When Uniform Tie-Breaking Preserves Global Score Structure

Anonymous authors
Paper under double-blind review

Abstract

Pairwise preference optimization learns from binary comparison outcomes, but many annotations also include ties. Prior work has incorporated ties into preference objectives and characterized their effect on population learning targets. Uniform tie-breaking preserves every comparison, but the conditions under which the resulting probabilities admit a single score vector have not been characterized. We establish this compatibility boundary for Davidson comparisons with a shared tie strength through cycle closure, which requires pairwise log-odds to sum to zero around every comparison cycle. We derive the exact residual sign and tight fixed-span extrema, and use inverse-link closure to characterize all feasible tie strengths, including multiple isolated solutions. The analysis explains DPO's population projection and separates reward-scale changes that temperature calibration can correct from distortions in score direction. On controlled graphs and shared policies, calibration reduces source-policy regret, and cycle mismatch remains distinct from the additional fitting cost of parameter sharing.

1 Introduction

Pairwise preference data connect human judgments to post-training objectives for language models (Tang et al., 2024; Meng et al., 2024). Direct preference optimization (DPO) learns a policy directly from ordered response pairs (Rafailov et al., 2023), but preference judgments also include ties. The Davidson model describes these outcomes through latent rewards and a nonnegative tie-strength parameter ν (Davidson, 1970). Tie-aware objectives retain the ternary outcome (Guo et al., 2025; Chen et al., 2025); binary pipelines can instead assign either winner to a tie with equal probability. This uniform tie-breaking preserves every comparison, but the resulting binary probabilities need not belong to the global-score model assumed by standard DPO.

We study when uniformly resolved ties still admit the global score differences required by DPO, and what DPO learns when they do not. For each prompt, one set of response scores must explain the preference probabilities for all observed pairs. Such probabilities are Bradley–Terry (BT) compatible (Sun et al., 2025; Pukdee et al., 2026) exactly when their signed log-odds sum to zero around each comparison cycle. Direct and indirect comparisons of the same response pair must therefore imply the same score difference. Uniform tie-breaking can violate this closure condition even though every pairwise probability remains valid. The mismatch arises before optimization, for both free scores and shared neural policies. Figure 1 illustrates the related shift in the population learning target.

For Davidson comparisons with a shared tie strength, our main results concern the exact structure of cycle residuals, their tight extrema, and inverse-link identification. We give the complete compatibility boundary, determine the residual sign on reward-ordered cycles, and show that balanced adjacent reward gaps attain the fixed-span extremum. Applying the inverse link turns the processed logits back into candidate reward gaps. Cycle closure on those gaps then characterizes every tie strength consistent with the observations, including isolated alternative values that remain indistinguishable without a comparison cycle. For DPO, these results explain its best-fitting global scores and distinguish correctable scale changes from comparison-design-dependent distortions. Controlled experiments verify this distinction numerically. Calibration removes the scale distortion; ranking reversals appear precisely in the designs predicted by the theory.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

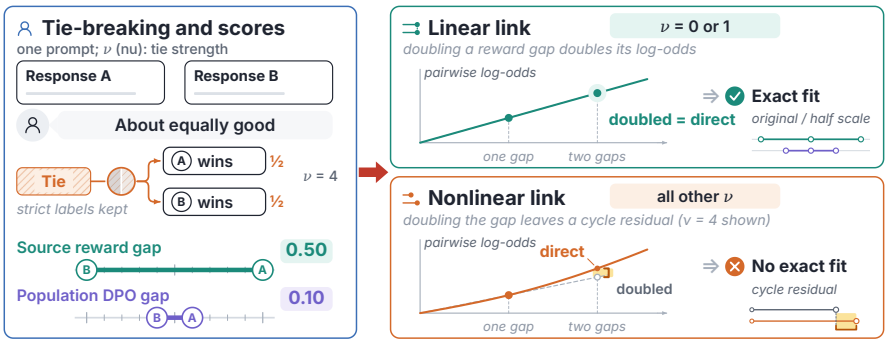


Figure 1: **Uniform tie-breaking can shrink score gaps.** With one score per response, direct log-odds must match the sum obtained through an intermediate response. The highlighted gap marks a nonzero cycle residual. Our analysis derives its exact form and identifies when global scores are possible.

Our contributions are as follows:

(1) Exact cycle structure. We give the complete universal BT-compatibility boundary for uniform tie-breaking and determine the residual sign on strictly reward-ordered cycles (Section 4).

