CAN LINKED BINARY FEEDBACK CERTIFY RANKING DISAGREEMENT?

Anonymous authors
Paper under double-blind review

ABSTRACT

Several comparisons from the same evaluator reveal response patterns that pooled labels discard. Do these patterns establish different rankings, or can they reflect differences in decisiveness and dependence between answers? Under a logistic response model, every fixed finite linked-binary protocol on at least three stimuli admits a shared-ranking population and a population with a fixed positive fraction of ranking reversals that generate exactly the same joint responses. One distribution of decisiveness serves all comparison bundles. This rules out a uniform certification guarantee from finite linkage alone. We characterize when a fixed shared-ranking direction imposes testable restrictions on the joint response law. In a three-stimulus benchmark, near ties reveal a signal quadratic in ranking heterogeneity, but dependence of the same order can exactly reproduce the complete repeated-comparison law under a shared ranking that permits ties. For a calibrated benchmark with a known dependence budget, matching detection bounds separate the effects of sampling noise and residual dependence. Computations corroborate these transitions. A controlled study of language-model judges illustrates the diagnostic: all three tested models, each using a fixed six-persona roster, violate necessary shared-order conditions on a preselected narrative-rhythm triple at simultaneous 95% confidence. The results connect informative comparison designs to the dependence calibration needed for ranking-disagreement certificates.

1 INTRODUCTION

Preference-learning pipelines use binary judgments to fit reward models or optimize language-model policies (Rafailov et al., 2023; Chidambaram et al., 2026). Retaining several judgments from each evaluator appears to offer a way to distinguish stable preferences from response noise. Suppose one evaluator compares A with B and A with C. Their joint answers reveal associations that pooled labels erase. Yet such associations have several explanations: evaluators may rank the items differently, share a ranking but differ in decisiveness, or carry an earlier answer into a later one. Which joint response patterns can rule out every shared-ranking explanation?

We study this question for a fixed finite collection of linked binary comparisons. The running example uses $e = (A, B)$ and $f = (A, C)$: the bundle $B_2 = (e, f)$ contains an anchor and a probe, and $B_3 = (e, f, f')$ adds a second occasion for the identical probe. The observable is the probability of each of the four or eight possible answer patterns. Evaluator decisiveness is an unknown positive scale on utilities. Multiplying every utility by that scale changes how reliably an evaluator chooses, while preserving the ranking.

Our first result shows that even the complete joint response distribution may leave ranking disagreement uncertified. For every fixed finite library on at least three stimuli, we construct two populations with identical observable laws: one shares a strict ranking and differs only in decisiveness; the other has a fixed positive fraction of ranking reversals. The construction matches all bundles with one decisiveness distribution. We call this local equivalence *Temperature Camouflage*. It extends known local unrestrictedness of binary latent-trait models (Haberman, 2005, Theorems 1–2) to comparisons constrained by one global utility vector, repeated edges, cycles, and a common mixing law across bundles.

Our second result identifies where this equivalence can break. Some designs and utility directions impose additional linear restrictions on the joint answers. We compute these restrictions through

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.

exponent collisions and the rank they remove. A three-stimulus example makes the signal concrete: near a tie between A and B, their average comparison can cancel while its association with the A–C probe remains. Under conditional independence, a shared ordering restricts that association. For the symmetric benchmark, the surviving signal is quadratic in the size of the ranking perturbation; a repeated probe calibrates its coefficient.

Our third result shows that this signal is also the scale at which dependence matters. A legal dependent response mechanism under a shared order that permits ties can reproduce the complete repeated-probe law, while preserving each comparison's logistic response probability. We prove that the smallest required dependence budget is quadratic, and obtain matching detection bounds when the budget is supplied for the calibrated near-tie class. This connects the information provided by a comparison design to the strength of the behavioural assumptions needed to interpret it.

The matched populations provide identically distributed data to a learner, limiting how its fitted types or policies can be interpreted (Section 8). Computations examine both predicted transitions. A fresh LLM-judge study illustrates successful pointwise certification: all three tested models separate the shared-order necessary set on one preselected narrative-rhythm triple, using fixed persona rosters and simultaneous 95% confidence.

2 Setup and Observation Regimes

Let $[N] = \{1, \dots, N\}$ index stimuli and let an evaluator have latent utilities $V = (V_1, \dots, V_N)$. For an oriented edge $e = (i, j)$, a binary judgment $Y_e \in \{0, 1\}$ chooses i over j; we use the signed response $D_e = 2Y_e - 1$. The conditional marginal model is

$$\mathbb{E}[D_e \mid V] = m_e(V) := \tanh\{(V_i - V_j)/2\}. \tag{1}$$

The scale of V controls decisiveness without changing its ordering. This is the binary logistic response induced by random-utility choice (McFadden, 1974); the inferential target below is weaker than recovering the utility distribution.

A *bundle* $B = (e_1, \dots, e_m)$ is a predeclared list of response occasions assigned to one evaluator. Its observable is the complete labelled law of $D_B = (D_{e_1}, \dots, D_{e_m})$, including every one of its 2^m cells. A finite library $\mathfrak{B}$ may contain several bundle types. Each of the K evaluators is fresh, receives one bundle, and evaluators are iid. Thus K counts evaluators (or bundles), while the judgment count is $M = \sum_{k=1}^{K} |B_k|$. Reusing a panel or identifying one persistent service instance is a different observation regime.

Under a shared ordering, evaluators agree on the relative order of all items, with ties allowed. The closed cone for permutation π is $\mathcal{C}_\pi = \{v : v_{\pi(1)} \geq \cdots \geq v_{\pi(N)}\}$. Our null and its order-reversal distance are

$$H_0^{\mathrm{ord}} = \{F : \exists \pi,\ F(\mathcal{C}_\pi) = 1\}, \qquad \rho_{\mathrm{ord}}(F) = \inf_{\pi} F(V \notin \mathcal{C}_\pi). \tag{2}$$

The positive-temperature submodel writes $V = As$, where s is one common direction, $A > 0$, and the same $G = \mathcal{L}(A)$ applies to every block in $\mathfrak{B}$. It lies inside H_0^{ord} whenever s has a fixed ordering.

Except in Section 5, response occasions are conditionally independent given V. Section 5 instead preserves each marginal mean in equation 1 but permits centred conditional interaction errors. Their externally supplied bound is the *dependence budget* b, which specifies the allowed departure from factorization for sensitivity analysis.

3 Temperature Camouflage

Each value of decisiveness produces a vector of probabilities for all bundle outcomes. Mixing decisiveness values takes convex combinations of these vectors. Although decisiveness is a scalar, its probability curve can span every observable direction available to a heterogeneous population. We extend the local unrestrictedness of logistic item models (Haberman, 2005, Theorems 1–2) to linked comparisons, where one global potential generates the logits and one mixing law serves all blocks.

Theorem 3.1 (Finite-library common-G exact saturation)**.** *For every fixed finite static linked-binary library $\mathfrak{B}$, there is a strict shared-order direction s (selectable from an open dense, full-measure set), positive supports $a_0, \dots, a_d$, and strictly positive weights $w_0, \dots, w_d$. Let $G_0 = \sum_r w_r \delta_{a_r}$ and let p_0 be its stacked bundle law. A relative neighbourhood $\mathcal{U}$ of p_0 in the common-law affine space has the following property: for every population law Q inducing a stacked law in $\mathcal{U}$, there is a mixing law G_Q on the same positive supports, common to all $B \in \mathfrak{B}$, such that*

$$\big(\mathcal{L}_Q(D_B)\big)_{B\in\mathfrak{B}} = \big(\mathcal{L}_{V=As,\, A\sim G_Q}(D_B)\big)_{B\in\mathfrak{B}}. \tag{3}$$

Equality holds cell by cell for the complete within-evaluator joint laws.

The proof establishes equality of two affine spans: that of all population response laws and that obtained by varying decisiveness along one direction s. An incidence–convolution matrix accounts for the identities forced by shared items, repeats, cycles, and the common mixing law. At a generic s, the two spans coincide. Selecting finitely many positive scales then forms a full-dimensional simplex. Around an interior mixture, changing its weights matches every nearby population law exactly (Figure 1). The matrix construction and proof are in Appendices A.1–A.2.

Corollary 3.2 (Fixed-positive reversal hard instances)**.** *If $N \geq 3$, then for every fixed $\mathfrak{B}$ there is a library-dependent $\rho_* > 0$ such that, for each fixed $0 < \rho < \rho_*$, an order-changing population F_ρ with $\rho_{\mathrm{ord}}(F_\rho) = \rho$ and a common positive-temperature law G_ρ induce exactly the same stacked law. Consequently, any level-α test of H_0^{ord} has power at most α on this hard alternative.*

The reversal fraction can stay positive as more evaluators are sampled from the same protocol: additional observations estimate an already identical law. Predictable allocation within the matched library preserves equality of the successive observation kernels (Kaufmann et al., 2016). Appendix A.2 gives this argument and the reduction of finite-depth query trees under conditional independence of potential occasions.

Implication and scope. Interpretable certification must use calibration or structure beyond finite linkage alone: for every fixed finite protocol, there exists a protocol-dependent hard direction and neighbourhood. This least-favourable overlap rules out a uniform guarantee while leaving pointwise identification and global convex-hull equality open. One G_Q serves all blocks. Growing libraries, unbounded sessions, fixed panels, and path-dependent kernels require separate models.

4 Collision Geometry

A joint pattern that shared order cannot explain. Label the running example's A, B, C as items $1, 2, 3$, with anchor $e = (1, 2)$ and probe $f = (1, 3)$. Suppose the population anchor mean is $\mu_e = \mathbb{E}D_e = 0$. Under any shared ordering, all conditional anchor means have the same sign; their zero average forces them to vanish almost surely. Conditional independence then gives $\tau_{ef} = \mathbb{E}[D_e D_f] = 0$. Thus a zero anchor mean together with a nonzero anchor–probe moment rules out shared order. Opposite ranking perturbations can produce precisely this pattern: the anchor answers cancel on average, while each perturbation moves the probe in the same direction as the anchor. Near a tie, the size of the permitted moment becomes the key quantity.

From a tie to a general design criterion. For a finite library, clearing the logistic cell denominators yields a finite sum of exponentials. The incidence–convolution matrix $R_{\mathfrak{B}}$ records their integer multiplicities, with one row per exponent vector d. Along $V = As$, equal projections $s^\top d$ produce identical exponential functions: these are *exponent collisions*. Let P_s sum the colliding rows. The loss of rank counts additional affine restrictions on the ray law, using the linear-independence approach to mixture geometry (Lindsay, 1983; Allman et al., 2009).

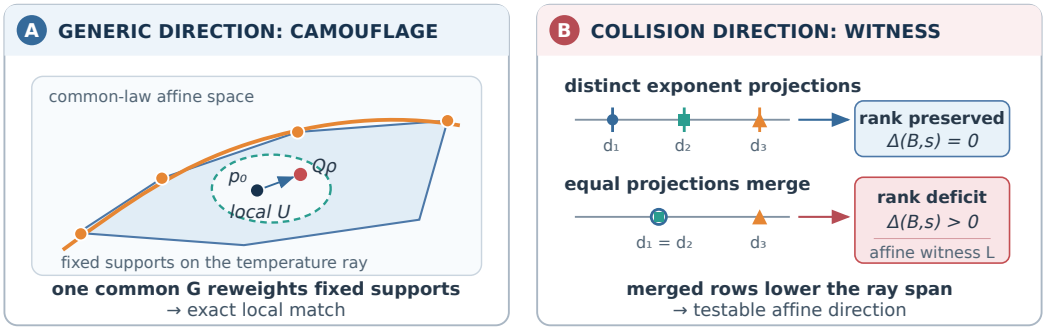


Figure 1: **Why camouflage works, and where it can fail.** At a generic direction, distinct exponents let a finite ray simplex contain a common-law neighbourhood, so one common law on fixed supports can match nearby laws by reweighting. Merged exponent rows leave a witness exactly when they create a rank deficit $\Delta > 0$.

Theorem 4.1 (Rank deficit and first non-nuisance order)**.** *For a fixed library $\mathfrak{B}$ and direction s,*

$$\Delta(\mathfrak{B}, s) = \operatorname{rank}(R_{\mathfrak{B}}) - \operatorname{rank}(P_s R_{\mathfrak{B}}) = \dim \mathcal{A}_{\text{common}} - \dim \mathcal{A}_{\text{ray}}(s). \tag{4}$$

Moreover, Δ is the dimension of the affine-witness quotient $\ker(P_s R_{\mathfrak{B}})/\ker(R_{\mathfrak{B}})$. If $\Delta = 0$, a rich ray-simplex interior point locally absorbs every common-population law. If $\Delta > 0$ and a path $Q_\varepsilon = p_0 + \sum_{k\geq 1} \varepsilon^k h_k$ first leaves the ray tangent space at $r_* = \min\{k : h_k \notin T_{\text{ray}}\} < \infty$, an affine witness L satisfies $L(Q_\varepsilon) = c\varepsilon^{r_*} + o(\varepsilon^{r_*})$, $c \neq 0$. For a fixed-s local ray-null experiment in which the alternative's in-plane projection lies in the same local ray simplex, and satisfying finite-cell interior, uniform remainder, signal-lower-bound, and sample-proportion conditions over a compact class, the conditional minimax scale is $K^{-1/(2r_*)}$.

An affine witness is a linear statistic of the outcome probabilities plus a constant that vanishes on the ray model. The quotient in the theorem removes identities that hold for every population, leaving restrictions specific to the ray. To compute it, enumerate labelled bundle cells and deduplicate their exponent sums; build $R_{\mathfrak{B}}$ from the resulting multiplicities; group rows by $s^\top d$; then compute the two ranks and a basis of $\ker(P_s R_{\mathfrak{B}})/\ker(R_{\mathfrak{B}})$. The supplied exact-rank audit implements these steps (Appendix B.1). A nonzero Δ returns candidate witnesses; the path expansion determines r_*, after which calibration and sampling cost enter the design choice. The rate statement follows the first surviving order of the signal, as in singular likelihood testing (Rotnitzky et al., 2000; Gu, 2016).

A testable near-tie benchmark. For the running example, collect the pairwise moment vector $\theta = (\mu_e, \mu_f, \tau_{ef}) = (\mathbb{E}D_e, \mathbb{E}D_f, \mathbb{E}D_eD_f)$. Define

$$\mathcal{P}_2 = \{(a,b,c) \in [0,1]^3 : \max(0, a+b-1) \leq c \leq \min(a,b)\}, \tag{5}$$

$$\mathcal{O}_2 = \bigcup_{\eta,\xi\in\{-1,1\}} \{(\eta a, \xi b, \eta\xi c) : (a,b,c) \in \mathcal{P}_2\}. \tag{6}$$

Thus $(x,y,z) \in \mathcal{O}_2$ iff $xyz \geq 0$, $|z| \leq \min(|x|,|y|)$, and $|z| \geq \max(0, |x|+|y|-1)$. Conditional factorization and any shared ordering imply $\theta \in \mathcal{O}_2$. These Fréchet moment restrictions (Fontana & Semeraro, 2018) extend the exact-tie example to nonzero means: in particular, $|\tau_{ef}| \leq |\mu_e|$. Distance from $\mathcal{O}_2$ measures how far the observed moments violate necessary shared-order conditions.

