic coverage ranges from 0.927 to 0.967, with Monte Carlo standard error at most 0.0151. Plug-in bias persists as cases grow at fixed runs. These checks assess the implementation under its known generator and do not calibrate coverage for the larger-model study.

Complete recombination. The main audit implementation supports incomplete injective tuples. For the coverage experiments we compute the complete U-statistic without enumerating all $(R)_m$ tuples. Expand $h(z) = \sum_{S \subseteq [m]} c_S \prod_{i \in S} z_i$, with coefficients obtained by a Boolean-lattice Möbius transform. Let $F_r(S)$ sum $\prod_{i \in S} Z_i^{(f(i))}$ over injections $f : S \to \{1, \ldots, r\}$. Initialize $F_0(\varnothing) = 1$, and update

$$F_r(S) = F_{r-1}(S) + \sum_{i \in S} Z_i^{(r)} F_{r-1}(S \setminus \{i\}).$$

The desired product reference is $\sum_S c_S F_R(S)/(R)_{|S|}$. Each subset monomial has this same expectation within the full m-coordinate injection average, so the computation equals Equation (18). Complexity is $O(Rm2^m)$ per case, with $O(2^m)$ working storage. `run_statistical.py` checks it against explicit tuple enumeration for $(m, R) = (2, 3), (3, 4), (4, 5)$; maximum discrepancy is 1.12×10^{-16}. This acceleration still requires moderate m and does not remove exponential dependence on agent count.

Coverage design. For each case draw $p \sim \mathrm{Uniform}(0.1, 0.5)$. In group a, the action vector is a synchronized Bernoulli(p) vector with probability λ_a, otherwise independent Bernoulli(p) coordinates; $(\lambda_0, \lambda_1) = (0.1, 0.7)$. Independent randomness is used across cases and runs. Common random numbers couple the two group variants within a case. Both groups have identical case-specific marginal rates. For majority, the target residual is

$$D_{\mathrm{dep}} = 0.6\,\mathbb{E}_p[p - \bar{h}(p\mathbf{1})].$$

Numerical polynomial integration gives 0.0408 for $m = 3$ and 0.063144 for $m = 5$. We run 300 independent datasets at each of eight (m, n, R) settings, with 399 case-block bootstrap replicates per dataset and estimator. The percentile interval is nominally 95%; bootstrap indices are shared between estimators for comparison. The complete estimator removes recombination Monte Carlo error. Simulation seeds and per-dataset estimates, inclusion indicators, and interval widths are retained.

m	n	R	Plug-in bias	U bias	Plug-in coverage	U coverage
3	200	3	-0.0317	-0.0003	0.010	0.967
3	200	32	-0.0038	-0.0000	0.703	0.933
3	800	3	-0.0312	0.0010	0.000	0.927
3	800	32	-0.0038	-0.0001	0.220	0.937
5	200	5	-0.0346	-0.0006	0.013	0.953
5	200	32	-0.0064	0.0003	0.583	0.937
5	800	5	-0.0352	-0.0006	0.000	0.950
5	800	32	-0.0068	-0.0001	0.053	0.927

Table 8: Bias and empirical coverage over 300 datasets per row. Coverage is a Monte Carlo outcome, not a theorem. The maximum coverage standard error for the U-statistic is about 0.015; values around the nominal level should be interpreted with that simulation uncertainty.

Bias and its Monte Carlo standard error are retained for every setting (Table 8). U-statistic coverage ranges from 0.927 to 0.967. The plug-in estimator can become more confidently wrong as n grows at fixed R. The bounded-case empirical-Bernstein interval for the residual uses half-range 2 and covers the target in every simulated dataset, illustrating its conservatism; this does not establish universal empirical coverage. We do not interpret 300/300 coverage as a probability of one.

Count-only computational scaling. Using the reduced LP, we fit monotone count-only rules to majority fidelity under $N \sim \mathrm{Binomial}(m, 1/2)$, with $\rho = 0.10$, $\varepsilon = 0.02$, and no empirical fairness constraint. Five consecutive solves per size include construction and solver time; Table 9 gives medians. The numerical runs set BLAS/OpenMP thread counts to one; environment versions and individual timings are saved. Timing is illustrative, not a comparison of optimized solvers. The implementation uses dense constraint storage, so its practical memory growth need not be linear despite the $O(m)$ variable and inequality counts. All $m + 1$ output probabilities and individual times are saved. These are coordinator solves, not LLM-panel scaling experiments.

m	Median seconds	Fit fidelity
3	0.0011	0.7892
5	0.0011	0.7365
9	0.0011	0.6864
19	0.0013	0.6392
49	0.0017	0.6010
99	0.0024	0.5826
199	0.0045	0.5698
499	0.0169	0.5585
999	0.0500	0.5528

Table 9: Executed count-only LP scaling; fidelity is expected agreement with majority, not labeled-task accuracy.