(2) Tight cycle bounds. We establish the fixed-span residual extremum at balanced adjacent gaps and the exact weighted score-projection error on graphs with one independent cycle (Theorem 2).

(3) Inverse-link identification. We characterize all tie strengths consistent with the processed logits, establish generic local identification on cyclic graphs, and give a bounded-interval condition for global uniqueness (Section 4.2).

(4) Consequences for learning and testing. We connect this cycle geometry to DPO projection error and ranking, and derive the Davidson-specific small-gap detection scale and information retained by tie handling (Section 5).

Sections 2 and 3 position and formalize the problem; Sections 4 and 5 develop compatibility, identification, projection, ranking, and finite-sample consequences. Section 6 reports the experiments, and the final section concludes.

2 Related Work

Tie-breaking changes the comparison probabilities that a global-score model must fit. Prior work describes these probabilities, the fitted score class, and the information available from comparison graphs.

Global score models and representability. The Bradley–Terry (BT) model expresses binary log-odds as differences of latent scores (Bradley & Terry, 1952). Pukdee et al. (2026) characterize representability from comparison data; their Theorem 5.2 expresses BT fitting as a comparison-weighted Bernoulli KL projection. We use this projection for the tie-broken law; the channel-specific cycle analysis determines its exact-fit feasibility and the source parameters consistent with the binary probabilities.

Paired-comparison models with ties. Rao–Kupper and Davidson extend binary score models with an explicit tied outcome (Rao & Kupper, 1967; Davidson, 1970). The Elo–Davidson analysis of Szczecinski & Djebbi (2020) already derives the expected-score link that equals our processed probability under $\kappa = 2\nu$ and $d = z\log(10)/\sigma_{\mathrm{Elo}}$, where σ_{Elo} is their rating scale. Their $\kappa = 2$ case gives the half-scale logistic relation. These link identities settle the single pair but leave the graph level open; we establish the complete compatibility boundary and the exact residual on ordered comparison cycles.

Tie-aware preference optimization. DPO embeds a binary BT likelihood in a policy-to-reference log ratio (Rafailov et al., 2023). Tie-rank Oriented DPO and Davidson-based DPO variants instead retain the ternary outcome in their objectives (Guo et al., 2025; Chen et al., 2025). Chen et al. (2025, Section 3.3 and Appendix G) also explain the regularization effect of ties through ideal-policy theory.

We keep binary DPO fixed and study whether the induced pairwise targets can hold simultaneously across a graph, and which source parameters remain recoverable after randomization.

Cycles, ranking, and model tests. HodgeRank supplies the decomposition into global-score and cyclic components (Jiang et al., 2011). The Davidson channel makes that decomposition explicit: the cycle residual in Section 4 measures the inconsistency tie-breaking creates. Sun et al. (2025) study order consistency under sparse comparisons, while Chen et al. (2024) analyze DPO-induced policy rankings. Makur & Singh (2025) give minimax tests of the Bradley–Terry–Luce (BTL) class, with matching bounds on complete graphs. Our exact Davidson residual supplies the channel-specific geometry and the small-gap separation scale used in the learning and testing consequences below.

The missing gap and our motivation. The link and projection results alone do not give the Davidson channel's complete compatibility boundary or resolve its inverse ambiguity. We derive the exact cycle residual and use its sign and curvature to obtain the boundary and tight ordered-cycle bounds. Inverse-link closure then characterizes all source parameters consistent with the processed comparisons, including isolated alternatives. Comparing their predictions identifies additional comparisons that can distinguish otherwise equivalent sources.

3 Tie-Broken Comparisons and Global Scores

Figure 1 shows a target-scale shift on a single pair. With multiple pairs, DPO also requires their logits to be differences of one global score. This requirement is cycle closure; Figure 2 connects it to learning and inference.

3.1 From Davidson Ties to Binary Labels

For one prompt, let $G = (V, E)$ be a connected undirected comparison graph on $n = |V|$ responses and $m = |E|$ edges. Each edge has population weight $w_e > 0$, with $\sum_e w_e = 1$. Orient edge e from u to v and give the incidence matrix $B \in \mathbb{R}^{m\times n}$ row $e_u^\top - e_v^\top$, where e_u is the standard basis vector. The reward vector $r \in \mathbb{R}^n$ induces gaps $d = Br$, so $d_e = r_u - r_v$. Since only differences are observable, impose $\mathbf{1}^\top r = 0$, where $\mathbf{1} \in \mathbb{R}^n$ is the all-ones vector. The columns of $C \in \mathbb{R}^{m\times(m-n+1)}$ span the cycle space $\ker(B^\top)$: signed edge flows with zero net flow at each vertex. This space is zero-dimensional on a tree.