The canonical family starts with items 1 and 2 tied above item 3. Set $s_0 = (1,1,-2)/\sqrt{6}$, $q_0 = (1,-1,0)/\sqrt{2}$, and

$$s_\delta = \frac{s_0 + \delta q_0/\sqrt{2}}{\sqrt{1+\delta^2/2}}, \qquad q_\delta = \frac{q_0 - \delta s_0/\sqrt{2}}{\sqrt{1+\delta^2/2}}.$$

The baseline anchor gap is $\delta/\sqrt{1+\delta^2/2}$; ε is the amplitude of the symmetric perturbation across evaluators. The two signs of Z in the following benchmark exchange the leading anchor preference.

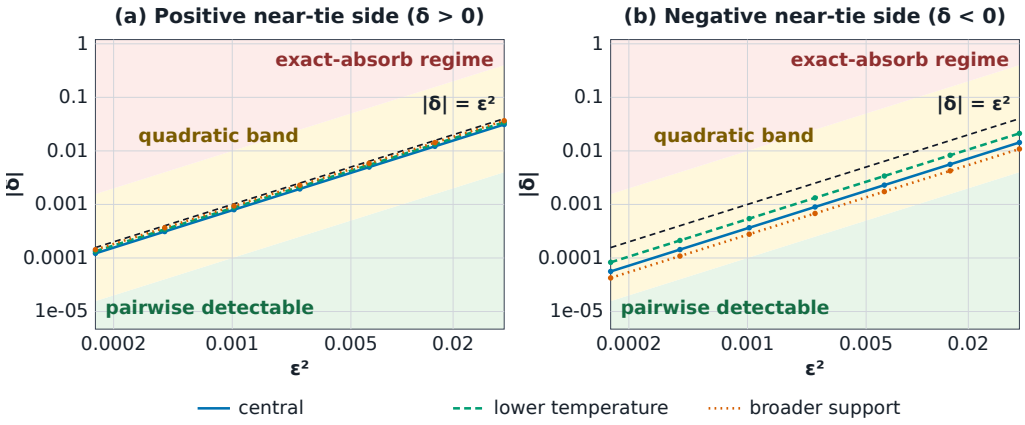


Figure 2: **Canonical collision phase.** Critical pairwise $|\delta|$ is plotted against ε^2 for both near-tie sides and three fixed compact temperature mixtures. The transition follows the predicted quadratic band, while constants vary with sign and nuisance geometry. The full-law absorption background comes from a separate fixed-four-support theorem; plotted onsets give finite-outcome corroboration of the pairwise transition.

Benchmark 4.2 (Canonical three-stimulus boundary)**.** Let $V_{\varepsilon,\delta} = As_\delta + \varepsilon Z q_\delta$, where A has a fixed law G_0 on a compact positive interval, $Z \in \{-1, 1\}$ is conditionally symmetric, and s_δ, q_δ form the canonical normalized near-tie pair specified in Appendix A.4. Under conditional independence, write $\Gamma_{ef} = \mathrm{dist}_\infty(\theta, \mathcal{O}_2)$. If $\delta/\varepsilon^2 \to r$, then

$$\Gamma_{ef} = \varepsilon^2 G_{G_0}(r) + o(\varepsilon^2), \qquad G_{G_0}(0) = \kappa(G_0)/2 > 0. \tag{7}$$

The positive region of G_{G_0} is a finite, side-dependent interval whose constants depend on G_0. Hence the canonical observable transition is $|\delta| \asymp \varepsilon^2$. There also exists a fixed-support hard sequence with exact positive-ray matching when $|\delta| \gg \varepsilon^2$, whereas $\Gamma_{ef} \asymp \varepsilon^2$ when $|\delta| \ll \varepsilon^2$. In the stated compact relative class $|\delta| \le c_0\varepsilon^2$, root-$K$ estimation of the observable gap and a fixed-null lower bound give $\varepsilon_{*,K} \asymp K^{-1/4}$.

The mechanism is visible in the leading moments. At $\delta = 0$, symmetry cancels the anchor mean exactly, while $\tau_{ef} = \kappa(G_0)\varepsilon^2 + O(\varepsilon^4)$ with $\kappa(G_0) > 0$. Moving the baseline away from the tie creates an anchor mean of order δ, which allows a shared-order explanation to absorb a moment of that size. The competition is therefore between $|\delta|$ and ε^2. Inside the positive-gap band, estimation error is of order $K^{-1/2}$ for K fresh evaluators. Equating $\varepsilon^2 \asymp K^{-1/2}$ gives the fourth-root amplitude scale; Appendix A.4 proves the matching upper and lower bounds.

At the base collision law ($\varepsilon = \delta = 0$), a true repeated probe, $B_3 = (e, f, f')$, yields

$$\kappa(G_0) = \tfrac{1}{4}\{1 - \mathbb{E}[D_f D_{f'}]\} \tag{8}$$

under factorization. It calibrates the quadratic coefficient and adds a collision witness, improving fixed-null information per evaluator while the $|\delta| \asymp \varepsilon^2$ band and fourth-root exponent stay unchanged. For perturbed laws, Proposition A.12 subtracts an independently valid calibration tube to obtain a finite-sample lower confidence bound. This is a local use of repeated measurements to calibrate an otherwise unknown coefficient; identification of entire finite mixtures uses stronger latent-structure conditions (Bonhomme et al., 2016).

Implication and scope. The rank criterion covers general finite designs. The quadratic band, $K^{-1/4}$ scale, and repeated-probe identity use the canonical three-stimulus, conditionally symmetric, compact-nuisance benchmark. The gap remains root- K; fourth-root scaling is its $\Gamma \asymp \varepsilon^2$ reparameterization. Exact absorption is least-favourable and its constants are nuisance-dependent. Since $\mathcal{O}_2$ is a necessary outer set, $\Gamma = 0$ is inconclusive about the null.

5 DEPENDENCE BARRIER AND A ROBUST TEST

At the exact tie above, conditional independence forces the anchor–probe moment to vanish under shared order. Dependence can create that same moment without changing either comparison's marginal probability. To account for the entire repeated-probe law, write the conditional means as (x, y, y) and the centred conditional pair and triple moments as $c_{12}, c_{13}, c_{23}, c_{123}$. These interaction errors vanish under factorization. The exact conditional third moment is

$$\mathbb{E}[D_1 D_2 D_3 \mid V] = xy^2 + xc_{23} + y(c_{12} + c_{13}) + c_{123}. \tag{9}$$

This identity shows how pair dependence also affects the third moment. Matching a full law requires fitting all these moments while keeping the eight cell probabilities nonnegative.

Theorem 5.1 (Complete-law dependence counterfeit)**.** *In the exact-collision B_3 benchmark, there is a finite-support interior G_0 such that, for every sufficiently small $\varepsilon > 0$, the complete eight-cell law of the conditionally independent order-changing alternative $Q^{(3)}_{\varepsilon,0}$ is reproduced exactly by a shared-order null on the ray $V = As_0$. The null changes the temperature weights by $O(\varepsilon^2)$ and uses a legal exchangeable conditional kernel that preserves all logistic conditional marginal means. If $b_*(Q_\varepsilon)$ is the smallest pointwise radius over its centred pair and third-order errors, then*

$$b_*(Q_\varepsilon) = \Theta(\varepsilon^2). \tag{10}$$

Exact equality holds for the weak-order, tie-inclusive null. Restricting the null to strict orders gives closure and hence arbitrarily accurate overlap.

The construction perturbs positive support weights by order ε^2, takes $c_{12} = c_{13} = \tau_\varepsilon$ and $c_{23} = 0$, and chooses c_{123} to match the remaining third moment in equation 9. Positivity holds because the base support is interior. Matching the five exchangeable moments identifies the full eight-cell law. The lower bound in equation 10 follows because any counterfeit must cancel a quadratic collision witness. Thus an unknown, unrestricted dependence budget removes a local uniform guarantee along this hard sequence.

Approximate closure with arbitrarily different slopes is known for two-period binary panels with conditionally independent errors and unrestricted covariate-dependent random effects (Forchini & Jiang, 2019, Theorems 2–3). Here complete-law matching also quantifies the smallest departure from conditional independence.

For a supplied pair budget b, define the safe expansion

$$\mathcal{E}_2(b) = \{(x, y, z) : \exists z_0, (x, y, z_0) \in \mathcal{O}_2, |z - z_0| \le b,\ (x, y, z) \text{ is cell-feasible}\}. \tag{11}$$

Theorem 5.2 (Known-budget fixed-risk boundary)**.** *Fix a common finite-support compact nuisance container $\mathcal{G}$, a bundle B_m with $m \in \{2, 3\}$, and a relative canonical class $|\delta| \le c_0\varepsilon^2$, with c_0 strictly inside the uniform positive-gap thresholds. Suppose fresh evaluators are iid, the null permits weak orders, and the externally specified deterministic isotropic budgets obey*

$$b_{12,K} = b_{13,K} = b_{23,K} = b_{123,K} = b_K \to 0.$$

For fixed $0 < \alpha, \beta < 1$ with $\alpha + \beta < 1$, constants $0 < c < C < \infty$ exist such that, for all large K, size α and minimum power $1 - \beta$ are incompatible when

$$\varepsilon \le c \max\{\sqrt{b_K}, K^{-1/4}\},$$

whereas the budgeted distance test achieves them when

$$\varepsilon \ge C \max\{\sqrt{b_K}, K^{-1/4}\}. \tag{12}$$

Thus the fixed-risk boundary in this class is $\varepsilon_{*,K} \asymp \max\{\sqrt{b_K}, K^{-1/4}\}$.

The two terms have different sources. A quadratic signal must exceed sampling error of order $K^{-1/2}$ and the moment error allowed by b_K. The first comparison gives $K^{-1/4}$; the second gives the robustness floor $\sqrt{b_K}$. The exact matching construction and a budget-expanded moment test attain the corresponding lower and upper bounds.

Computing a certificate. Use the tighter necessary set $\mathcal{R}_2(b) \subseteq \mathcal{E}_2(b)$ of Lemma A.14, with $\mathcal{R}_2(0) = \mathcal{O}_2$. For one predeclared pair, compute

$$\widehat{\theta} = \frac{1}{K}\sum_{k=1}^{K}(D_{k,e}, D_{k,f}, D_{k,e}D_{k,f}), \qquad \text{reject if } \operatorname{dist}_\infty(\widehat{\theta}, \mathcal{R}_2(b)) > \sqrt{\frac{2\log(6/\alpha)}{K}}. \tag{13}$$

The coordinates lie in $[-1, 1]$, so Hoeffding's inequality and a union bound give a confidence cube with coverage at least $1 - \alpha$. Null containment in $\mathcal{R}_2(b)$ then gives level α. Multiple pairs or strata require a simultaneous radius, as in the LLM study. This approximate-moment approach (Armstrong & Kolesár, 2021) is safe but can be conservative for the full B_3 geometry.

The *breakdown budget* is the smallest b at which rejection disappears (Masten & Poirier, 2020): the moment dependence this certificate tolerates. Actual session dependence requires separate calibration.

An analysis predeclares a grid of b values and reports the margin and breakdown budget. To use a selected b as a guarantee, a design argument or independent calibration must bound the relevant conditional errors for the same within-bundle kernel and strata. The pairwise test uses the pair error; the isotropic B_3 rate theorem also bounds the other pair and triple errors. Cross-session repeats describe a different kernel.

Implication and scope. A robust certificate must supply its dependence class. The exact counterfeit uses a specified finite-support hard sequence and a tie-inclusive null; the strict-order restriction gives closure overlap. The rate theorem assumes the written common nuisance container, relative calibrated class, known deterministic shrinking *isotropic* budget, and fresh evaluators. Sharp anisotropic B_3 geometry and data-driven budget calibration remain open. With an unspecified budget, the hard sequence rules out a local uniform guarantee.

6 Computational and LLM evidence

We first examine the predicted collision and dependence transitions, then use the moment diagnostic on controlled LLM judgments. A public-data audit identifies the observation fields needed to reuse this design.

Finite-outcome phase and robustness calculations. The deterministic phase audit evaluates seven ε values, both signs of δ, and three compact three-support temperature geometries (42 rows). The onset ratios are nuisance- and sign-dependent: the computed positive-side critical ratios range from 0.779 to 0.915, and the negative-side ratios from 0.271 to 0.531. Figure 2 shows their bounded variation around the $|\delta|/\varepsilon^2$ organization in Benchmark 4.2.

The dependence grid uses $K \in \{4096, 16384, 65536\}$, 105 grid points per scenario, and 800 Monte Carlo replicates per cell. All conditional laws are legal (minimum cell probability 0.0182). A naive test rejects in $800/800$ draws at a dependence-boundary null. The budget-aware test rejects in at most $10/800 = 0.0125$ draws there (95% Wilson interval $[0.0068, 0.0229]$). Across observable-overlap cells, mean worst-direction robust power is 0.0017; when $\sqrt{K}(\text{signal} - 2b)_+ \geq 4$, mean worst-direction power is 0.984. Wilson intervals accompany the finite-sample audit; the theorem supplies size control. Figure 3 shows the corresponding power phase against the theorem-predicted order $\varepsilon^2 \asymp b + K^{-1/2}$.

Earlier latent-mixture fitting experiments failed their calibration gates and remain scientifically inconclusive. Appendix C records their status and a separate deterministic exact-overlap certificate.

A controlled LLM-judge example. An independent pilot selected two literary triples; no pilot judgment entered confirmation. The narrative-rhythm triple retells the same night-bus story in fragmented or continuous prose. Fresh confirmation used a predeclared, equal-weight roster of six literary personas within each of three requested models, including personas favouring each of these rhythms. Here $Z = (\text{requested model}, \text{persona})$ indexes a known response-kernel type. One recorded bundle contains three stateless requests sharing Z with independent message histories, so K counts bundles. The common model–persona type supplies the linkage; each response is obtained in a fresh context. Appendix D specifies the protocol, sample split, and record schema.

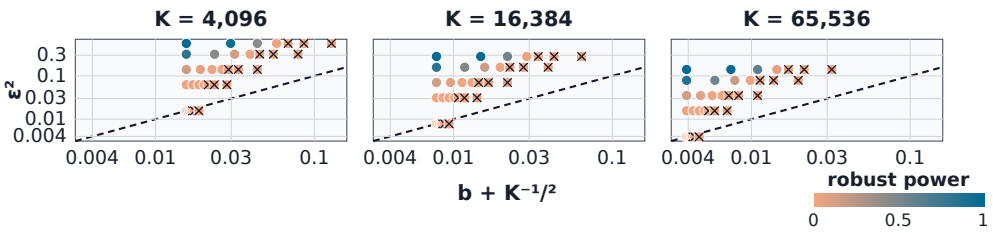


Figure 3: **Dependence-robust finite-sample phase.** Worst-direction budget-aware power for three evaluator counts (800 trials per cell); crosses mark population overlap. The dashed diagonal is the theorem-predicted order $\varepsilon^2 \asymp b + K^{-1/2}$, with constants unspecified. Wilson intervals and boundary-null size are in the released table. The plot corroborates the predicted scaling at finite sample sizes.