L Structural Consequences and Additional Audited Moments

L.1 Proof of Corollary 2

For $m \geq 2$, $\max(k, m-k+1) \geq 2$, so Theorem 2 gives $\mathcal{V}_0(h_k) \geq 1/2$. Now let g be monotone and select an optimal envelope $\ell(\mathbf{z}) = a + b^\top \mathbf{z}$ with $\|g - \ell\|_\infty \leq t$ and $2t + \sum_i \varepsilon_i |b_i| \leq \rho$. We can choose $b_i \geq 0$. To see this, if $b_i < 0$, replace b_i by zero and a by $a + b_i/2$. For each fixed $\mathbf{z}_{-i}$ the two old affine values are decreasing, whereas the two values of g are increasing. Both old affine values lie in the intersection $[g(1, \mathbf{z}_{-i}) - t,\ g(0, \mathbf{z}_{-i}) + t]$, so their midpoint remains within t of both target values. This replacement does not increase the approximation error and reduces the slope penalty. Repeat for every negative slope.

The resulting ℓ lies in $[-t, 1+t]$. Hence

$$q(\mathbf{z}) = \frac{\ell(\mathbf{z}) + t}{1 + 2t}$$

is monotone, affine, and probability-valued. It is a convex mixture of constants and single-agent rules by Appendix E. Moreover,

$$|g - q| = \frac{|g - \ell + t(2g-1)|}{1+2t} \leq \frac{2t}{1+2t} \leq \frac{\rho}{1+\rho}.$$

The same construction preserves the robustness budget:

$$\mathcal{V}_{\boldsymbol{\varepsilon}}(q) = \frac{\sum_i \varepsilon_i b_i}{1+2t} \leq \frac{\rho - 2t}{1+2t} \leq \rho.$$

Here an affine rule's vulnerability is the weighted absolute slope sum, since the admissible marginal differences can simultaneously attain their extremal signs. At exact local parity, $\mathcal{V}_0(g) = 0$ if and only if g is affine. This proves the exact statement at $\rho = 0$. The $\rho/2$ distance from Theorem 1 is to unrestricted affine functions; it is not automatically a distance to valid probability-valued mixtures.

L.2 An accuracy ceiling for nonlinear robust coordination

The affine rule in Corollary 2 satisfies the same structural budget as the original coordinator. Consequently, proximity to mixtures also bounds the utility that additional action-pattern flexibility can retain.

Proposition 5 (Utility advantage at a matched robustness budget)**.** *Fix a distribution of action vectors and a linear utility $U(g) = \mathbb{E}[g(Z)u_1 + (1 - g(Z))u_0]$, where $|u_1 - u_0| \leq L$ almost surely. Let $\mathcal{G}_{\boldsymbol{\varepsilon},\rho}$ be all monotone, probability-valued coordinators with $\mathcal{V}_{\boldsymbol{\varepsilon}}(g) \leq \rho$, and let $\mathcal{A}_{\boldsymbol{\varepsilon},\rho}$ be its affine members. Then*

$$0 \leq \max_{g \in \mathcal{G}_{\boldsymbol{\varepsilon},\rho}} U(g) - \max_{q \in \mathcal{A}_{\boldsymbol{\varepsilon},\rho}} U(q) \leq L \frac{\rho}{1+\rho}. \tag{28}$$

The same upper bound applies when the affine baseline is enlarged to the matched mixture of a fixed monotone rule, single agents, and constants. For ordinary accuracy, $L = 1$.

Proof. For any feasible g, the proof of Corollary 2 supplies $\ell = a + b^\top z$ with $b_i \geq 0$, $\|g - \ell\|_\infty \leq t$, and $2t + \sum_i \varepsilon_i b_i \leq \rho$. Its probability-valued affine rule $q = (\ell + t)/(1 + 2t)$ satisfies

$$\mathcal{V}_{\boldsymbol{\varepsilon}}(q) = \frac{\sum_i \varepsilon_i b_i}{1 + 2t} \leq \frac{\rho - 2t}{1 + 2t} \leq \rho, \qquad \|g - q\|_\infty \leq \frac{2t}{1 + 2t} \leq \frac{\rho}{1 + \rho}. \tag{29}$$