The setting fixes a correctly specified Davidson model and varies only the preprocessing channel. A tie $u \sim v$ is an observed third outcome, distinct from abstention or a missing judgment. For reward gap d, Davidson's original tie coefficient is 2ν, shared across pairs, with $\nu \geq 0$ (Davidson, 1970):

$$\big(p_\nu^+(d), p_\nu^-(d), p_\nu^0(d)\big) = \frac{\big(e^d, 1, 2\nu e^{d/2}\big)}{e^d + 1 + 2\nu e^{d/2}}. \tag{1}$$

The superscripts $+$, $-$, and 0 denote a win for the first endpoint, a loss, and a tie. Equal-reward responses tie with probability $\nu/(1+\nu)$; hence $\nu = 0$ removes ties and $\nu = 1$ gives tie probability $1/2$ at zero gap.

Uniform tie-breaking leaves strict outcomes unchanged and gives either endpoint of a tie the preferred label with probability $1/2$. Let $Y_e \in \{0, 1\}$ indicate whether the first endpoint receives that label. Define $\sigma(t) = (1 + e^{-t})^{-1}$, $\mathrm{logit}(p) = \log\{p/(1-p)\}$, the processed win probability $q_\nu(d) = p_\nu^+(d) + p_\nu^0(d)/2$, and its binary log-odds $h_\nu(d) = \mathrm{logit}\, q_\nu(d)$. For finite d, both labels have positive probability and $h_\nu(-d) = -h_\nu(d)$. The edge-logit vector $\ell = h_\nu(Br) \in \mathbb{R}^m$, with h_ν applied componentwise, has signed cycle sums $C^\top \ell$. Assigning each tie the fractional target $1/2$ gives the same mean q_ν and expected logistic loss.

3.2 When a Global Score Exists

An oriented edge-logit vector $\ell \in \mathbb{R}^m$ is Bradley–Terry (BT) compatible when some score vector $s \in \mathbb{R}^n$ satisfies $\ell = Bs$. Because the graph is connected, s is unique up to an additive constant. A global score exists exactly when every cycle sum vanishes:

$$\ell \in \mathrm{range}(B) \quad \Longleftrightarrow \quad C^\top \ell = 0. \tag{2}$$

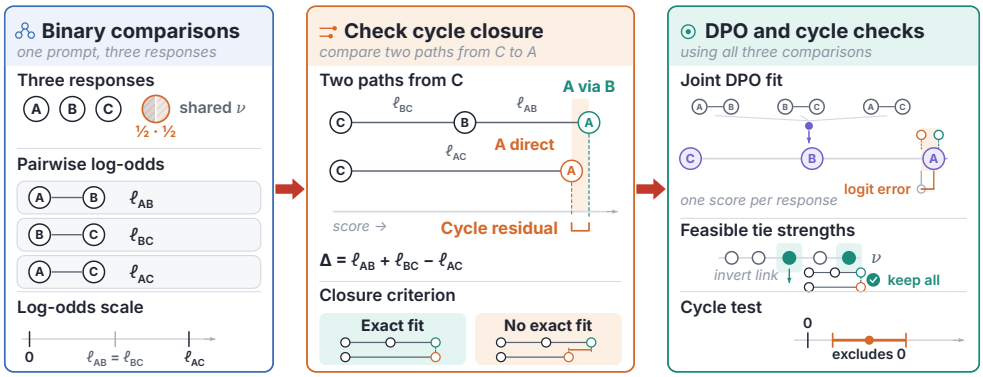


Figure 2: **Checking score consistency.** Path sums between the same responses must agree for an exact fit. DPO fits the pairs jointly; the inverse link checks which tie strengths fit the observations. The confidence interval is schematic.

Every finite edge-logit vector is compatible on a tree. On a graph with cycles, Equation (2) is an observable consistency restriction: identification asks whether gaps recovered from the observed logits obey it, while misspecified DPO measures how far ℓ lies from $\mathrm{range}(B)$.