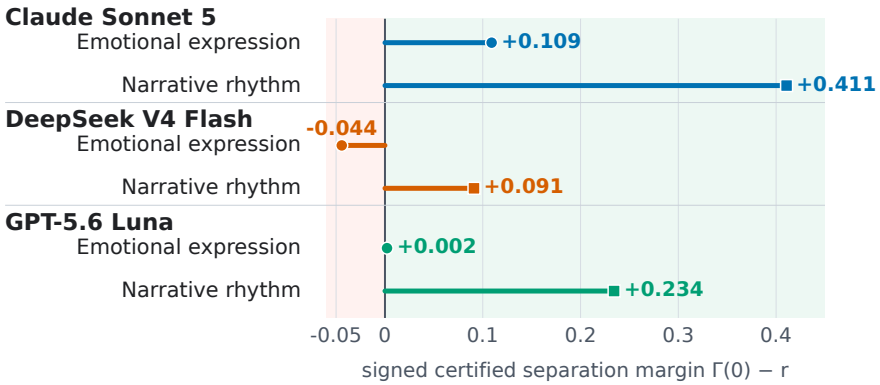


Figure 4: **Fixed-stratum separation margins.** Signed margins $\Gamma(0) - r$ for six fixed requested-model–triple strata; positive values certify separation from the shared-order necessary set at simultaneous 95% confidence. Every row has $K = 4,104$ fresh type-linked bundles, and the common radius jointly protects 18 randomized-presentation moment coordinates. The tested response law averages over randomized presentation; position effects are included in this estimand.

The confirmation retained 24,624 valid bundles and 73,872 judgments. Each of the six fixed model–triple strata has $K = 4,104$. Presentation orientation was balanced at random. Repeated probe slots were symmetrized, and the primary estimand was the three-coordinate moment vector $\theta = (\mu_e, \mu_f, \tau_{ef})$. A common Hoeffding radius $r = 0.0566$ protects all 18 primary coordinates simultaneously at familywise 95% confidence. We numerically certify $\Gamma(0) = \mathrm{dist}_\infty(\theta, \mathcal{O}_2)$ and report the signed separation margin $\Gamma(0) - r$.

All three narrative-rhythm strata separate the necessary set; Claude also separates on emotional expression, while Luna clears the simultaneous boundary by 0.0021 and DeepSeek does not separate at this precision. Thus five of the six tested model–triple combinations reject. Under an externally imposed sensitivity budget $b = 0.20$, four fixed strata remain separated, including all three narrative-rhythm strata. Appendix D reports both budgets; the supplement contains the full sensitivity grid.

Balanced randomization is part of the stated estimand: across the six fixed strata, the mean absolute persona–slot presentation contrast ranges from 0.103 to 0.909 on the signed-response scale. The result concerns the randomized-presentation response kernels of these rosters and stimuli; randomization does not establish presentation-invariant literary preferences.

Public releases and the linked observation unit. Ten purposively selected releases were checked for evaluator linkage, fixed objects, native binary edges, multiple linked edges, a true repeat, and raw presentation metadata. None documents all six fields. The core matrix and evidence in Appendix E locate the missing information for this sample of releases. Reliable evaluator identifiers alone do not reconstruct the joint law required by the diagnostic.

7 Related work

Classical paired-comparison models infer rankings from binary judgments (Thurstone, 1927; Bradley & Terry, 1952). Makur & Singh (2025) test whether independent edge counts fit one Bradley–Terry–Luce matrix, with a $\Theta((nk)^{-1/2})$ complete-graph threshold. Our observation is a within-evaluator joint law under a population mixture of response scales; shared-order certification therefore concerns a different null and separation.

Random-utility models admit nonparametric specification tests (Kitamura & Stoye, 2018); sufficient covariate support can identify random coefficients under additional assumptions (Fox et al., 2012). Under their linear-reward heterogeneity model, Chidambaram et al. (2026) identify latent preferences from one preferred response in a three-option menu when one binary comparison per user is insufficient. Their categorical choice differs from our three linked binary slots $B_3 = (e, f, f')$. Feedback formats must be compared together with their response models and support assumptions.

The saturation argument builds on Haberman (2005, Theorems 1–2): locally unrestricted logistic item models already use finite-support simplices and exponential Vandermonde matrices. Section 3 extends that argument to the structural constraints of a linked library. Related mixture results address affine menu approximation (Chang et al., 2026) and identification from grouped measurements (Allman et al., 2009; Bonhomme et al., 2016; Vandermeulen & Scott, 2019). The latter recovers a finite mixture from conditionally iid observations sharing a component; linked slots here use different edge kernels and allow an unrestricted scale distribution. Our collision criterion tracks the observable restrictions created by this design and connects them to the canonical phase and the dependence budget.

LLM-as-a-judge studies document scalable evaluation and substantial response biases (Zheng et al., 2023; Chiang et al., 2024). These motivate the presentation controls in our response-kernel diagnostic and the release-schema audit.

8 Implications, limitations, and conclusion

What follows for preference learning. For the two populations matched by Corollary 3.2, let $\mathcal{D}_K$ denote the complete data from the fixed protocol. Any learner $\mathsf{A}(\mathcal{D}_K, U)$, with the same independent randomization U in both populations, has identically distributed outputs. In particular, with the same prompts, candidate responses, reference policy, and training rule, reward-model and DPO losses computed from these labels have the same distribution (Rafailov et al., 2023). A good fit, stable latent clusters, or additional evaluators on that protocol cannot resolve the matched ranking structures. This is a limit on interpreting the learned structure; policy utility or regret would require a specified downstream objective.

The constructive consequence is to design data collection around a testable restriction. Preserve the labelled joint answers, use independent pilot and confirmation units to select promising comparisons, and report the dependence budgets under which a shared-order explanation is rejected. A repeated probe can calibrate the canonical signal coefficient under factorization. A bound on residual dependence needs its own design or calibration argument.

Scope of the evidence. The LLM study tests randomized-presentation response kernels for six-persona rosters and two literary triples; its strong position effects remain part of that estimand. The public-release audit concerns a purposive sample checked in August 2026. Extending certification to human populations requires an appropriate sampling frame and dependence calibration; data-driven budget calibration remains open.

Linked feedback can reveal ranking disagreement when its joint patterns escape the shared-order restrictions. The results identify two obstacles to that interpretation: flexible decisiveness can match complete response laws, and dependence can reproduce the remaining collision signal at the same quadratic scale. Designing informative comparisons and controlling dependence must therefore accompany the collection of linked labels.

AI USE STATEMENT

We used generative AI tools for the following required-disclosure tasks: developing conceptual and theoretical frameworks; formulating mathematical claims; contributing proof ideas and draft derivations; proposing and refining hypotheses; designing and reviewing experimental protocols; implementing methods; assisting with translation; and interpreting results. We did not use generative AI tools to generate synthetic data sets, clean and reformat datasets, or support qualitative and thematic data analysis.

Additionally, we used generative AI tools for recommended-disclosure tasks, including literature search and bibliographic checks, code and test authoring, figure rendering, LaTeX editing, drafting literary stimuli that the authors later approved, and drafting, restructuring, and copy-editing the manuscript. Language-model outputs also constitute the recorded observations in the LLM-evaluator study; those observations are study data, not research assistance.

The authors set the research questions and claim boundaries, approved all formal stimuli and designs before collection, checked the theorems against the appendix proofs, reran the deterministic validators, and verified citations against primary sources. We have reviewed all AI-assisted work and take responsibility for the final content, including text, claims, and artifacts. No language model is an author.

ETHICS STATEMENT

The empirical study evaluates LLM response kernels directly and involves no human participants. It does not treat LLM judgments as a proxy for human preferences. The public-release audit used dataset documentation and aggregate schema metadata; no individual-level records were downloaded or analysed. Requests were stateless and used fixed literary comparison prompts; the analysis retained only the fields required for the study and its released presentation summary contains no raw response text. Model labels denote requested API strata, not persistent identities.

REPRODUCIBILITY STATEMENT

The supplementary project contains deterministic source-hashed experiment packages, fixed seeds and configurations where randomness is used, machine-readable CSV/JSON outputs, Wilson intervals for Monte Carlo results, and validators for cell positivity, mass conservation, complete-law matching, and output invariants. Appendix B maps each main-text number and figure to its artifact; Appendix D specifies the LLM pilot/confirmation split and estimand; Appendix E records the public-release coding rules and sources; Appendix A maps every claim to its proof.

REFERENCES

Elizabeth S. Allman, Catherine Matias, and John A. Rhodes. Identifiability of parameters in latent structure models with many observed variables. *The Annals of Statistics*, 37(6A):3099–3132, 2009. doi: 10.1214/09-AOS689.

Timothy B. Armstrong and Michal Kolesár. Sensitivity analysis using approximate moment condition models. *Quantitative Economics*, 12(1):77–108, 2021. doi: 10.3982/QE1609.

Stéphane Bonhomme, Koen Jochmans, and Jean-Marc Robin. Non-parametric estimation of finite mixtures from repeated measurements. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(1):211–229, 2016. doi: 10.1111/rssb.12110.

Ralph Allan Bradley and Milton E. Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952. doi: 10.1093/biomet/39.3-4.324.

Haoge Chang, Yusuke Narita, and Kota Saito. Approximating choice data by discrete choice models. *arXiv preprint arXiv:2205.01882v5*, 2026. URL https://arxiv.org/abs/2205.01882.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Banghua Zhu, Hao Zhang, Michael Jordan, Joseph E. Gonzalez, and Ion Stoica.

Chatbot arena: An open platform for evaluating LLMs by human preference. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 8359–8388. PMLR, 2024. URL https://proceedings.mlr.press/v235/chiang24b.html.

Keertana Chidambaram, Karthik Vinay Seetharaman, and Vasilis Syrgkanis. Direct preference optimization with unobserved preference heterogeneity: The necessity of ternary preferences. In *Proceedings of The 29th International Conference on Artificial Intelligence and Statistics*, volume 300 of *Proceedings of Machine Learning Research*, pp. 4555–4563. PMLR, 2026. URL https://proceedings.mlr.press/v300/chidambaram26a.html.

Roberto Fontana and Patrizia Semeraro. Representation of multivariate bernoulli distributions with a given set of specified moments. *Journal of Multivariate Analysis*, 168:290–303, 2018. doi: 10.1016/j.jmva.2018.08.003.

Giovanni Forchini and Bin Jiang. Fragility of identification in panel binary response models. *The Econometrics Journal*, 22(3):282–291, 2019. doi: 10.1093/ectj/utz011.

Jeremy T. Fox, Kyoo il Kim, Stephen P. Ryan, and Patrick Bajari. The random coefficients logit model is identified. *Journal of Econometrics*, 166(2):204–212, 2012. doi: 10.1016/j.jeconom.2011.09.002.

Jiaying Gu. Neyman's $C(\alpha)$ test for unobserved heterogeneity. *Econometric Theory*, 32(6):1483–1522, 2016. doi: 10.1017/S0266466615000249.

Shelby J. Haberman. Identifiability of parameters in item response models with unconstrained ability distributions. Research Report RR-05-24, Educational Testing Service, 2005. URL https://www.ets.org/research/policy_research_reports/publications/report/2005/hzdl.html.

Emilie Kaufmann, Olivier Cappé, and Aurélien Garivier. On the complexity of best-arm identification in multi-armed bandit models. *Journal of Machine Learning Research*, 17(1):1–42, 2016. URL https://jmlr.org/papers/v17/kaufman16a.html.

Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-pic: An open dataset of user preferences for text-to-image generation. In *Advances in Neural Information Processing Systems*, volume 36, 2023. doi: 10.52202/075280-1594. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/73aacd8b3b05b4b503d58310b523553c-Abstract-Conference.html.

Yuichi Kitamura and Jörg Stoye. Nonparametric analysis of random utility models. *Econometrica*, 86(6):1883–1909, 2018. doi: 10.3982/ECTA14478.

Shu Kong, Xiaohui Shen, Zhe Lin, Radomir Mech, and Charless Fowlkes. Photo aesthetics ranking network with attributes and content adaptation. In *Computer Vision – ECCV 2016*, volume 9905 of *Lecture Notes in Computer Science*, pp. 662–679. Springer, 2016. doi: 10.1007/978-3-319-46448-0_40. URL https://link.springer.com/chapter/10.1007/978-3-319-46448-0_40.

Bruce G. Lindsay. The geometry of mixture likelihoods, part II: The exponential family. *The Annals of Statistics*, 11(3):783–792, 1983. doi: 10.1214/aos/1176346245.

Anuran Makur and Japneet Singh. Minimax hypothesis testing for the Bradley–Terry–Luce model. *IEEE Transactions on Information Theory*, 71(12):9163–9202, 2025. doi: 10.1109/TIT.2025.3605846. URL https://arxiv.org/abs/2410.08360.

Matthew A. Masten and Alexandre Poirier. Inference on breakdown frontiers. *Quantitative Economics*, 11(1):41–111, 2020. doi: 10.3982/QE1288.

Daniel McFadden. Conditional logit analysis of qualitative choice behavior. In Paul Zarembka (ed.), *Frontiers in Econometrics*, pp. 105–142. Academic Press, New York, 1974.

Lester James Validad Miranda, Yizhong Wang, Yanai Elazar, Sachin Kumar, Valentina Pyatkin, Faeze Brahman, Noah A. Smith, Hannaneh Hajishirzi, and Pradeep Dasigi. Hybrid preferences: Learning to route instances for human vs. AI feedback. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7162–7200, 2025. doi: 10.18653/v1/2025.acl-long.355. URL https://aclanthology.org/2025.acl-long.355/.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems*, volume 36, pp. 53728–53741, 2023. doi: 10.52202/075280-2338. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/a85b405ed65c6477a4fe8302b5e06ce7-Abstract-Conference.html.

Jian Ren, Xiaohui Shen, Zhe Lin, Radomir Mech, and David J. Foran. Personalized image aesthetics. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 638–647, 2017. URL https://openaccess.thecvf.com/content_iccv_2017/html/Ren_Personalized_Image_Aesthetics_ICCV_2017_paper.html.

Andrea Rotnitzky, David R. Cox, Matteo Bottai, and James Robins. Likelihood-based inference with singular information matrix. *Bernoulli*, 6(2):243–284, 2000. doi: 10.2307/3318576.

Louis L. Thurstone. A law of comparative judgment. *Psychological Review*, 34(4):273–286, 1927. doi: 10.1037/h0070288.

Robert A. Vandermeulen and Clayton D. Scott. An operator theoretic approach to nonparametric mixture models. *The Annals of Statistics*, 47(5):2704–2733, 2019. doi: 10.1214/18-AOS1762. URL https://arxiv.org/abs/1607.00071.

Zhilin Wang, Alexander Bukharin, Olivier Delalleau, Daniel Egert, Gerald Shen, Jiaqi Zeng, Oleksii Kuchaiev, and Yi Dong. Helpsteer2-preference: Complementing ratings with preferences. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=MnfHxPP5gs.

Yuzhe Yang, Liwu Xu, Leida Li, Nan Qie, Yaqian Li, Peng Zhang, and Yandong Guo. Personalized image aesthetics assessment with rich attributes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19861–19869, 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/Yang_Personalized_Image_Aesthetics_Assessment_With_Rich_Attributes_CVPR_2022_paper.html.