The equality follows because the vulnerability of an affine rule is its weighted absolute slope sum: the marginal differences can attain the desired signs simultaneously. Hence $q \in \mathcal{A}_{\boldsymbol{\varepsilon},\rho}$ and $U(g) - U(q) \leq \mathbb{E}[|u_1 - u_0||g - q|] \leq L\rho/(1 + \rho)$. Optimize over g and use the inclusions of the affine, matched-mixture, and full monotone classes. □

The conclusion compares optima on the same utility distribution and with the same structural constraint. It does not bound the difference between two independently estimated test accuracies, nor does it automatically preserve an additional empirical-disparity constraint. For a returned certificate, $2Lt/(1 + 2t)$ is the more informative rule-specific ceiling.

Proposition 3 identifies a tight-budget regime in which nonlinear flexibility has exactly zero optimal-accuracy value. Its best agent is defined on the specified optimization law and need not be the agent selected on a different fitting distribution.

Proof of Proposition 3. Use a nonnegative-slope envelope $\ell = a + b^\top z$ as above and set $B = \sum_i b_i$. Probability-valuedness of g gives $-t \leq a \leq 1 + t - B$. Let $\pi = \Pr(Y = 1)$ and $A_i = \Pr(Z_i = Y)$. Since accuracy is 1-Lipschitz in the output probability,

$$U(g) \leq (1 - \pi) + a(2\pi - 1) + \sum_i b_i\big(A_i - (1 - \pi)\big) + t.$$

If $\pi \geq 1/2$, substitute $a \leq 1 + t - B$; otherwise substitute $a \geq -t$. Both cases yield $U(g) \leq C + 2Ct + B(A_{\max} - C)$. The certificate $2t + \varepsilon B \leq \rho$ proves Equation (15). Under the additional assumptions, its larger coefficient is $(A_{\max} - C)/\varepsilon$, and the claimed mixture is a valid monotone affine coordinator with vulnerability exactly ρ and the stated accuracy. This attains the upper bound.

For the equality conditions, keep $0 \leq \rho \leq \varepsilon$ and let $d = (A_{\max} - C)/\varepsilon > C$. The same inequalities give

$$U(g) \leq C + d\rho - 2t(d - C).$$

Every exact optimum therefore has $t = 0$ and $B = \rho/\varepsilon$. It is a valid monotone affine mixture, and positive agent weights can only be assigned to agents attaining $A_{\max}$. At $\rho = \varepsilon$, $B = 1$ forces the intercept to zero; if the best agent j is unique, $g(z) = z_j$ on the entire cube. For $0 < \rho < \varepsilon$, uniqueness of the best constant also fixes the remaining mixture weight. These conclusions apply when the fixed law is the empirical fit distribution. They concern the learned functions, not just equal fit objective values. □

L.3 An exact hierarchy of moment uncertainty sets

Let $\mathcal{X}$ be a finite state space, $h : \mathcal{X} \to [0, 1]$, and $\phi : \mathcal{X} \to \mathbb{R}^q$ a fixed feature map. Define

$$\mathcal{U}_\eta^\phi = \{(P_0, P_1) : |\mathbb{E}_{P_1}\phi_j - \mathbb{E}_{P_0}\phi_j| \leq \eta_j,\ j = 1, \ldots, q\}.$$

Proposition 6 (Finite moment-constrained disparity)**.** *For nonnegative η, the maximum absolute disparity over $\mathcal{U}_\eta^\phi$ equals Equation (9); primal and dual optima are attained. Replacing h by a fitted rule g gives an exact robust-repair LP with $|\mathcal{X}|$ rule variables and an envelope in $1, \phi_1, \ldots, \phi_q$.*

Proof. The primal feasible set is compact, nonempty, and closed under exchanging the groups. Its signed objective is therefore sufficient. Attach nonnegative multipliers to the two inequalities for each feature and free multipliers to the probability normalizations. Eliminating the latter gives $\min_\theta\{\mathrm{osc}(h - \theta^\top\phi) + \sum_j \eta_j|\theta_j|\}$. Centering the oscillation gives Equation (9). Finite LP duality supplies equality and attainment, exactly as in Appendix A. □

For binary actions choose all monomials $\phi_S(\mathbf{z}) = \prod_{i \in S} z_i$ with $1 \leq |S| \leq d$. When all differences vanish, the value is twice the best degree-d uniform approximation error. The uncertainty sets are

nested, so their values are nonincreasing in d; at $d = m$ the moments identify the entire law and the value is zero. With nonzero tolerances, the feature coefficients receive the corresponding weighted absolute-value penalty. This is a finite-state moment specialization, not a new general moment-duality theorem.