To connect this condition to DPO, let the reference policy π_{ref} have full support on V, and let $\beta > 0$ be the DPO temperature. Using full sequence probabilities, a full-support policy π induces

$$z_{uv}(\pi) = \beta \left(\log \frac{\pi(u)}{\pi_{\mathrm{ref}}(u)} - \log \frac{\pi(v)}{\pi_{\mathrm{ref}}(v)} \right). \tag{3}$$

This is a difference of node scores, so DPO can fit every processed edge exactly if and only if ℓ satisfies Equation (2).

For the processed edge probabilities $q_e = q_\nu(d_e)$, DPO minimizes the weighted binary cross-entropy

$$\mathcal{R}(s) = \sum_{e \in E} w_e \left\{ \log \left(1 + e^{(Bs)_e} \right) - q_e (Bs)_e \right\}. \tag{4}$$

Definition 1 (Fisher consistency for the oracle policy)**.** A tie-handling rule and loss are Fisher-consistent for a reward vector r when their population risk over full-support policies has the unique normalized minimizer

$$\pi^*(u) = \frac{\pi_{\mathrm{ref}}(u) e^{r_u/\beta}}{\sum_{v \in V} \pi_{\mathrm{ref}}(v) e^{r_v/\beta}}. \tag{5}$$

Here β is fixed for both training and the Kullback–Leibler (KL)-regularized oracle; DPO uses the risk in Equation (4).

More generally, if the compatible logits equal $c(r_u - r_v)$ for $c > 0$, write π_c for the normalized policy proportional to $\pi_{\mathrm{ref}}(u) e^{c r_u/\beta}$. Thus compatibility and Fisher consistency are different questions when tie-breaking changes the global reward scale.

4 Cycle Geometry and Tie-Strength Identification

The closed form and half-scale case of the processed link appear in the Elo–Davidson analysis (Szczecinski & Djebbi, 2020); its exact cycle structure is new. The residual sign gives the complete compatibility boundary, its curvature the tight fixed-span extremum, and inverse-link closure the feasible source parameters.

4.1 Exact Compatibility Boundary

A linear processed logit integrates to global scores on every graph; the following result identifies when this occurs.

Theorem 1 (Compatibility boundary for uniform tie-breaking)**.** *For every finite $d \in \mathbb{R}$ and $\nu \geq 0$,*

$$q_\nu(d) = \frac{e^d + \nu e^{d/2}}{e^d + 1 + 2\nu e^{d/2}}, \qquad \boxed{h_\nu(d) = \log \frac{e^{d/2} + \nu}{e^{-d/2} + \nu}} \tag{6}$$

has the following properties.

(a) A positive constant c_ν satisfies $h_\nu(d) = c_\nu d$ for every real d if and only if $\nu \in \{0, 1\}$. The two maps are d and $d/2$.

(b) On a strictly reward-ordered triangle with $r_i > r_j > r_k$, let $x = r_i - r_j$ and $y = r_j - r_k$. The cycle sum

$$\Delta_\nu(x, y) = h_\nu(x) + h_\nu(y) - h_\nu(x + y) \tag{7}$$

obeys

$$\operatorname{sign} \Delta_\nu(x, y) = \begin{cases} +1, & 0 < \nu < 1, \\ 0, & \nu \in \{0, 1\}, \\ -1, & \nu > 1. \end{cases} \tag{8}$$

Thus every strictly ordered triangle rules out a global score representation when $\nu \notin \{0, 1\}$.

Proof sketch of Theorem 1. Set $A_x = e^{x/2} > 1$ and $A_y = e^{y/2} > 1$. Substitution first gives Equation (6); exponentiating the cycle sum and simplifying then yields

$$\operatorname{sign} \Delta_\nu(x, y) = \operatorname{sign}\{\nu(1 - \nu)(A_x - 1)(A_y - 1)(A_x A_y - 1)\}.$$

The gap factors are positive. A linear link must make every triangle sum vanish, leaving only $\nu = 0, 1$; substitution gives $h_0(d) = d$ and $h_1(d) = d/2$. The full calculation and boundary cases are in Appendix A.