Lily Hong Zhang, Smitha Milli, Karen Jusko, Jonathan Smith, Brandon Amos, Wassim Bouaziz, Manon Revel, Jack Kussman, Yasha Sheynin, Lisa Titus, Bhaktipriya Radharapu, Jane Yu, Vidya Sarma, Kris Rose, and Maximilian Nickel. Cultivating pluralism in algorithmic monoculture: The community alignment dataset. In *The Fourteenth International Conference on Learning Representations*, 2026a. URL https://openreview.net/forum?id=4NtoAVqfHA.

Xueyao Zhang, Chaoren Wang, Huan Liao, Ziniu Li, Yuancheng Wang, Li Wang, Dongya Jia, Yuanzhe Chen, Xiulin Li, Zhuo Chen, and Zhizheng Wu. Speechjudge: Towards human-level judgment for speech naturalness. In *The Fourteenth International Conference on Learning Representations*, 2026b. URL https://openreview.net/forum?id=I9ED9VWZq6.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-bench and chatbot arena. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*, volume 36, pp. 46595–46623, 2023. doi: 10.52202/075280-2020. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/91f18a1287b398d378ef22505bf41832-Abstract-Datasets_and_Benchmarks.html.

Haonan Zhu, Elad Hirsch, Alexandria Minetti, Allison Nulty, and Purvanshi Mehta. TASTE: A designer-annotated multi-dimensional preference dataset for AI-generated graphic design. *arXiv preprint arXiv:2605.20731*, 2026. URL https://arxiv.org/abs/2605.20731.

A PROOFS

The proofs proceed from the common likelihood coordinates to saturation, collision geometry, and dependence robustness. All neighbourhoods are relative to the indicated affine space. All norms on a fixed finite-dimensional outcome space are equivalent; we name a norm only when its constant matters.

Table 1: Proof dependencies.

Result	Main proof step
Exact camouflage	Match nearby laws by interior barycentric coordinates on one stacked ray simplex.
Reversal and allocation	Cross an ordering cone with fixed mass; propagate equality through matched transition kernels.
Collision rank and rate	Group exponent rows, identify the witness quotient, and project finite-cell KL.
Canonical boundary	Expand the symmetric logistic path and control the shrinking ray simplex.
Dependence counterfeit	Match all Walsh–Fourier moments with a positive conditional law.
Known-budget boundary	Combine robust outer-set separation with exact adversarial cancellation.

A.1 FINITE-LIBRARY LIKELIHOOD COORDINATES

Write $\mathbb{U} = \{v \in \mathbb{R}^N : \mathbf{1}^\top v = 0\}$. Adding a common constant to every utility changes no response probability, so restricting to $\mathbb{U}$ loses nothing. For an oriented edge $e = (i, j)$, let $x_e = \mathbf{e}_i - \mathbf{e}_j$, and put $Y_e = (D_e + 1)/2$. With $\ell(t) = (1 + \exp(-t))^{-1}$, conditional independence gives, for a bundle $B^{(j)} = (e_{j1}, \dots, e_{jm_j})$ and cell $y \in \mathcal{Y}_j := \{0, 1\}^{m_j}$,

$$\psi_j(v)_y = \prod_{t=1}^{m_j} \ell(x_{e_{jt}}^\top v)^{y_t} \{1 - \ell(x_{e_{jt}}^\top v)\}^{1-y_t} = \frac{\exp(v^\top c_{jy})}{Z_j(v)}, \tag{14}$$

where

$$c_{jy} = \sum_{t=1}^{m_j} y_t x_{e_{jt}}, \qquad Z_j(v) = \sum_{z \in \mathcal{Y}_j} \exp(v^\top c_{jz}).$$

This formula permits repeated edges and arbitrary incidence or cycle identities. Let

$$\Psi_{\mathfrak{B}}(v) = (\psi_1(v), \dots, \psi_J(v)), \qquad \Phi_{\mathfrak{B},s}(a) = \Psi_{\mathfrak{B}}(as), \quad a > 0,$$

and define

$$\mathcal{A}_{\text{common}} = \operatorname{aff}\{\Psi_{\mathfrak{B}}(v) : v \in \mathbb{U}\}, \qquad \mathcal{A}_{\text{ray}}(s) = \operatorname{aff}\{\Phi_{\mathfrak{B},s}(a) : a > 0\}. \tag{15}$$

The first affine space is deliberately not the Cartesian product of the individual bundle simplices: the same latent vector generates every block.

To capture all cross-block affine identities, form the finite exponent set

$$\mathcal{D} = \left\{ \sum_{j=1}^{J} c_{jz_j} : (z_1, \dots, z_J) \in \prod_{j=1}^{J} \mathcal{Y}_j \right\},$$

where equal vectors are stored once. Their multiplicities are

$$N_d = \#\left\{(z_1,\dots,z_J) : \sum_j c_{jz_j} = d\right\},$$
$$N_{jy,d} = \#\left\{(z_k)_{k\neq j} : c_{jy} + \sum_{k\neq j} c_{kz_k} = d\right\}.$$

The incidence–convolution matrix $R_{\mathfrak{B}}$ has rows indexed by $d \in \mathcal{D}$, a constant column $\star$, and columns (j,y):

$$R_{d,\star} = N_d, \qquad R_{d,(j,y)} = N_{jy,d}. \tag{16}$$

For $\lambda \in \Lambda_s := \{s^\top d : d \in \mathcal{D}\}$, define the row-grouping matrix

$$(P_s)_{\lambda,d} = \mathbf{1}\{s^\top d = \lambda\}. \tag{17}$$

Lemma A.1 (Exact affine annihilators). *For an affine functional*

$$L_{\alpha,u}(p) = \alpha + \sum_{j=1}^{J}\sum_{y\in\mathcal{Y}_j} u_{jy}p_{jy}, \qquad \vartheta = (\alpha, (u_{jy})_{j,y}),$$

the following equivalences hold:

$$L_{\alpha,u}(\Psi_{\mathfrak{B}}(v)) = 0 \quad \forall v \in \mathbb{U} \iff R_{\mathfrak{B}}\vartheta = 0, \tag{18}$$
$$L_{\alpha,u}(\Phi_{\mathfrak{B},s}(a)) = 0 \quad \forall a > 0 \iff P_s R_{\mathfrak{B}}\vartheta = 0. \tag{19}$$

Consequently,

$$\dim \mathcal{A}_{\text{common}} = \operatorname{rank}(R_{\mathfrak{B}}) - 1, \tag{20}$$
$$\dim \mathcal{A}_{\text{ray}}(s) = \operatorname{rank}(P_s R_{\mathfrak{B}}) - 1. \tag{21}$$

Proof. Multiply $L_{\alpha,u}(\Psi_{\mathfrak{B}}(v))$ by the strictly positive product $\prod_{k=1}^{J} Z_k(v)$. Expanding with equation 14 yields

$$\{L_{\alpha,u}(\Psi_{\mathfrak{B}}(v))\}\prod_{k=1}^{J} Z_k(v) = \sum_{d\in\mathcal{D}} \beta_d \exp(v^\top d), \qquad \beta_d = \alpha N_d + \sum_{j,y} u_{jy}N_{jy,d}. \tag{22}$$

Thus $\beta = R_{\mathfrak{B}}\vartheta$. The functions $v \mapsto \exp(v^\top d)$, for distinct $d \in \mathcal{D} \subset \mathbb{U}$, are linearly independent. Indeed, choose $w \in \mathbb{U}$ outside the finitely many hyperplanes $w^\top(d - d') = 0$; restricting to $v = tw$ reduces the claim to independence of univariate exponentials with distinct exponents. Hence equation 22 vanishes for all v exactly when every β_d vanishes, proving equation 18.

Along $v = as$, collect terms with the same scalar exponent:

$$\sum_{d\in\mathcal{D}} \beta_d e^{as^\top d} = \sum_{\lambda\in\Lambda_s}\left(\sum_{d:s^\top d=\lambda} \beta_d\right) e^{a\lambda}.$$

Univariate exponential independence on the open interval $(0,\infty)$ proves equation 19. Finally, the linear annihilators of the augmented sets $\{(1,\Psi(v))\}$ and $\{(1,\Phi_s(a))\}$ are the two kernels above. Rank–nullity gives their augmented linear dimensions, and subtracting the constant coordinate gives equations 20–21. □

Lemma A.2 (Generic directions). *Let*

$$\mathcal{E}_{\mathfrak{B}} = \bigcup_{\substack{d,d'\in\mathcal{D}\\ d\neq d'}} \{s \in \mathbb{U} : s^\top(d - d') = 0\} \cup \bigcup_{1\le i<j\le N} \{s \in \mathbb{U} : s_i = s_j\}. \tag{23}$$

Its complement is open, dense, and full Lebesgue measure in $\mathbb{U}$. *Every* $s \notin \mathcal{E}_{\mathfrak{B}}$ *has a strict item order and*

$$\mathcal{A}_{\text{ray}}(s) = \mathcal{A}_{\text{common}}. \tag{24}$$

Proof. The exceptional set is a finite union of proper linear hyperplanes. Its complement therefore has the stated topological and measure properties. There P_s merely relabels distinct exponent rows, so $\operatorname{rank}(P_s R) = \operatorname{rank}(R)$. Lemma A.1 then gives equal affine dimensions; the inclusion $\mathcal{A}_{\mathrm{ray}}(s) \subseteq \mathcal{A}_{\mathrm{common}}$ gives equality. The added item-tie hyperplanes make the ordering strict. □

The sufficient condition in Lemma A.2 is stronger than necessary. At a colliding direction the exact criterion remains $\operatorname{rank}(P_s R) = \operatorname{rank}(R)$; counting colliding pairs of exponent rows is not a rank calculation.

A.2 Proof of Temperature Camouflage

Lemma A.3 (A common interior ray simplex)**.** *Suppose $\mathcal{A}_{\mathrm{ray}}(s) = \mathcal{A}_{\mathrm{common}}$, and put $d = \dim \mathcal{A}_{\mathrm{common}}$. There are positive temperatures $a_0, \dots, a_d$ for which the stacked vectors $\Phi_{\mathfrak{B},s}(a_r)$ are affinely independent. For any weights $w_r > 0$ summing to one, the point*

$$p_0 = \sum_{r=0}^{d} w_r \Phi_{\mathfrak{B},s}(a_r) \tag{25}$$

has a relative neighbourhood $\mathcal{U} \subset \mathcal{A}_{\mathrm{common}}$ such that every $Q \in \mathcal{U}$ has a unique representation

$$Q = \sum_{r=0}^{d} w_r(Q) \Phi_{\mathfrak{B},s}(a_r), \qquad w_r(Q) > 0, \qquad \sum_r w_r(Q) = 1. \tag{26}$$

Proof. Equality of the affine hulls and the definition of affine dimension allow us to select $d+1$ affinely independent points from the positive ray. Their convex hull is a full-dimensional simplex in $\mathcal{A}_{\mathrm{common}}$. The positive weights in equation 25 put p_0 in its relative interior. Barycentric coordinates relative to an affine basis are affine, hence continuous, on the whole affine hull. Since every coordinate is positive at p_0, all remain positive in a sufficiently small relative neighbourhood. Affine independence gives uniqueness. □

Proof of Theorem 3.1. Fix the finite library. By Lemma A.2, choose a unit $s \notin \mathcal{E}_{\mathfrak{B}}$; it has a strict order and its positive ray has affine hull $\mathcal{A}_{\mathrm{common}}$. Apply Lemma A.3, choose any positive base weights, and set

$$G_0 = \sum_{r=0}^{d} w_r \delta_{a_r}.$$

Let a population law Q have stacked bundle law $q \in \mathcal{U}$. Because every evaluator population produces

$$q = \int \Psi_{\mathfrak{B}}(v)\, dQ(v),$$

and finite-dimensional affine subspaces are closed, q belongs to $\mathcal{A}_{\mathrm{common}}$. Define

$$G_Q = \sum_{r=0}^{d} w_r(q) \delta_{a_r}. \tag{27}$$

Equation 26, applied to the entire stack rather than block by block, gives

$$\int \Phi_{\mathfrak{B},s}(a)\, dG_Q(a) = q.$$

Every coordinate of this identity is a labelled outcome-cell probability. Thus it is exactly equation 3, and the single weight vector in equation 27 serves all bundles. The complement of the finite hyperplane set used to select s is open, dense, and full measure, which also proves the genericity assertion. □

We next construct the fixed-positive reversal in a centred rank-one residual family. This also makes explicit why its admissible mass depends on the chosen library and interior point.

Lemma A.4 (A centred transverse direction crosses both ways). *Let $N \geq 3$, let $s \in \mathbb{U}$ have distinct coordinates, and let $q \neq 0$ belong to $\mathbb{U} \cap s^{\perp}$. After relabelling so that $s_1 < \cdots < s_N$, there are pairs $i < j$ and $k < \ell$ such that*

$$q_i > q_j, \qquad q_k < q_\ell.$$

Consequently, for every finite set of positive temperatures there is $L < \infty$ such that both $as + Lq$ and $as - Lq$ lie outside the ordering cone of s, for every temperature in that set.

Proof. For centred s, q,

$$Ns^{\top}q = \sum_{i<j}(s_j - s_i)(q_j - q_i). \tag{28}$$

If nonzero q were nondecreasing, the right side would be strictly positive; if it were nonincreasing, it would be strictly negative. Both contradict $q \perp s$. Hence the two displayed inequalities exist.

Let $A_{\max}$ be the largest selected temperature. Choose

$$L > A_{\max} \max\left\{\frac{s_j - s_i}{q_i - q_j}, \frac{s_\ell - s_k}{q_\ell - q_k}\right\}.$$

Then $(as + Lq)_i > (as + Lq)_j$ and $(as - Lq)_k > (as - Lq)_\ell$, each contrary to the order of s. Avoiding the finitely many values of L that create an item tie makes every displaced support point strict, if desired. □

Proof of Corollary 3.2. Use the direction, supports, weights, and interior point from the theorem. Because $N \geq 3$, $\mathbb{U} \cap s^{\perp}$ contains a nonzero unit vector q. Choose L by Lemma A.4. With $A \sim G_0$, define a conditionally centred residual by

$$\Pr(Z = 0 \mid A) = 1 - \rho, \qquad \Pr(Z = L \mid A) = \Pr(Z = -L \mid A) = \rho/2, \tag{29}$$

and set $V = As + Zq$. Thus $E[Z \mid A] = 0$, and the population law is a centred rank-one residual family.

Let $\mathcal{C}_s$ be the unique strict ordering cone containing every as. Exactly the mass ρ in equation 29 lies outside $\mathcal{C}_s$. Any other ordering cone excludes all of the strict base-ray mass, and therefore excludes mass at least $1 - \rho$. Hence, for $0 < \rho < 1/2$,

$$\rho_{\mathrm{ord}}(F_\rho) = \min\{\rho, 1 - \rho\} = \rho. \tag{30}$$

The stacked law of this population is

$$Q_\rho = (1 - \rho)p_0 + \rho R_L, \qquad R_L = \frac{1}{2}\sum_r w_r \left\{\Psi_{\mathfrak{B}}(a_r s + Lq) + \Psi_{\mathfrak{B}}(a_r s - Lq)\right\}. \tag{31}$$

Both p_0 and R_L belong to $\mathcal{A}_{\mathrm{common}}$, and $\|Q_\rho - p_0\| = \rho\|R_L - p_0\|$. Choose a relative ball of radius $r_0 > 0$ inside the neighbourhood in Lemma A.3. If $R_L \neq p_0$, take

$$\rho_* < \min\left\{\frac{1}{2}, \frac{r_0}{\|R_L - p_0\|}\right\}; \tag{32}$$

if $R_L = p_0$, any $\rho_* < 1/2$ works. For every fixed $0 < \rho < \rho_*$, the common barycentric construction supplies one positive-temperature law G_ρ whose complete stacked law equals Q_ρ.