Raw moments versus covariance. Bounds on raw second-moment differences are linear in the two probability tables. If both marginal vectors p_0, p_1 are fixed, a covariance-difference constraint is also linear, since

$$\Delta \operatorname{Cov}_{ij} = \Delta\mathbb{E}[Z_i Z_j] - (p_{i,1}p_{j,1} - p_{i,0}p_{j,0})$$

has a known offset. When only marginal-gap bounds are known, the product offset varies with the optimization variables, so a general covariance constraint is not an LP constraint. Under exact local parity its offset cancels, and equal pairwise covariances are equivalent to equal raw second moments.

Proof of Proposition 1. Any quadratic approximant p can first be made multilinear on the cube. Define $p_*(\mathbf{z}) = [p(\mathbf{z}) + 1 - p(\mathbf{1} - \mathbf{z})]/2$. The complement identity gives $\|h - p_*\|_\infty \le \|h - p\|_\infty$. In $p(\mathbf{z}) - p(\mathbf{1} - \mathbf{z})$ every quadratic term cancels, so p_* is affine. Affine polynomials are a subclass of quadratics, giving equality of the two best errors. Apply Proposition 6 and Theorem 2. □

Coordinator ($m = 5$)	Degree 1	Degree 2	Degree 3	Degree 4	Degree 5
1-of-5	0.8000	0.4444	0.2000	0.0625	0.0000
2-of-5	0.7500	0.5625	0.4667	0.2500	0.0000
3-of-5	0.6667	0.6667	0.3750	0.3750	0.0000
4-of-5	0.7500	0.5625	0.4667	0.2500	0.0000
5-of-5	0.8000	0.4444	0.2000	0.0625	0.0000

Table 10: **Sharp gaps when all moments through degree d match exactly.** Independent primal/dual finite LP calculations. Degree 1 is exact local parity; degree 2 adds all pairwise moment equalities. Values are rounded to four decimals.

The full computation includes $m \in \{3, 5\}$, every threshold, every degree, and a separate degree-two grid with local tolerance 0.02 and pairwise tolerances $0, 0.02, 0.10, 1$: 66 cases in total. Primal witnesses satisfy the stated constraints, independent dual objectives agree within 10^{-8}, the zero-tolerance hierarchy is nonincreasing, and the full-degree values vanish. These are executed finite-law calculations; no language-model measurements enter them.

L.4 Retaining information in marginal reports

The global guarantee permits both dependence changes and changes in common selection rates. When deployment rates are known to remain near audited reports, the same finite-state optimization can retain that information. This localizes the uncertainty set without requiring independence.

Let $\mathcal{K} \subseteq [0,1]^{2m}$ be a nonempty polytope of permitted pairs (p_0, p_1), and define

$$\mathcal{V}_{\mathcal{K}}(h) = \max_{P_0, P_1 :\, (\mathbb{E}_{P_0} Z, \mathbb{E}_{P_1} Z) \in \mathcal{K}} |\mathbb{E}_{P_1} h - \mathbb{E}_{P_0} h|, \qquad \sigma_{\mathcal{K}}(b_0, b_1) = \max_{(p_0,p_1)\in\mathcal{K}} (b_0^\top p_0 + b_1^\top p_1).$$

For example, $\mathcal{K}$ may impose group-specific intervals $\underline{p}_a \le p_a \le \overline{p}_a$ together with $|p_1 - p_0| \le \varepsilon$. Fixed reports are singleton polytopes.

Proposition 7 (Report-aware exact characterization)**.** *For $s \in \{-1, 1\}$, let*

$$W_s(h; \mathcal{K}) = \min_{b_0, b_1 \in \mathbb{R}^m} \Big\{ \max_z [-sh(z) - b_0^\top z] + \max_z [sh(z) - b_1^\top z] + \sigma_{\mathcal{K}}(b_0, b_1) \Big\}. \tag{30}$$

Then $\mathcal{V}_{\mathcal{K}}(h) = \max\{W_+(h; \mathcal{K}), W_-(h; \mathcal{K})\}$, and all optima are attained. If $\mathcal{K}$ is invariant under exchanging p_0 and p_1, the two signed optima coincide.