Theorem 2 (Tight ordered-cycle residual)**.** *For $k \geq 2$, let $d_1, \ldots, d_k > 0$, $S = \sum_i d_i$, and $R_{\nu,k} = \sum_i h_\nu(d_i) - h_\nu(S)$. For $\nu \in (0, \infty) \setminus \{1\}$,*

$$\operatorname{sign} R_{\nu,k} = \operatorname{sign}(1 - \nu), \qquad 0 < |R_{\nu,k}| \leq |k h_\nu(S/k) - h_\nu(S)|. \tag{9}$$

The residual is zero at $\nu = 0, 1$; otherwise equality holds exactly for $d_1 = \cdots = d_k$. Suppose the graph has exactly one independent cycle $\mathcal{C}$ of length $k + 1$, with successive reward gaps $d_1, \ldots, d_k$, closing gap $-S$, and the corresponding signed vector $c_\mathcal{C} \in \{-1, 0, 1\}^m$ spanning $\ker(B^\top)$. Then $c_\mathcal{C}^\top \ell = R_{\nu,k}$ and $\min_s \sum_e w_e \{\ell_e - (Bs)_e\}^2 = R_{\nu,k}^2 / \sum_{e \in \mathcal{C}} w_e^{-1}$. For fixed incompatible ν, k, total span, and weights, balanced adjacent gaps uniquely maximize this exact weighted score-projection error.

Proof sketch. For $\nu > 0$ and positive gaps, $\operatorname{sign} h_\nu'' = \operatorname{sign}(\nu^2 - 1)$. Strict convexity or concavity and Jensen's inequality give the sign, bound, and equality case; the one-dimensional cycle-space dual gives the projection identity. See Appendix A.

Corollary 3 (Population DPO targets at the compatible values)**.** *Population DPO after uniform tie-breaking targets the oracle policy π^* at $\nu = 0$ and the rescaled policy $\pi_{1/2}$ at $\nu = 1$. For every other ν, a strict observed triangle prevents an exact edgewise fit by any full-support policy. Changing a single global DPO temperature cannot remove a nonzero cycle sum.*

Proof sketch of Corollary 3. Substituting the linear links from Theorem 1 into Equation (3) gives the stated policies. Every policy induces a score-difference field, so a nonzero cycle rules out exact edgewise fitting at any temperature (Appendix C).

Temperature calibration. At $\nu = 1$, the processed logits are halved. Write β_{obj} for the KL-regularization coefficient and β_{train} for the DPO temperature. Setting $\beta_{\text{train}} = \beta_{\text{obj}}/2$ restores the oracle policy; without this correction, compatibility is not Fisher consistency. A nonzero cycle still prevents exact edgewise fitting. Policy recovery after calibration depends instead on whether the fitted score is a positive multiple of the source reward (Proposition 21).

Beyond the Davidson model. Translation invariance means dependence only on the reward gap; endpoint-reversal symmetry means that swapping endpoints swaps the oriented labels. If the induced binary logit is continuous, universal BT compatibility under these conditions is equivalent to linearity of that logit. Fixed-graph heterogeneous criteria and sign-stability bounds are in Theorems 8 and 14 and Proposition 7.

4.2 Identification from Cycles

A single processed probability cannot distinguish the reward gap from the tie strength. Across a graph, inverse-link closure restricts which tie strengths agree with all observed comparisons. Since h_ν is strictly increasing for finite ν, write $D_\nu := h_\nu^{-1}$ and apply it componentwise. The identified set is

$$\mathcal{I}_G(\ell) = \{\nu \geq 0 : C^\top D_\nu(\ell) = 0\}. \tag{10}$$

For each admissible ν, the zero-sum solution of $Br_\nu = D_\nu(\ell)$ is the unique compatible reward. Identification is characterized on cyclic graphs with at least three distinct reward levels. The boundary cases are trees, two-level reward vectors, and the observationally equivalent endpoints $\nu = 0, 1$ when reward scale is unknown; see Theorem 10.

Local identification asks whether changing ν can be mimicked by a reward change. At the true (ν_0, r), let $j = (\partial_\nu D_\nu(\ell_e)|_{\nu_0})_{e\in E}$. The profiled Fisher information I_{eff} measures the variation in this direction left after allowing reward perturbations. Its weighted projection formula is given in Theorem 11.

Proposition 4 (Local identification from cycle variation)**.** *If $C^\top j \neq 0$, then (ν_0, r) is locally identified up to the additive reward anchor. The information is zero on trees, whereas on a cyclic graph the condition holds for almost every anchored reward vector at fixed finite ν_0.*

Proof sketch of Proposition 4. Inverse-linking makes reward perturbations span $\mathrm{range}(B)$ and the ν-direction equal j; hence full rank is equivalent to $C^\top j \neq 0$, and weighted projection gives I_{eff}. The inverse function theorem yields local identification. An explicit reward configuration on a cycle gives $C^\top j \neq 0$; analyticity then makes the rank-failure set measure zero.