The corresponding positive-ray population lies in H_0^{ord}. An observable test has the same rejection probability under two identical data laws, so a test with null size at most α has power at most α at this F_ρ. Equations 32 and 31 also show explicitly that ρ_* is a least-favourable, library-dependent constant, not a uniform endpoint. □

Lemma A.5 (Predictable allocation preserves equality). *Suppose two experiments have the same complete outcome law for every bundle b in a finite library, using one common nuisance law in the null experiment. If at round t a policy selects B_t by a kernel depending only on the past transcript and then observes a fresh evaluator, the two complete transcript laws agree at every finite horizon. They also agree after any transcript-measurable bounded stopping time.*

Proof. Given the same past history h, both experiments apply the same selection kernel $\Pi_t(db \mid h)$. Conditional on the selected b, freshness makes the current outcome independent of the past given the model, and the assumed complete bundle laws are equal. Hence the transition kernels from h to (B_t, D_{B_t}) coincide. Induction from the empty history proves finite-horizon equality. A stopped transcript is a measurable function of the full finite transcript, so equality is preserved under pushforward. □

Finite-depth within-evaluator trees. Fix a finite edge set and depth h. A deterministic binary query tree has at most $2^h - 1$ internal nodes. Give every potential node its own labelled response occasion, order these occasions canonically, and call their list B_T^{super}. Under counterfactual occasion independence given V, pre-generating the responses at every potential node produces the product law equation 14; following the realized branches is a deterministic map from this super-bundle vector to the observed tree transcript. There are only finitely many labelled trees at fixed depth, so their super-bundles form a finite static library. Applying Theorem 3.1 and Corollary 3.2 to that enlarged library matches every super-bundle with one common G, and therefore matches every tree transcript under the deterministic pushforward. External randomization is handled by conditioning on a seed independent of the evaluator. Cross-evaluator adaptivity then follows from Lemma A.5.

This reduction uses a fixed finite depth, a finite action set, fresh latent draws, and conditional independence across potential response occasions. A realized branch that changes prompt history, memory, or action persistence changes the current kernel. Such paths must be encoded as distinct interventions in a new finite library and the global-potential and occasion- independence assumptions rechecked there. Depth-uniform bounds, fixed panels, and evaluator reuse therefore remain separate problems.

A.3 Collision quotient and fixed-ray local rate

Fix s and a point $p_0 \in \mathcal{A}_{\mathrm{ray}}(s)$. Denote the direction spaces by

$$T_{\mathrm{common}} = \mathrm{lin}(\mathcal{A}_{\mathrm{common}} - p_0), \qquad T_{\mathrm{ray}} = \mathrm{lin}(\mathcal{A}_{\mathrm{ray}}(s) - p_0).$$

For an affine functional $L(p) = \alpha + u^\top p$, write $\ell_L(h) = u^\top h = L(p_0 + h) - L(p_0)$ for its linear part.

Lemma A.6 (Witness quotient duality)**.** *There is a natural nondegenerate pairing*

$$\frac{\ker(P_s R_{\mathfrak{B}})}{\ker(R_{\mathfrak{B}})} \times \frac{T_{\mathrm{common}}}{T_{\mathrm{ray}}} \longrightarrow \mathbb{R}, \qquad ([L], [h]) \longmapsto \ell_L(h). \tag{33}$$

In particular, every nonzero normal class $[h]$ is detected by an affine functional that vanishes on the entire fixed- s ray affine hull but is not a universal common-law identity.

Proof. Lemma A.1 identifies $\ker(P_s R)$ with affine functionals vanishing on $\mathcal{A}_{\mathrm{ray}}(s)$, and $\ker(R)$ with those vanishing on $\mathcal{A}_{\mathrm{common}}$. Their linear parts vanish on T_{ray} and T_{common}, respectively. Thus adding an element of T_{ray} to h, or a universal common identity to L, does not change the pairing. After restricting the linear parts to T_{common}, ordinary finite-dimensional quotient duality identifies the left quotient with $(T_{\mathrm{common}}/T_{\mathrm{ray}})^*$; hence the pairing is nondegenerate. □

Algebraic and geometric parts of Theorem 4.1. Equations 20–21 give

$$\begin{aligned} \dim \mathcal{A}_{\mathrm{common}} - \dim \mathcal{A}_{\mathrm{ray}}(s) &= \mathrm{rank}(R_{\mathfrak{B}}) - \mathrm{rank}(P_s R_{\mathfrak{B}}) \\ &= \Delta(\mathfrak{B}, s), \end{aligned}$$

which proves equation 4. Since $R\vartheta = 0$ implies $P_s R\vartheta = 0$, the kernel inclusion is automatic. If the coefficient space has dimension C, rank–nullity yields

$$\begin{aligned} \dim \frac{\ker(P_s R)}{\ker(R)} &= \{C - \mathrm{rank}(P_s R)\} - \{C - \mathrm{rank}(R)\} \\ &= \Delta(\mathfrak{B}, s). \end{aligned}$$

If $\Delta = 0$, the two nested affine spaces have equal dimension and are therefore equal. Select $d+1$ affinely independent ray points, where $d = \dim \mathcal{A}_{\text{common}}$, and put positive mass on all of them. The proof of Lemma A.3 then gives a relative neighbourhood in which every common-population law has an exact positive-ray representation.

Now let $Q_\varepsilon = p_0 + \sum_{k\geq 1} \varepsilon^k h_k$, with $h_k \in T_{\text{common}}$, and let $r_* < \infty$ be the first index for which $h_{r_*} \notin T_{\text{ray}}$. By Lemma A.6, choose a ray witness L satisfying $\ell_L(h_{r_*}) \neq 0$. Since $L(p_0) = 0$ and $\ell_L(h_k) = 0$ for every $k < r_*$,

$$L(Q_\varepsilon) = \varepsilon^{r_*} \ell_L(h_{r_*}) + o(\varepsilon^{r_*}). \tag{34}$$

This is the asserted expansion with $c = \ell_L(h_{r_*}) \neq 0$. □

We finish the theorem's conditional testing assertion. Suppose bundle j receives K_j fresh evaluators, with $K_j/K \to \lambda_j > 0$. For block-zero-sum directions define the Fisher inner product

$$\langle h, g \rangle_{p_0,\lambda} = \sum_j \lambda_j \sum_y \frac{h_{jy} g_{jy}}{p_{0,jy}}. \tag{35}$$

Let Π_T be its orthogonal projection onto T_{ray}, and put

$$h_{r_*,\text{eff}} = (I - \Pi_T) h_{r_*} \neq 0.$$

Lemma A.7 (Finite-cell nuisance projection)**.** *Assume all cells of p_0 are bounded below by a positive constant, the chosen ray simplex contains a relative ball about p_0, and $p_0 + \Pi_T(Q_\varepsilon - p_0)$ remains in that simplex. Then*

$$\inf_{P \in H_0^{\text{ray}}(s)} \mathrm{KL}_\lambda(Q_\varepsilon \| P) = \frac{1}{2} \varepsilon^{2r_*} \|h_{r_*,\text{eff}}\|^2_{p_0,\lambda} + o(\varepsilon^{2r_*}). \tag{36}$$

Proof. Write $d_\varepsilon = Q_\varepsilon - p_0$, $t_\varepsilon = \Pi_T d_\varepsilon$, and $n_\varepsilon = (I - \Pi_T) d_\varepsilon$. The interior-simplex assumption makes $P^*_\varepsilon = p_0 + t_\varepsilon$ an actual ray-null law. Uniformly for local $u \in T_{\text{ray}}$, the finite-alphabet Taylor formula is

$$\mathrm{KL}_\lambda(p_0 + d_\varepsilon \| p_0 + u) = \frac{1}{2} \|d_\varepsilon - u\|^2_{p_0,\lambda} + o(\|d_\varepsilon - u\|^2). \tag{37}$$

Orthogonality gives

$$\|d_\varepsilon - u\|^2 = \|t_\varepsilon - u\|^2 + \|n_\varepsilon\|^2,$$

so $u = t_\varepsilon$ attains the quadratic minimum. This also gives an upper bound tending to zero. Any approximate global minimizer must then tend to p_0, by Pinsker's inequality and the positive limiting stratum proportions, and hence enters the local expansion; this proves the matching lower bound. Finally, projection removes all h_k with $k < r_*$, giving $n_\varepsilon = \varepsilon^{r_*} h_{r_*,\text{eff}} + o(\varepsilon^{r_*})$. □

For completeness, consider a compactly indexed local class satisfying the conditions stated in Theorem 4.1, uniformly, and suppose its normal gap obeys

$$c_- \varepsilon^{r_*} \leq \|(I - \Pi_T)(Q - p_0)\|_{p_0,\lambda} \leq c_+ \varepsilon^{r_*}. \tag{38}$$

The lower bound follows by selecting a path in the class and its in-plane null P^*_ε. Lemma A.7 and product additivity give $\mathrm{KL}(Q^K \| (P^*_\varepsilon)^K) \leq CK\varepsilon^{2r_*}$. When $\varepsilon = o(K^{-1/(2r_*)})$, Pinsker's inequality makes the two product laws indistinguishable. At a fixed risk pair (α, β), choosing the constant small enough makes their total variation less than $1 - \alpha - \beta$, precluding the requested size and power.

For the upper bound, let $\widehat{Q}$ be the stacked empirical cell frequencies and use

$$T_K = \|(I - \Pi_T)(\widehat{Q} - p_0)\|_{p_0,\lambda}.$$

Fixed dimension, the uniform cell lower bound, and the positive lower bound on K_j/K imply $\sup_{P \in H_0^{\text{ray}}} E_P T_K^2 \leq C_0/K$. Markov's inequality gives a level-α threshold $(C_0/\alpha K)^{1/2}$; the same second-moment bound and equation 38 give power at least $1-\beta$ once $\varepsilon \geq CK^{-1/(2r_*)}$. This proves the conditional minimax statement under the stated uniform assumptions.

A.4 CANONICAL THREE-STIMULUS GEOMETRY

We now prove Benchmark 4.2. Throughout this subsection,

$$\phi(t) = \tanh(t/2), \qquad \phi'(t) = \tfrac{1}{2}\{1 - \phi(t)^2\}.$$

Lemma A.8 (The pairwise outer set)**.** *Under a shared ordering and conditional factorization,* $(\mu_e, \mu_f, \tau_{ef}) \in \mathcal{O}_2$*. Moreover, in a neighbourhood of* $(0, m, 0)$*, for any fixed* $m \in (0, 1)$*, membership reduces in the* (μ_e, τ_{ef}) *coordinates to*

$$\mathcal{W} = \{(x, z) : x \geq 0,\ 0 \leq z \leq x\} \cup \{(x, z) : x \leq 0,\ x \leq z \leq 0\}. \tag{39}$$

Proof. Orient the two edges consistently with a witnessing common order and write their nonnegative conditional means as $U, W \in [0, 1]$. Conditional factorization gives oriented population moments $(a, b, c) = (EU, EW, E[UW])$. Pointwise,

$$0 \leq UW \leq \min(U, W), \qquad UW \geq U + W - 1.$$

Taking expectations places (a, b, c) in $\mathcal{P}_2$; restoring the two orientation signs gives $\mathcal{O}_2$.

The union representation is equivalent to the absolute-value description in the main text. Membership in a signed component immediately gives $xyz \geq 0$ and the three inequalities for $(|x|, |y|, |z|)$. Conversely, given those inequalities, choose the signs of the two nonzero singleton means (arbitrarily when one is zero); $xyz \geq 0$ supplies the compatible interaction sign, and the absolute values form a point of $\mathcal{P}_2$.

Near $(0, m, 0)$, the sign of the second mean is fixed and positive. By shrinking the neighbourhood, $|x| + |y| - 1 < 0$, so the lower Fréchet constraint is inactive. The sign condition and $|z| \leq |x|$ then give exactly equation 39. □

The containment is not sufficient for an item-potential shared-order law. To see this directly for $e = (1, 2)$, $f = (1, 3)$, consider $(\mu_e, \mu_f, \tau) = (1/4, 1/4, 0)$, which belongs to $\mathcal{O}_2$. In any shared order producing the two positive means, item 1 is above items 2 and 3. Either the conditional oriented means satisfy $U_f \geq U_e$ almost surely or $U_e \geq U_f$ almost surely. Their equal expectations force $U_e = U_f$ almost surely, whence conditional factorization gives $\tau = E[U_e^2] \geq (EU_e)^2 = 1/16$, a contradiction.

Let

$$s_0 = \frac{(1, 1, -2)}{\sqrt{6}}, \qquad q_0 = \frac{(1, -1, 0)}{\sqrt{2}}, \qquad n_\delta = \sqrt{1 + \delta^2/2}, \tag{40}$$

and define

$$s_\delta = \frac{s_0 + \delta q_0/\sqrt{2}}{n_\delta}, \qquad q_\delta = \frac{q_0 - \delta s_0/\sqrt{2}}{n_\delta}. \tag{41}$$

These are unit vectors in $\mathbb{U}$, with $s_\delta^\top q_\delta = 0$. For $e = (1, 2)$, $f = (1, 3)$, and $a_\circ = \sqrt{3/2}$, direct calculation gives

$$a_e(\delta) := x_e^\top s_\delta = \frac{\delta}{n_\delta}, \qquad b_e(\delta) := x_e^\top q_\delta = \frac{\sqrt{2}}{n_\delta}, \tag{42}$$

$$a_f(\delta) := x_f^\top s_\delta = \frac{a_\circ + \delta/2}{n_\delta}, \qquad b_f(\delta) := x_f^\top q_\delta = \frac{1/\sqrt{2} - (\sqrt{3}/2)\delta}{n_\delta}. \tag{43}$$

In particular, the anchor ray margin is exactly δ/n_δ.

Let $A \sim G_0$ be supported in $[A_{\min}, A_{\max}] \subset (0, \infty)$, and assume $\Pr(Z = 1 \mid A) = \Pr(Z = -1 \mid A) = 1/2$. Define

$$m_0 = E\phi(Aa_\circ), \qquad a_1 = \tfrac{1}{2}EA, \qquad \beta = \tfrac{1}{2}E\{A\phi(Aa_\circ)\}, \qquad \kappa = \tfrac{1}{2}E\phi'(Aa_\circ). \tag{44}$$

Then

$$0 < m_0 < 1, \qquad 0 < \beta < a_1, \qquad \kappa > 0. \tag{45}$$

Lemma A.9 (Uniform canonical expansion)**.** *Uniformly over the stated compact support and sufficiently small* (ε, δ)*,*

$$\mu_e = a_1\delta + O(\delta^3 + \varepsilon^2|\delta| + \varepsilon^4|\delta|), \tag{46}$$

$$\mu_f = m_0 + O(|\delta| + \varepsilon^2), \tag{47}$$

$$\tau_{ef} = \beta\delta + \kappa\varepsilon^2 + O(\delta^2 + \varepsilon^2|\delta| + \varepsilon^4). \tag{48}$$

The constants may be chosen uniformly over all laws supported on the fixed interval.