Proof. Introduce $p_a = \mathbb{E}_{P_a} Z$ in the finite LP maximizing $s(\mathbb{E}_{P_1} h - \mathbb{E}_{P_0} h)$ with $(p_0, p_1) \in \mathcal{K}$. Dual variables b_a for these moment equalities yield the two envelopes in Equation (30); maximizing over

the allowed reports yields their support function. The primal is nonempty because every vector in $[0,1]^m$ is realizable by a product law. It is compact, and finite LP duality establishes equality and attainment. Maximizing the two signs gives the absolute disparity. Swapping the two laws proves the final assertion. □

An exact repair LP. Write $\mathcal{K} = \{p : Ap \le d\}$, including the cube bounds, and partition $A = [A_0\ A_1]$ according to the two marginal vectors. The support-function dual is $\sigma_{\mathcal{K}}(b_0, b_1) = \min_{y \ge 0:\, A^\top y = (b_0, b_1)} d^\top y$. Thus $\mathcal{V}_{\mathcal{K}}(g) \le \rho$ is equivalent to the following linear constraints for each $s \in \{-1, 1\}$:

$$\begin{aligned} &y_s \ge 0, \quad t_{0,s}, t_{1,s} \in \mathbb{R}, \quad 0 \le g(z) \le 1, \\ &t_{0,s} + (A_0^\top y_s)^\top z \ge -s g(z), \quad t_{1,s} + (A_1^\top y_s)^\top z \ge s g(z) \quad \text{for every } z, \\ &t_{0,s} + t_{1,s} + d^\top y_s \le \rho. \end{aligned} \tag{31}$$

For an exchange-symmetric set, one sign suffices.

Relation to the global guarantee. The singleton choice recovers the fixed-report expression $\max\{H^+(p_1) - H^-(p_0), H^+(p_0) - H^-(p_1)\}$. For the global set $\mathcal{K}_{\boldsymbol{\varepsilon}} = \{(p_0, p_1) \in [0,1]^{2m} : |p_1 - p_0| \le \boldsymbol{\varepsilon}\}$, the slopes can be chosen as $(b_0, b_1) = (-b, b)$, and $\sigma_{\mathcal{K}_{\boldsymbol{\varepsilon}}}(-b, b) = \sum_i \varepsilon_i |b_i|$. The two envelopes combine into an oscillation, recovering Theorem 1. If $\mathcal{K} \subseteq \mathcal{K}'$, then $\mathcal{V}_{\mathcal{K}}(h) \le \mathcal{V}_{\mathcal{K}'}(h)$; at the same budget and utility objective, the smaller uncertainty set therefore has weakly higher optimal utility. This states explicitly the tradeoff between retaining audited rate information and protecting against broader rate drift. Interval reports must describe the deployment rates for this narrower guarantee to apply; a fitting estimate alone does not provide that premise.

L.5 Other complexity measures and finite outputs

Influence and noise sensitivity need a reference law. Ordinary uniform-cube influence does not control our distribution-free disparity in a dimension-free way that vanishes with influence. For AND, total influence is $m/2^{m-1} \to 0$, while $\mathcal{V}_0 = 1 - 1/m \to 1$. Under the usual uniform-marginal noise coupling, its noise sensitivity is at most $2\mathbb{P}(h = 1) = 2^{1-m}$, yet the same worst-case disparity tends to one. A useful influence-based bound would therefore need additional restrictions on the compared laws or a different complexity measure.

Multiple output categories. For probabilities $h_1(\mathbf{z}), \dots, h_L(\mathbf{z})$ summing to one, apply the scalar theorem to each output event $S \subseteq [L]$, using $h_S = \sum_{j \in S} h_j$. The exact worst-case total-variation distance between output distributions is $\max_{S \subseteq [L]} \mathcal{V}_{\boldsymbol{\varepsilon}}(h_S)$; the two finite maxima commute. Per-category parity uses only singleton events. Joint repair keeps the simplex constraints and each chosen event's linear envelope constraints. The number of events can be exponential in L.

Bounded cohorts. Proposition 6 holds on any finite state space, including a bounded cohort's joint action table. One must specify which cohort features are audited and which output events define disparity. Probabilities over feasible allocations can impose a capacity constraint, but their state/output tables grow combinatorially. A single-subject marginal audit is not automatically a cohort fairness constraint.

When count-only optimization is lossless. If the feasible set and the fitted linear objective are invariant under every permutation of agents, averaging an optimizer over permutations preserves feasibility and objective value, producing a count-only optimum. It is enough for the final objective coefficient on each action cell to depend only on its count; individual components of the fitted utility need not separately be symmetric. Without that invariance, count-only optimization is guaranteed exact only within its restricted class. Equality with the full optimum can still be checked by solving both LPs, but does not follow solely from a symmetric fairness constraint.