On a strict triangle generated with $0 < \nu_0 < 1$, the inverse-link cycle profile has the same nonzero sign at 0 and 1. A root with nonzero derivative therefore requires another root, despite local identification. The population example in Appendix F.5 resolves its two branches with an added cycle; Appendix B.3 retains all feasible values under sampling uncertainty. A fixed-sign profile derivative gives interval-wise uniqueness (Equation (19)). The expansion in Equation (20) identifies when information decays as t^6, and when $\nu = 1/2$ causes the faster t^{10} decay.

5 Consequences of Incompatible Comparisons

The cycle results determine the learning and statistical consequences below. Population DPO minimizes weighted binary cross-entropy, whose target is the comparison-weighted KL projection of Pukdee et al. (2026, Theorem 5.2). The Davidson residual then determines exact-fit feasibility and approximation error. The comparison design controls the fitted score direction, ranking, and calibrated policy regret.

5.1 Best-Fitting Global Scores

Let $q = \sigma(\ell) \in (0, 1)^m$ collect the processed edge probabilities and let $W = \mathrm{diag}(w_e)$. Parameterize a full-support policy by its unique centered DPO score $s \in \mathbf{1}^\perp := \{s : \mathbf{1}^\top s = 0\}$, so that $\pi_s(u) \propto \pi_{\text{ref}}(u)e^{s_u/\beta}$ and its edge logits are Bs. DPO minimizes the weighted binary cross-entropy $\mathcal{R}(s)$ between q and $\sigma(Bs)$, defined in Equation (4).

Proposition 5 (Population DPO is a graph-weighted projection)**.** *The risk has a unique anchored minimizer $s^\dagger$. Its policy $\pi^\dagger(u) \propto \pi_{\text{ref}}(u)e^{s^\dagger_u/\beta}$ satisfies*

$$\boxed{B^\top W\{\sigma(Bs^\dagger) - q\} = 0}. \tag{11}$$

It fits every edge exactly if and only if $C^\top \ell = 0$.

Proof sketch of Proposition 5. On $\mathbf{1}^\perp$, the positive-definite Hessian and growth at infinity give a unique minimizer; differentiation yields Equation (11). Exact fitting is equivalent to $\ell \in \mathrm{range}(B)$, hence cycle closure (Appendix C). Compatible data zero every edge residual; otherwise balance holds only after graph-weighted aggregation.

5.2 Fit Error and Ranking

To separate irreducible mismatch from the fitted score, measure the distance from the induced logits to the global-score subspace. For an edge vector u_E, write $\|u_E\|_W^2 = u_E^\top W u_E$. The graph Hodge projection separates an edge field into score-difference and cycle components (Jiang et al., 2011). The squared weighted Hodge projection error is

$$Q_W(\ell) = \min_{s \in \mathbb{R}^n} \|\ell - Bs\|_W^2.$$

Its equivalent cycle formulation is derived in Equation (25). Thus Q_W isolates the cycle mismatch and bounds the associated logistic and policy/log-ratio errors; source-reward regret additionally contains the temperature-scale term characterized below, which remains positive at $\nu = 1$ with nonconstant rewards. Promptwise errors cannot cancel under shared DPO scores (Theorem 16 and Corollary 17).

For $\mathcal{J}_r(\pi) = \sum_u \pi(u) r_u - \beta D_{\mathrm{KL}}(\pi \| \pi_{\mathrm{ref}})$, the source-reward regret is $\mathcal{J}_r(\pi^*) - \mathcal{J}_r(\pi^\dagger) = \beta D_{\mathrm{KL}}(\pi^\dagger \| \pi^*)$. At a common temperature it is positive for nonconstant rewards and positive tie mass (Theorem 18). At fixed objective temperature, regret after optimizing the training temperature vanishes exactly when $s^\dagger = ar$ for some $a > 0$; this condition is necessary and sufficient for source-policy recovery and can hold despite a nonzero cycle (Proposition 21).

Ranking is design-dependent: complete equal-weight comparison designs preserve every strict reward order, and the sparse four-response path characterizes the reversal that can occur for each finite $\nu > 0$, $\nu \neq 1$ even when $Q_W = 0$ (Theorem 19 and Proposition 20). For bounded logits, Theorem 22 bounds score changes under comparison perturbations and preserves baseline gaps larger than its pairwise bound.