Proof. Conditional on A, put

$$r_e = Aa_e(\delta), \quad r_f = Aa_f(\delta), \quad c_e = b_e(\delta), \quad c_f = b_f(\delta).$$

Conditional symmetry gives, for every smooth h,

$$\frac{h(r + \varepsilon c) + h(r - \varepsilon c)}{2} = h(r) + \frac{\varepsilon^2 c^2}{2} h''(r) + O(\varepsilon^4), \tag{49}$$

and, for the product,

$$\begin{aligned} &\frac{1}{2} \sum_{z\in\{-1,1\}} \phi(r_e + z\varepsilon c_e)\phi(r_f + z\varepsilon c_f) \\ &= \phi(r_e)\phi(r_f) + \varepsilon^2 \left\{ \frac{c_e^2}{2}\phi''(r_e)\phi(r_f) + \frac{c_f^2}{2}\phi(r_e)\phi''(r_f) + c_e c_f \phi'(r_e)\phi'(r_f) \right\} + O(\varepsilon^4). \end{aligned} \tag{50}$$

At $\delta = 0$, $r_e = 0$, $c_e c_f = 1$, $\phi(0) = \phi''(0) = 0$, and $\phi'(0) = 1/2$. Expanding equations 42–43 around zero therefore gives the coefficients in equation 44. Taylor's theorem gives the displayed remainders. Compactness bounds A and all required derivatives, so the same constants survive integration over G_0. □

The even expansion in equation 49 makes the role of symmetry explicit: conditional sign symmetry removes every odd power of the residual amplitude in this benchmark, and the quotient witness first appears at order two. Without that cancellation, the first derivative gives $r_* = 1$ whenever it has a component outside T_{ray}; if several successive derivatives remain tangent, Theorem 4.1 uses the next surviving order, including $r_* > 2$. Compact support supplies the uniform remainder used here. For a fixed unbounded G_0, analogous pointwise differentiation under the integral requires suitable moments; uniform statements over a class additionally need a common integrability envelope.

Lemma A.10 (Limit gap)**.** *If* $\delta/\varepsilon^2 \to r$*, then*

$$\Gamma_{ef} = \varepsilon^2 G_{G_0}(r) + o(\varepsilon^2), \tag{51}$$

where, for $r \geq 0$*,*

$$G_{G_0}(r) = \frac{1}{2}[\kappa - (a_1 - \beta)r]_+, \tag{52}$$

and, for $r = -u < 0$*, writing* $b(u) = \kappa - \beta u$*,*

$$G_{G_0}(-u) = \begin{cases} 0, & b(u) \leq 0, \\ \{b(u) + \min(a_1 u, b(u))\}/2, & b(u) > 0. \end{cases} \tag{53}$$

Thus $G_{G_0}(0) = \kappa/2$*, and its positive region is precisely*

$$-\frac{\kappa}{\beta} < r < \frac{\kappa}{a_1 - \beta}. \tag{54}$$

Proof. Lemma A.9 gives

$$\varepsilon^{-2}(\mu_e, \tau_{ef}) \longrightarrow (a_1 r, \beta r + \kappa).$$

The probe mean remains in a compact subinterval of $(0, 1)$, so Lemma A.8 applies. Distance to the closed cone $\mathcal{W}$ is continuous and positively homogeneous.

For $r \geq 0$, both limiting coordinates are nonnegative. The distance of (x, z) to $0 \leq z \leq x$ is $[z - x]_+/2$ in the sup norm, giving equation 52. For $r = -u < 0$, the limit is $(-a_1 u, b(u))$. If $b(u) \leq 0$, then $-a_1 u \leq b(u) \leq 0$, so the point lies in the negative wedge. If $b(u) > 0$, its distances to the negative and positive wedges have minimum $\{b(u) + \min(a_1 u, b(u))\}/2$. This proves equation 53; the remaining claims use equation 45. □

Lemma A.10 proves equation 7 and the side-dependent critical interval in Benchmark 4.2. We next prove the asserted exact-overlap hard sequence, including the complete three-slot law rather than only its pair marginal.

For deterministic temperature t, set

$$x_\delta(t) = \phi(ta_e(\delta)), \qquad y_\delta(t) = \phi(ta_f(\delta)), \qquad g(t) = \phi(a_\circ t).$$

The augmented ray-moment columns for B_2 and B_3 are

$$R_\delta^{(2)}(t) = (1, x_\delta, y_\delta, x_\delta y_\delta)^\top, \tag{55}$$

$$R_\delta^{(3)}(t) = (1, x_\delta, y_\delta, y_\delta^2, x_\delta y_\delta, x_\delta y_\delta^2)^\top. \tag{56}$$

At $\delta = 0$, the augmented B_2 ray columns have rank two and the augmented B_3 columns have rank three, since they reduce respectively to $(1, 0, g, 0)$ and $(1, 0, g, g^2, 0, 0)$. Their common-law augmented ranks are four and six: the two edge margins range over an open rectangle, and the polynomials $1, x, y, xy$ and $1, x, y, y^2, xy, xy^2$ are linearly independent. Thus the exact-collision witness bases may be chosen as

$$(\mu_e, \tau_{ef}) \quad \text{for } B_2, \qquad (\mu_e, \tau_{ef}, \zeta_{eff'}) \quad \text{for } B_3.$$

This is the additional collision witness supplied by the repeated probe.

Lemma A.11 (Fixed supports through a shrinking collision)**.** *Inside any open positive-temperature interval, one can choose four distinct supports for which the evaluation matrix of*

$$1, \quad t/2, \quad g(t), \quad tg(t)/2$$

is nonsingular, and six distinct supports for which the evaluation matrix of

$$1, \quad t/2, \quad g(t), \quad g(t)^2, \quad tg(t)/2, \quad tg(t)^2/2$$

is nonsingular. At those fixed supports, the matrices whose columns are equation 55 or equation 56 are nonsingular for all sufficiently small $\delta \neq 0$, and their inverses have norm $O(|\delta|^{-1})$.

Proof. Both function lists are linearly independent on every positive open interval. Indeed, a linear relation can be written $P(g(t)) + tQ(g(t)) = 0$, with P, Q of degree at most one for B_2 and at most two for B_3. If $Q \neq 0$, substituting $t = 2\,\mathrm{artanh}(g)/a_\circ$ makes $\mathrm{artanh}(g)$ a rational function on an interval. Analytic continuation contradicts its logarithmic singularity at $g = 1$. Thus $Q = 0$, and then $P = 0$. The standard evaluation lemma for independent continuous functions—proved by selecting points inductively while increasing evaluation rank—gives the desired supports.

At each fixed support,

$$\frac{x_\delta(t)}{\delta} \to t/2, \qquad y_\delta(t) \to g(t), \qquad \frac{x_\delta(t) y_\delta(t)^k}{\delta} \to tg(t)^k/2.$$

Thus the two augmented matrices have factorizations

$$B_\delta^{(2)} = \mathrm{diag}(1, \delta, 1, \delta)\{\overline{B}^{(2)} + O(\delta)\},$$

$$B_\delta^{(3)} = \mathrm{diag}(1, \delta, 1, 1, \delta, \delta)\{\overline{B}^{(3)} + O(\delta)\},$$

with nonsingular limit matrices. Inverting these factorizations proves the claim. □

Choose the six-support construction, assign weights $w_j > 0$, and use this fixed G_0 in the canonical alternative. Conditional independence and exchangeability imply that the complete B_3 law is determined by

$$(\mu_e, \mu_f, r_f, \tau_{ef}, \zeta_{eff'}), \tag{57}$$

where $r_f = E[D_f D_{f'}]$ and $\zeta_{eff'} = E[D_e D_f D_{f'}]$. This follows directly from Fourier inversion:

$$p(d_e, d_f, d_{f'}) = \frac{1}{8}\big[1 + \mu_e d_e + \mu_f(d_f + d_{f'}) + r_f d_f d_{f'} + \tau_{ef} d_e(d_f + d_{f'}) + \zeta_{eff'} d_e d_f d_{f'}\big]. \tag{58}$$

The corresponding three-moment formula determines the complete B_2 law.

Relative to the base ray mixture at the same s_δ, the symmetric residual changes the five moments by

$$\Delta\mu_e = O(\varepsilon^2\delta), \qquad \Delta\mu_f = O(\varepsilon^2), \qquad \Delta r_f = O(\varepsilon^2), \tag{59}$$

$$\Delta\tau_{ef} = \kappa\varepsilon^2 + O(\varepsilon^2|\delta| + \varepsilon^4), \qquad \Delta\zeta_{eff'} = \eta\varepsilon^2 + O(\varepsilon^2|\delta| + \varepsilon^4), \tag{60}$$

where $\eta = E\{g(A)\phi'(Aa_\circ)\} > 0$. These formulas follow from equation 49 applied to $\phi(r_f)^2$ and from equation 50 applied to $\phi(r_e)\phi(r_f)^2$; compact finite support makes the remainders uniform.

For $\delta \neq 0$, affine weights on the selected supports are unique. The matrix factorization in Lemma A.11 and equations 59– 60 give

$$\|w(\varepsilon,\delta) - w\| = O\left(\varepsilon^2 + \frac{\varepsilon^2}{|\delta|}\right). \tag{61}$$

If $|\delta|/\varepsilon^2 \to \infty$, this correction tends to zero, so all weights remain positive. Equation 58 then proves exact equality of all eight cells with one positive-ray mixture; its B_2 marginal is exact as well. If $|\delta|/\varepsilon^2 \to 0$, Lemma A.10 instead gives $\Gamma_{ef} \sim (\kappa/2)\varepsilon^2$, excluding every shared-order law, hence every ray mixture.

When $\varepsilon^2/\delta \to \lambda$ is finite, multiplying the normal shift in equation 60 by the limiting inverse matrix gives

$$w(\varepsilon,\delta) = w + \lambda c^{(3)} + o(1), \qquad c^{(3)} = (\overline{B}^{(3)})^{-1}(0,0,0,0,\kappa,\eta)^\top.$$

The normalization row gives $\mathbf{1}^\top c^{(3)} = 0$, while $c^{(3)} \neq 0$; it therefore has both signs. Positivity holds on a finite open interval $\{\lambda : w_j + \lambda c_j^{(3)} > 0\ \forall j\}$, with endpoints depending on the selected supports, weights, and sign of δ. This proves the benchmark's hard-sequence and critical-band assertions for the selected supports.

A.4.1 Repeated-probe calibration and fourth-root testing

At exact collision, put $g(A) = \phi(Aa_\circ)$. Under conditional factorization,

$$E[D_f D_{f'}] = E[g(A)^2].$$

Since $\phi'(t) = \{1 - \phi(t)^2\}/2$,

$$\kappa = \frac{1}{2}E\phi'(Aa_\circ) = \frac{1}{4}\{1 - E[D_f D_{f'}]\},$$

which proves equation 8. This calculation uses the base collision law and factorization; away from it, a Taylor expansion of $E[D_f D_{f'}]$ has generally nonzero $O(\delta)$ and $O(\varepsilon^2)$ biases.

Proposition A.12 (Finite-sample repeated-probe calibration)**.** *For K iid fresh evaluators, let*

$$\widehat{r}_f = K^{-1}\sum_{k=1}^{K} D_{k,f}D_{k,f'}, \qquad h_{K,\eta} = \sqrt{2\log(2/\eta)/K}.$$

At exact collision,

$$\underline{\kappa}_K^{\text{exact}} = \left[\frac{1 - \widehat{r}_f - h_{K,\eta}}{4}\right]_+ \tag{62}$$

satisfies $\Pr\{\underline{\kappa}_K^{\text{exact}} \le \kappa(G_0)\} \ge 1 - \eta$. For the compact near-collision class, choose a known constant C_r such that

$$\left|\frac{1 - r_f(\varepsilon,\delta)}{4} - \kappa(G_0)\right| \le C_r(|\delta| + \varepsilon^2). \tag{63}$$

If an independent pilot, a design constraint, or a simultaneous confidence procedure supplies $\overline{b}_{\text{cal}}$ with

$$\Pr\{|\delta| + \varepsilon^2 \le \overline{b}_{\text{cal}}\} \ge 1 - \eta_{\text{cal}},$$

then

$$\underline{\kappa}_K^{\text{near}} = \left[\frac{1 - \widehat{r}_f - h_{K,\eta}}{4} - C_r\overline{b}_{\text{cal}}\right]_+ \tag{64}$$

obeys $\Pr\{\underline{\kappa}_K^{\text{near}} \le \kappa(G_0)\} \ge 1 - \eta - \eta_{\text{cal}}$.

Proof. The variables $D_{k,f}D_{k,f'}$ lie in $[-1,1]$, so Hoeffding's inequality gives $|\widehat{r}_f - r_f| \le h_{K,\eta}$ with probability at least $1-\eta$. At collision combine this event with $\kappa = (1-r_f)/4$. In the near-collision class, intersect it with the stated calibration event and use equation 63; a union bound gives the claimed coverage. The truncation at zero preserves both inequalities because $\kappa(G_0) > 0$. □

We finally establish the stated fourth-root boundary. Fix a compactly supported G_* and define its coefficients by equation 44. Let

$$c_{\mathrm{crit}} = \min\left\{\frac{\kappa}{a_1-\beta}, \frac{\kappa}{\beta}\right\}, \qquad 0 < c_0 < c_{\mathrm{crit}}. \tag{65}$$

Continuity of G_{G_*}, its strict positivity on the compact interval $[-c_0, c_0]$, and the uniform Taylor remainder imply constants $c_\Gamma, \varepsilon_0 > 0$ such that

$$\inf_{|\delta| \le c_0\varepsilon^2} \Gamma_{ef} \ge c_\Gamma \varepsilon^2, \qquad 0 < \varepsilon \le \varepsilon_0. \tag{66}$$

For K iid fresh evaluators, estimate the three pair moments by $\widehat{\theta}$ and reject when

$$\mathrm{dist}_\infty(\widehat{\theta}, \mathcal{O}_2) > r_{K,\alpha}, \qquad r_{K,\alpha} = \sqrt{2\log(6/\alpha)/K}. \tag{67}$$

Each coordinate is an average of $[-1,1]$-valued iid variables. Hoeffding's inequality and a union bound give

$$\Pr\{\|\widehat{\theta} - \theta\|_\infty > r\} \le 6e^{-Kr^2/2}. \tag{68}$$

Lemma A.8 and the 1-Lipschitz property of distance therefore give finite-sample size at most α. Equations 66– 68 give minimum power at least $1-\beta$ once $c_\Gamma\varepsilon^2 \ge r_{K,\alpha} + r_{K,\beta}$. The same test applies to B_3 after marginalizing f'.