5.3 Information Preserved by Tie Handling

Tie handling changes both the population target and retained information (Table 4 in Appendix D). Conditioning on a strict Davidson outcome cancels the tie term and preserves Bradley–Terry odds e^d, but discards samples; a correct ternary likelihood also recovers the original reward. In contrast, a fractional half-label and uniform random tie-breaking share mean $q_\nu(d)$ and therefore the projection in Equation (11).

Encode a win, loss, or tie as $X = 1, 0, 1/2$. Let $\widetilde{X} = X$ on strict outcomes and draw $\widetilde{X} \mid \{X = 1/2\} \sim \mathrm{Bernoulli}(1/2)$ on a tie. This preserves expected logistic loss but adds $p_\nu^0(d)/4$ to both label and stochastic-gradient variance. With ν known and d the scalar fixed-edge parameter, the randomized label and the ternary outcome differ in Fisher information only through the coin-variance term $p_\nu^0(d)/4$, and Appendix D gives the exact accounting.

5.4 Detecting Incompatibility from Finite Data

Makur & Singh (2025) establish general minimax BTL testing bounds. Here we fix an ordered triangle with gaps $x, y > 0$ and $\nu \in (0, \infty) \setminus \{1\}$. Scaling the gaps by $t \downarrow 0$ gives a nonzero cubic residual (Equation (14)), which determines the local comparison budget.

Finite-sample cycle certificates. An interval excluding zero certifies BT incompatibility; on an ordered Davidson triangle, the sign distinguishes $0 < \nu < 1$ from $\nu > 1$. Edgewise coverage permits data-selected cycles.

The sixth-power boundary. With n_{rep} independent Bernoulli observations per edge, each with mean q_e, Cycle-Wald (Section 6) attains the local boundary below.

Corollary 6 (Sixth-power small-gap boundary)**.** *Along this path, the paired-null type-I and type-II error limits of every test sum to at least one if $n_{\mathrm{rep}} t^6 \to 0$. At any fixed significance level in $(0, 1)$, Cycle-Wald has power tending to one if $n_{\mathrm{rep}} t^6 \to \infty$. Thus t^{-6} is the sharp local boundary.*

Proof sketch. The cubic expansion and Theorem 25 give the error-sum lower bound. When $n_{\mathrm{rep}} t^6 \to \infty$, the Cycle-Wald noncentrality diverges (Theorem 24). Full proofs, constants, allocation, and coverage are in Appendix E.

6 EXPERIMENTS

Controlled graphs test the residual sign, detection power, and tie-handling estimates; trained policies test whether the same mismatch persists after optimization. Dense cyclic designs isolate the cycle residual, while instruction-tuning logs reflect the reward-scale and coin-variance channel (Section 5.3). Temperature calibration removes the scale component of regret, and promptwise scores separate that component from parameter sharing.

6.1 EXPERIMENTAL SETUP

The finite-sample estimators are Davidson maximum likelihood (Dav-MLE), Bradley–Terry after tie deletion (BT-Drop), half-label logistic regression (Logit-Half), and fair-coin logistic regression (Logit-Coin). Exact-Cycle uses simultaneous exact-binomial intervals; Cycle-Wald standardizes a cycle sum. Equal allocation (EA) is uniform, while oracle cycle-information allocation (OCI) minimizes cycle variance using generating probabilities. The reported model comparisons use the binary baseline DPO and the two tie-aware objectives Tie-rank Oriented DPO (TODO) (Guo et al., 2025) and Davidson DPO (DPO-D) (Chen et al., 2025); together they separate tie handling from other design choices, since DPO tie-breaks while TODO and DPO-D retain the ternary outcome. We fit the four estimators to 4,320 paired datasets and train DPO, TODO, and DPO-D on the two held-out splits; grids and optimization settings are in Appendix F.

6.2 CONTROLLED GRAPH EXPERIMENTS

At $\nu = 0, 1$, the largest computed $|\Delta_\nu|$ is 4.44×10^{-16}. All 295,200 grid evaluations have the predicted sign, and fitted slopes 2.9994–3.0000 match the cubic signal in Corollary 6. Complete graphs have $0/630$ reversals; all $15/15$ constructed paths reverse the unobserved endpoint order (Theorem 19 and Proposition 20). At $t = 1$ with 2^{20} comparisons per edge, Cycle-Wald reaches 99.50% power and Exact-Cycle reaches 71.78%; the gap is the cost of Exact-Cycle's simultaneous control over all cycles. At one million comparisons, OCI reaches 71.11% versus EA's 62.89%, after matching EA at small budgets, as the oracle allocation predicts (Table 6 and Figure 4).