For the matching lower bound, compare the exact-tie independent alternative $Q^{(m)}_{\varepsilon,0}$, $m = 2, 3$, with the fixed ray null $P^{(m)}_0 = Q^{(m)}_{0,0}$. Conditional symmetry and compact support give a cellwise expansion

$$Q^{(m)}_{\varepsilon,0} = P^{(m)}_0 + \varepsilon^2 h_m + O(\varepsilon^4). \tag{69}$$

All base cells are bounded away from zero, and $h_m \ne 0$ because its D_eD_f moment is $\kappa > 0$. The finite-cell logarithm expansion gives

$$\mathrm{KL}(Q^{(m)}_{\varepsilon,0} \| P^{(m)}_0) = \frac{1}{2} J_m \varepsilon^4 + o(\varepsilon^4), \qquad J_m = \sum_y \frac{h_m(y)^2}{P^{(m)}_0(y)} > 0. \tag{70}$$

For K evaluators, product KL is $O(K\varepsilon^4)$. Pinsker's inequality makes null and alternative total variation vanish when $K\varepsilon^4 \to 0$, and gives the fixed-risk small-constant lower bound when $\varepsilon \le cK^{-1/4}$. Combined with the preceding uniform upper bound, this proves $\varepsilon_{*,K} \asymp K^{-1/4}$ in the stated relative class.

The repeated observation can increase the fixed-null information constant, but not the exponent. Indeed the B_2 score is the conditional expectation of the B_3 score given (D_e, D_f), so variance decomposition gives $J_3 - J_2 = E_0\{\mathrm{Var}(S_3 \mid D_e, D_f)\} \ge 0$. To see strictness, write ℓ and ρ_2 for the second-order coefficients of μ_f and r_f. For $D_f = 1$, constancy of S_3 across $D_{f'} = 1, -1$, separately for each D_e, would require

$$\frac{2\ell + \rho_2 + D_e(2\kappa+\eta)}{1 + 2m_0 + r_0} = \frac{-\rho_2 - D_e\eta}{1 - r_0}.$$

Subtracting the equation for $D_e = -1$ from that for $D_e = 1$ gives

$$\frac{2\kappa+\eta}{1+2m_0+r_0} = -\frac{\eta}{1-r_0},$$

which is impossible because every denominator and κ, η are positive. Hence $J_3 > J_2$ for this fixed-null canonical path. This is a simple-vs-simple per-evaluator comparison only; it supplies no nuisance-uniform per-judgment ordering.

A.5 DEPENDENCE COUNTERFEIT AND ROBUST BOUNDARY

We first record the complete conditional-moment algebra. Orient the anchor and probe consistently with a shared order and write the three signed responses as Y_1, Y_2, Y_3, with conditional means (x, y, y). Let $\xi_1 = Y_1 - x$, $\xi_2 = Y_2 - y$, and $\xi_3 = Y_3 - y$, and define

$$c_{ij} = E[\xi_i \xi_j \mid V], \qquad c_{123} = E[\xi_1 \xi_2 \xi_3 \mid V].$$

Expanding products and using $E[\xi_j \mid V] = 0$ gives

$$E[Y_1 Y_2 \mid V] = xy + c_{12}, \qquad E[Y_1 Y_3 \mid V] = xy + c_{13}, \tag{71}$$

$$E[Y_2 Y_3 \mid V] = y^2 + c_{23}, \tag{72}$$

$$E[Y_1 Y_2 Y_3 \mid V] = xy^2 + xc_{23} + y(c_{12} + c_{13}) + c_{123}. \tag{73}$$

Equation 73 is equation 9. In particular, a third raw-moment error cannot be represented by c_{123} alone when the pair errors are nonzero.

A.5.1 EXACT EIGHT-CELL CONSTRUCTION

At exact collision let $g(a) = \phi(aa_\circ)$. Choose three distinct positive temperatures a_1, a_2, a_3 and put $g_j = g(a_j)$. The matrix

$$H = \begin{pmatrix} 1 & 1 & 1 \\ g_1 & g_2 & g_3 \\ g_1^2 & g_2^2 & g_3^2 \end{pmatrix} \tag{74}$$

is a Vandermonde matrix in the distinct g_j's and is nonsingular. Fix weights $w_j > 0$ summing to one and let $G_0 = \sum_j w_j \delta_{a_j}$. Under the canonical independent alternative using this same G_0, denote the exact five moments by

$$\mu_e^\varepsilon = 0, \quad \mu_f^\varepsilon, \quad r_f^\varepsilon, \quad \tau_\varepsilon, \quad \zeta_\varepsilon.$$

The expansions in equations 59– 60, now at $\delta = 0$, imply

$$\mu_f^\varepsilon = \sum_j w_j g_j + O(\varepsilon^2), \qquad r_f^\varepsilon = \sum_j w_j g_j^2 + O(\varepsilon^2), \tag{75}$$

$$\tau_\varepsilon = \kappa \varepsilon^2 + O(\varepsilon^4), \qquad \zeta_\varepsilon = \eta \varepsilon^2 + O(\varepsilon^4), \tag{76}$$

with $\kappa, \eta > 0$. Define exact adjusted weights by

$$(w_1(\varepsilon), w_2(\varepsilon), w_3(\varepsilon))^\top = H^{-1}(1, \mu_f^\varepsilon, r_f^\varepsilon)^\top. \tag{77}$$

They equal $w + O(\varepsilon^2)$ and are therefore strictly positive for all sufficiently small ε.

At support a_j, define the following conditional distribution on $\{-1, 1\}^3$:

$$p_j^\varepsilon(d_e, d_f, d_{f'}) = \frac{1}{8}\big[1 + g_j(d_f + d_{f'}) + g_j^2 d_f d_{f'} + \tau_\varepsilon d_e(d_f + d_{f'}) + \zeta_\varepsilon d_e d_f d_{f'}\big]. \tag{78}$$

Lemma A.13 (Cell positivity of the counterfeit kernel)**.** *The kernel in equation 78 is normalized, exchangeable in f, f', and nonnegative whenever*

$$|2\tau_\varepsilon + \zeta_\varepsilon| \le (1 + g_j)^2, \tag{79}$$

$$|\zeta_\varepsilon| \le 1 - g_j^2, \tag{80}$$

$$|\zeta_\varepsilon - 2\tau_\varepsilon| \le (1 - g_j)^2. \tag{81}$$

All three inequalities are strict for every selected support when ε is sufficiently small.

Proof. When $d_f = d_{f'} = 1$, multiplying the two cells indexed by $d_e = \pm 1$ by eight gives $(1+g_j)^2 \pm (2\tau_\varepsilon + \zeta_\varepsilon)$. When $d_f \neq d_{f'}$, the corresponding expressions are $1 - g_j^2 \pm \zeta_\varepsilon$; when $d_f = d_{f'} = -1$, they are $(1 - g_j)^2 \pm (\zeta_\varepsilon - 2\tau_\varepsilon)$. This proves nonnegativity under equations 79– 81. All nonconstant Walsh terms sum to zero, proving normalization, and the formula is visibly exchangeable. The finite set of g_j's is contained in $(0, 1)$, whereas $\tau_\varepsilon, \zeta_\varepsilon = O(\varepsilon^2)$, so strict feasibility follows for small ε. □

Proof of Theorem 5.1. Use latent utilities $V = a_j s_0$ with mixing weights equation 77, and use equation 78 conditionally on support j. Fourier inversion shows that its conditional moments are

$$E[D_e \mid a_j] = 0, \qquad E[D_f \mid a_j] = E[D_{f'} \mid a_j] = g_j,$$
$$E[D_f D_{f'} \mid a_j] = g_j^2, \qquad E[D_e D_f \mid a_j] = E[D_e D_{f'} \mid a_j] = \tau_\varepsilon,$$
$$E[D_e D_f D_{f'} \mid a_j] = \zeta_\varepsilon.$$

Thus every conditional marginal equals its logistic mean on the ray. In the centred notation above, the dependence errors are exactly

$$c_{12} = c_{13} = \tau_\varepsilon, \qquad c_{23} = 0, \qquad c_{123}(a_j) = \zeta_\varepsilon - 2 g_j \tau_\varepsilon. \tag{82}$$

The last identity follows either from equation 73 with $x = 0, y = g_j$, or by direct expansion.

After mixing over j, equation 77 matches (μ_f, r_f) exactly, while the conditional construction matches (μ_e, τ, ζ) exactly. The five-moment Fourier formula equation 58 therefore gives equality of all eight labelled cells. Lemma A.13 proves that the kernel is legal. The adjusted weights differ from the base weights by $O(\varepsilon^2)$, and every error in equation 82 is $O(\varepsilon^2)$.

The ray direction s_0 ties items 1 and 2, so this is an exact member of the weak-order null. For a strict-order-only null, take $\delta_n > 0$ with $\delta_n \to 0$ and use $V = a_j s_{\delta_n}$. Let its two logistic means be $x_{jn} > 0$ and y_{jn}. Starting from the same centred errors in equation 82, define the remaining conditional raw moments through equations 71– 73. Fourier inversion then gives conditional cells that converge entrywise to equation 78; because the latter cells are strictly positive, all approximating kernels are legal for large n and preserve the new logistic marginal means. Keeping the finite mixing weights fixed produces strict-null laws converging to the exact counterfeit law. Exact equality is impossible at finite n: after orienting a strict shared order, the anchor conditional mean is positive almost surely and hence its population mean cannot equal the alternative's exact zero. This proves closure, and only closure, for the strict-order null.

It remains to prove equation 10. The construction gives the upper bound $b_*(Q_\varepsilon) \le C\varepsilon^2$. Conversely, in any shared-order representation of the same law, orient the anchor so its conditional mean $x(V) \ge 0$. The observed anchor mean is zero, so $Ex(V) = 0$ and therefore $x(V) = 0$ almost surely. By equation 71,

$$\tau_\varepsilon = E[c_{12}(V)].$$

Hence

$$\|c_{12}\|_\infty \ge |\tau_\varepsilon| \ge (\kappa/2)\varepsilon^2$$

for small ε, using equation 76. Taking the infimum over all legal representations proves $b_*(Q_\varepsilon) = \Theta(\varepsilon^2)$. □

Marginalizing f' in this construction proves the analogous B_2 counterfeit: the conditional two-cell kernel has Fourier form

$$\tfrac{1}{4}\{1 + g_j d_f + \tau_\varepsilon d_e d_f\},$$

and is positive for small ε. It also has minimal pair-error radius $\Theta(\varepsilon^2)$. This hard sequence rules out a uniform guarantee under an unrestricted dependence budget.

A.5.2 Known budget: robust upper and matching lower bounds

We make the nuisance container in Theorem 5.2 explicit. Fix the preceding supports in $[A_{\min}, A_{\max}]$, and let $\mathcal{W}_G$ be a compact weight polytope around w such that every coordinate is at least $w_{\min} > 0$. Set

$$\mathcal{G} = \left\{ \sum_j v_j \delta_{a_j} : v \in \mathcal{W}_G \right\}, \tag{83}$$

choosing the polytope large enough to contain the $O(\varepsilon^2)$ corrections above. Compactness supplies uniform cell lower bounds, Taylor constants, moment-polygon interior, and positive lower bounds for κ, β, and $a_1 - \beta$.

Lemma A.14 (Budgeted null containment)**.** *For $0 \le b \le 1$, define*

$$L_b(a,d) = \max\{a+d-1, -b\},$$
$$U_b(a,d) = \min\{1-|a-d|,\ b+\sqrt{1-b}\min(a,d)\}, \tag{84}$$
$$\mathcal{R}_2(b) = \bigcup_{\eta_e,\eta_f\in\{-1,1\}} \{(\eta_e a, \eta_f d, \eta_e\eta_f t) : 0 \le a,d \le 1,\ L_b(a,d) \le t \le U_b(a,d)\}. \tag{85}$$

If a shared-order pair law preserves the logistic marginal means and has $\|c_{12}\|_\infty \le b$, its moment vector lies in $\mathcal{R}_2(b)$. Moreover, $\mathcal{R}_2(b) \subseteq \mathcal{E}_2(b)$, where $\mathcal{E}_2(b)$ is defined in equation 11, and $\mathcal{R}_2(0) = \mathcal{E}_2(0) = \mathcal{O}_2$.

Proof. Orient both edges by a witnessing order. Write their conditional means as $X, Y \in [0,1]$, their conditional product moment as M, and $a = EX, d = EY, t = EM$. Binary cell feasibility and the pointwise budget give

$$X + Y - 1 \le M \le 1 - |X - Y|, \qquad |M - XY| \le b.$$

Hence $t \ge \max\{a+d-1, -b\}$. Put $q = \sqrt{1-b}$. We claim $M \le b + qX$. If $Y \le q$, use $M \le XY + b \le qX + b$. If $Y > q$ and $X \le q$, then

$$M \le 1 - Y + X \le 1 - q + X \le b + qX,$$

where the last inequality is $(1-q)(X-q) \le 0$. If $X > q$, use $M \le 1 \le b + qX$. Exchanging X, Y, taking expectations and using $E|X-Y| \ge |a-d|$ proves $t \le U_b(a,d)$. Restoring the signs proves null containment, including ties.

For fixed orientation, the set $\mathcal{E}_2(b)$ is the interval

$$\max\{a+d-1, -b\} \le t \le \min\{1-|a-d|,\ b+\min(a,d)\}.$$

This follows by expanding the core interval $[\max\{0, a+d-1\}, \min(a,d)]$ by b and intersecting with binary cell feasibility. Since $q \le 1$, the claimed inclusion follows; at $b = 0$ the intervals coincide. Budgets above one can be truncated to one because a binary conditional covariance has absolute value at most one. These are necessary outer sets for the item-potential model. □

Let $\theta^{\mathrm{ind}}_{\varepsilon,\delta}$ be the independent canonical pair vector and $\theta^{\mathrm{dep}}_{\varepsilon,\delta}$ an alternative vector satisfying the same pair budget. Uniformly over $G \in \mathcal{G}$, the canonical expansion and the strict interior choice of c_0 yield constants $c_\Gamma, C_\delta, C_R > 0$ such that

$$\mathrm{dist}_\infty(\theta^{\mathrm{ind}}_{\varepsilon,\delta}, \mathcal{O}_2) \ge c_\Gamma\varepsilon^2 - C_\delta|\delta| - C_R(\varepsilon^4 + \delta^2). \tag{86}$$

By Lemma A.14, $\mathcal{R}_2(b_K) \subseteq \mathcal{E}_2(b_K)$. For every $r \in \mathcal{R}_2(b_K)$, some $o \in \mathcal{O}_2$ therefore differs from r only in the interaction coordinate by at most b_K; the alternative dependence changes its independent interaction by at most another b_K. The triangle inequality therefore gives

$$\mathrm{dist}_\infty(\theta^{\mathrm{dep}}_{\varepsilon,\delta}, \mathcal{R}_2(b_K)) \ge c_\Gamma\varepsilon^2 - C_\delta|\delta| - 2b_K - C_R(\varepsilon^4 + \delta^2). \tag{87}$$

Because $|\delta| \le c_0\varepsilon^2$ and c_0 is strictly inside the uniform positive-gap thresholds, shrink the local neighbourhood if needed to obtain

$$\mathrm{dist}_\infty(\theta^{\mathrm{dep}}_{\varepsilon,\delta}, \mathcal{R}_2(b_K)) \ge \bar{c}_\Gamma\varepsilon^2 - 2b_K \tag{88}$$

for some $\bar{c}_\Gamma > 0$.

Estimate the three pair moments from K iid fresh evaluators and reject when

$$\mathrm{dist}_\infty(\widehat{\theta}, \mathcal{R}_2(b_K)) > r_{K,\alpha}, \qquad r_{K,\alpha} = \sqrt{2\log(6/\alpha)/K}. \tag{89}$$

For B_3, use the (e,f) marginal. By Lemma A.14, the true null vector lies in the tested set. Equation 68 therefore proves finite-sample size at most α. Under the alternative, the reverse triangle inequality, equation 88, and the same concentration bound give power at least $1-\beta$ whenever

$$\bar{c}_\Gamma\varepsilon^2 - 2b_K \ge r_{K,\alpha} + r_{K,\beta}. \tag{90}$$

Thus a sufficiently large constant C_U makes

$$\varepsilon \geq C_U \max\{\sqrt{b_K}, K^{-1/4}\} \tag{91}$$

sufficient, uniformly over the calibrated class.

There are two matching lower bounds. First, the zero-error alternative and fixed exact-tie null both belong to the budgeted classes. Equation 70 gives, for some finite C_m,

$$\mathrm{TV}\{(Q_{\varepsilon,0}^{(m)})^K, (P_0^{(m)})^K\} \leq \sqrt{KC_m\varepsilon^4/2}.$$

Choosing $\varepsilon \leq c_{\mathrm{stat}} K^{-1/4}$ with c_{stat} small enough makes this total variation less than $1-\alpha-\beta$. No level-α test can then have power $1-\beta$ against that member of the alternative class.

Second, the exact construction above has a constant $C_{\mathrm{dep}} < \infty$ such that all four errors in equation 82 are bounded by $C_{\mathrm{dep}}\varepsilon^2$. The B_2 marginal has the same property. Under an isotropic budget, if

$$\varepsilon \leq c_{\mathrm{dep}}\sqrt{b_K}, \qquad c_{\mathrm{dep}} \leq C_{\mathrm{dep}}^{-1/2},$$

the dependent weak-order null is legal and has exactly the same complete law as the zero-error canonical alternative. Its rejection probability is therefore no larger than the null size α.

Proof of Theorem 5.2. Let $\rho_K = \max\{\sqrt{b_K}, K^{-1/4}\}$. The upper assertion follows from equation 91. For the lower assertion, if $\sqrt{b_K} \geq K^{-1/4}$, take the constant no larger than c_{dep} and apply the exact-counterfeit lower bound. If $K^{-1/4} \geq \sqrt{b_K}$, take it no larger than c_{stat} and apply the fixed-null statistical lower bound. These cases exhaust all K, and both use the same evaluator-count cost, nuisance container, canonical alternative class, null, and isotropic budgets. Since $b_K \to 0$, all selected laws eventually lie inside the local neighbourhoods used above. This proves both fixed-risk inequalities and equation 12. □

The upper procedure only tests a necessary pairwise set, so it remains safe for the full B_3 null but may be conservative. The proof also shows why the clean boundary needs an externally specified deterministic isotropic budget: with vector budgets, the exact counterfeit requires capacity in the (12), (13), and third-order coordinates, while increasing b_{23} alone does not reproduce it.

B Computational details

The phase and robustness calculations use Python's standard library. Validators check cell positivity, mass conservation, moments, and complete-law matching; configurations and Monte Carlo seeds accompany the output tables.

Artifact map. Under `experiments/paper_experiments/`, the collision ratios and Figure 2 map to `collision_phase_boundaries.csv` and `fig1_collision_phase.svg`; the robustness counts, Wilson intervals, and power summaries map to `dependence_phase.csv` and `e1_summary.json`. The LLM counts, radius, margins, and Figure 4 map to `e3_linked_law_v5/results/e3_v5_formal_confirmation_summary.json` and its source-hashed SVG; the independent output validator reconstructs retained denominators and confirmation statistics. Table 3 maps to `e4_observation_regime_audit/e4_dataset_matrix.csv`, checked by `validate_e4_audit.py`. Figure 1 maps to `paper/latex/manuscript/figures/camouflage_collision_schematic.svg` and contains no empirical data.

Collision phase. The figure is generated from 42 deterministic rows. There is no random seed. The released machine-readable supplement records the three fixed temperature supports and weights, seven ε values, both signs, and all displayed boundary ratios.

B.1 EXACT COLLISION-RANK DESIGN AUDIT

The standard-library utility under `experiments/paper_experiments/collision_rank_audit/` accepts a finite oriented-edge library and rational candidate directions. It enumerates each labelled cell exponent c_{jy}, deduplicates the cross-block sums d, and constructs the integer multiplicity matrix $R_{\mathfrak{B}}$ from equation 16. Exact rational grouping by $s^\top d$ gives $P_s R_{\mathfrak{B}}$. Gaussian elimination then reports both ranks, Δ, collision groups, and deterministic representatives of the witness quotient. The canonical examples reproduce $(4, 2, 2)$ for B_2 and $(6, 3, 3)$ for B_3, where each triple lists $(\operatorname{rank} R, \operatorname{rank} P_s R, \Delta)$.

The returned witnesses screen proposed libraries and directions. Their signal order r_* is obtained from the intended alternative path; the audit alone does not determine a sampling rate.

Dependence phase. The source grid is seeded by 20260817. It contains four scenarios on 105 grid points each and 800 replicates per point (420 output rows and 336,000 interaction draws in the source suite). The paper plot retains the boundary null and worst alternative. The largest legality error is 1.11×10^{-16} in the unified audit; the broader source validator's maximum probability or moment error is 2.22×10^{-16}.

C CALIBRATION AUDIT AND DETERMINISTIC OVERLAP CERTIFICATE

Earlier fitting experiments. Two latent-type logistic-mixture experiments examined apparent ranking reversals under a shared-order data-generating population. The first failed its positive-control calibration (22/100); the second failed its optimizer certificate gate (2,237 of 24,600 certificates failed). Both are scientifically inconclusive and supply no target evidence. Their full protocols, gate order, and outcomes remain in the supplementary experiment package `experiments/paper_experiments/e2_spurious_plurality/`.

Deterministic overlap certificate. Two explicit finite-atom populations—a shared-order temperature mixture F_{null} and a matched population F_{matched} with $\rho_{\text{ord}} = 0.03$—have the same complete laws on the predeclared anchor–probe and probe–repeat dictionary to maximum error 1.11×10^{-16}. Therefore

$$F_\lambda = (1-\lambda)F_{\text{null}} + \lambda F_{\text{matched}}, \qquad \rho_{\text{ord}}(F_\lambda) = 0.03\lambda, \qquad 0 \le \lambda \le 1,$$

has the same complete linked law throughout. This certifies an inner interval $[0, 0.03]$ for that predeclared finite dictionary; the full identified set remains uncharacterized. All 21 displayed grid points pass, with maximum reversal- identity error 2.88×10^{-16}; convex linearity establishes the continuum. The certificate uses the explicit populations without sampled counts or fitting.

D LLM-EVALUATOR PROTOCOL AND COMPLETE FIXED-STRATUM RESULTS

D.1 PREDECLARED DESIGN AND ESTIMAND

The design separated selection from confirmation. A pilot selected two fixed triples—emotional expression and narrative rhythm—under a predeclared numerical rule. No pilot judgment entered confirmation. Fresh confirmation crossed three requested-model specifications, two triples, six fixed personas, three binary slots, and balanced presentation orientation. The persona pairs favour implicit versus explicit emotion, sensory versus abstract expression, and continuous versus fragmented rhythm. The retained confirmation sample contains 24,624 bundles and 73,872 judgments. The released design and stimulus files specify the persona prompts and all three texts in each comparison triple.

For each fixed model–triple stratum, the six personas form a predeclared equal-weight roster. A bundle shares the known type Z but the three stateless requests have independent histories. Orientation-standardized signed outcomes are symmetrized across the repeated slots. The primary 18-dimensional family contains $(\mu_e, \mu_f, \tau_{ef})$ for six strata. Hoeffding's inequality and a Bonferroni allocation at familywise $\alpha = 0.05$ give the common radius 0.0566238711. The numerical solver brackets $\operatorname{dist}_\infty(\theta, \mathcal{O}_2)$ from above and below; separation requires its certified lower endpoint to exceed the simultaneous radius.

Table 2: Complete primary and $b = 0.20$ sensitivity results. "Distance" is the certified numerical lower endpoint for $\Gamma(0)$; conservative columns subtract the simultaneous familywise radius and truncate at zero.

Model	Triple	Distance	$b = 0$ lower	$b = .20$ lower
Claude	Emotional	0.165691	0.109068	0.023238
Claude	Rhythm	0.467348	0.410724	0.328249
DeepSeek	Emotional	0.012426	0	0
DeepSeek	Rhythm	0.147660	0.091036	0.004487
Luna	Emotional	0.058722	0.002099	0
Luna	Rhythm	0.290936	0.234312	0.154946

The formal result is thus five separated strata at $b = 0$ and four at $b = 0.20$. Seven sensitivity brackets slightly exceed the requested 10^{-6} width but have certified signs; they support sign-stable sensitivity conclusions. The primary $b = 0$ brackets all meet the width target.

D.2 Minimal collection contract

A directly reusable protocol is specified before confirmation: the candidate library and direction, pilot selection rule, primary outer-set statistic, multiplicity allocation, sensitivity grid, and the distinction between evaluator count K and judgment count M. Each confirmation unit receives one bundle. Display side is balanced for every oriented edge, legal slot orders are block-randomized, and f' repeats the same ordered comparison as f; pilot units do not reappear in confirmation.

One row per response slot records: (i) a bundle/evaluator key and analysis stratum Z; (ii) fixed stimulus and oriented-edge identifiers; (iii) slot index, assigned order, and repeat-group identifier; (iv) the displayed left/right mapping; (v) the signed binary outcome and validity status; and (vi) attempt/replacement provenance without response-dependent deletion. These fields reconstruct labelled bundle cells, presentation randomization, K, M, and the separate API-attempt count. Participant or service identifiers can be released as within-study pseudonyms; the statistical requirement is reliable linkage within a bundle and fresh assignment across bundles.

Dependence sensitivity. The positive-budget columns use $\mathcal{R}_2(b)$ from Lemma A.14, not the wider expansion $\mathcal{E}_2(b)$. Each anchor–probe conditional error is bounded pointwise by b. Averaging the repeated-slot pair vectors preserves containment because each fixed-orientation piece of $\mathcal{R}_2(b)$ is convex. Apply equation 13 with the simultaneous radius for all 18 coordinates. The released confirmation summary reports the full grid $b \in \{0, 0.01, 0.02, 0.05, 0.10, 0.20\}$. Interpreting a selected budget as a guarantee requires calibration for the same stratum, presentation protocol, and within-bundle kernel, as described in Section 5.

D.3 Audits and interpretation

The product-law reference has $\Gamma = 0$ by algebraic construction and serves as a solver unit test in every stratum. Repeated-probe agreement and the unused Walsh coordinates serve as descriptive audits outside the rejection rule. Position auditing compares signed response means under left and right placement within fixed persona–slot units. Several differences are large: across the six fixed strata, the mean absolute persona–slot contrast ranges from 0.1033 to 0.9094, and the maximum absolute contrast reaches the full signed-response range of 2. Randomization defines an orientation-averaged estimand in which that effect remains.

The inferential result concerns these pairwise moment projections. Complete bundle cells and repeat agreement are retained for descriptive checks. The released protocol records transport retries separately from the scientific sample: 73,891 attempts yielded the 73,872 retained judgments, with 19 transport errors and no format-invalid responses.

E PUBLIC-RELEASE OBSERVATION-REGIME AUDIT

E.1 SELECTION AND CODING

We reviewed ten purposively selected releases spanning language/LLM preference, image aesthetics and generation, graphic design, and speech. The selection includes both widely used preference releases and datasets chosen because their documentation suggested unusually strong evaluator linkage or session structure. Evidence was checked through August 25, 2026 against primary papers, official data cards, repositories, project pages, and released schemas.

For each release, "present" requires a stable evaluator key; verifiably fixed candidates; a native, separately retained binary response for each edge; at least two fixed edges linked within a common evaluator/session unit; two labelled occasions for the identical edge; and enough trial-order, side, and assignment metadata to reconstruct presentation. "Partial" marks a derivable or incomplete component, "absent" an explicit mismatch, and "unresolved" a field not established by public documentation. A session identifier is only a proxy for common latent state. Even a release satisfying all six fields would still require separate sampling, behavioural, and power checks.

Table 3: **Public-release observation-regime audit.** Core-field documentation status as of August 25, 2026. Y: present; P: partial or derivable; N: absent; U: unresolved from public documentation; –: not applicable. The sample was selected purposively.

Release	ID	Fixed	Binary	Linked	Repeat	Order
AADB	Y	–	N	P	U	U
Community Alignment	Y	P	N	P	N	P
FLICKR-AES	Y	–	N	U	U	U
Chatbot Arena 33K	Y	P	P	N	N	P
MultiPref	Y	Y	P	P	U	P
HelpSteer2-Preference	N	Y	P	N	N	P
PARA	Y	–	N	P	P	P
Pick-a-Pic	Y	P	P	P	U	P
SpeechJudge-Data	Y	Y	P	N	U	P
TASTE	Y	Y	P	Y	N	P

E.2 DATASET-SPECIFIC EVIDENCE

Language and LLM preference. Community Alignment links multi-turn categorical winner choices by annotator, but candidate sets change across turns and a four-way winner is not a vector of separately elicited binary edges (Zhang et al., 2026a). Chatbot Arena 33K records one anonymous battle per row; repeated use of a judge proxy does not create a fixed multi-edge bundle (Zheng et al., 2023). MultiPref releases stable evaluator keys and four aspects for a fixed response pair, but the aspects share one annotation event rather than form distinct preference- graph edges (Miranda et al., 2025). HelpSteer2-Preference releases a fixed pair and ordinal preference strength but no evaluator identifier that links multiple judgments (Wang et al., 2025).

Image and design preference. AADB and FLICKR-AES retain rater-aware scalar aesthetic scores, from which pairwise rankings may be constructed only after collection (Kong et al., 2016; Ren et al., 2017). PARA adds session identifiers and deliberately repeated images, but repeats scalar ratings rather than the same pairwise edge (Yang et al., 2022). Pick-a-Pic records user-linked image-pair choices, yet its interaction generates changing candidates and no documented fixed bundle or true repeat (Kirstain et al., 2023). TASTE comes closest: one designer/session covers all six unordered pairs among four fixed outputs. Its canonical release stores strict ranks derived from those comparisons, not the six raw labelled occasions, and each edge occurs only once (Zhu et al., 2026).

Speech preference. SpeechJudge-Data preserves rater-linked judgments for fixed audio pairs, but the public unit has one naturalness edge and no cross-row bundle identifier; intentional identical-pair repeats and presentation randomization are not established in the released documentation (Zhang et al., 2026b).

The clearest missing design element is a documented repeat of the same binary edge, with both labelled responses retained. The minimum collection schema in Appendix D.2 specifies the information needed to reconstruct the joint law. Detailed access and provenance records remain with the released dataset matrix and source audit.