Table 1 separates target shift from sampling error. Logit-Half and Logit-Coin fit the processed binary distribution, so their source-reward error is larger by construction; around their shared population target both are accurate, and the small difference between them matches the added coin variance in Section 5.3. Logit-Half therefore removes coin noise while retaining the same population projection.

Across these controls, one closed-form residual determines the compatibility boundary, the cubic detection signal, the fitted targets, and the ranking reversals.

Table 1: **Tie handling.** Means over 4,320 paired datasets. Dav-MLE uses the known generating ν; method-target RMSE uses each estimator's own population target. Blue and underlined amber mark the best and second-best comparable values.

Method	Source-reward RMSE ↓	Method-target RMSE	Order acc. (%) ↑
Dav-MLE	**0.0933**	0.0933	**95.58**
BT-Drop	**0.0984**	0.0984	95.52
Logit-Half	0.3119	0.0481	**95.63**
Logit-Coin	0.3146	0.0652	93.44

Temperature calibration. Individual calibration reduces source-policy regret on complete graphs to a mean of 4.44×10^{-6} across all nonlinear configurations (Tables 2 and 7), matching the scale-recovery prediction of Proposition 21. A single temperature correction thus recovers the source policy on complete designs, while cycle, Erdős–Rényi, and path graphs retain the design-dependent regret ordering in Table 7; the remaining error is direction distortion, not scale error.

6.3 SHARED-POLICY EXPERIMENTS

Shared fits one reward scale across prompts and Separate one per prompt, while Free fits independent centered scores; in this controlled design Separate and Free coincide, isolating the logistic fitting cost of sharing a scale (Table 3).

Table 2: **Regret after temperature calibration.** Complete six-response graphs with a uniform reference and $\beta_{\text{obj}} = 1$; three-seed means. Common calibration uses one temperature per ν; individual calibration uses one per graph–reward instance. Ray distance is $\min_{a \geq 0} \|s^\dagger - ar\|_2 / \|s^\dagger\|_2$.

ν	Fixed regret $\times 10^{-3}$	Common regret $\times 10^{-6}$	Individual regret $\times 10^{-6}$	Ray distance $\times 10^{-3}$
0	0.0000	0.0000	0.0000	0.0000
0.1	1.9894	1.8055	0.2422	0.8354
0.25	9.4353	4.5782	0.6006	1.3329
0.5	25.7583	3.9253	0.5077	1.2344
1	57.3947	0.0000	0.0000	0.0000
2	102.5369	18.1631	2.3599	2.6249
4	150.0711	95.7785	12.5169	5.9543

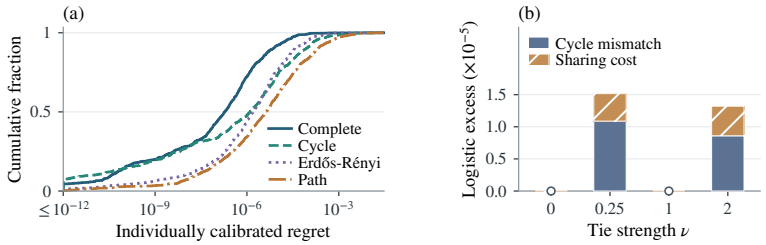


Figure 3: (a) Distribution of regret after individual calibration across the nonlinear designs; values at or below the first tick are grouped there. (b) Logistic excess in the shared-policy example: free-score excess (cycle mismatch) and the additional cost of sharing a scale across prompts.

At $\nu = 0.25$, sharing adds about 39% to the free-score logistic excess (Figure 3(b)); even with independent promptwise scores, the nonzero cycle sums leave positive excess (Corollary 17). At $\nu = 0, 1$, both Free and Shared have zero excess, matching the compatibility boundary in Theorem 1.

Table 3: **Parameter sharing and cycle mismatch.** Population means on the same three prompts, at fixed temperature one. Excess is Bernoulli KL; regret is source-policy KL. Separate and Free coincide. Common calibration uses one temperature; promptwise calibration gives zero regret throughout.

ν	Scores	Excess $\times 10^{-5}$	Logit RMSE $\times 10^{-2}$	Fixed regret $\times 10^{-2}$	Common regret $\times 10^{-6}$
0	All	0.0000	0.0000	0.0000	0.0000
0.25	Shared	1.5158	1.6261	1.0957	0.0000
0.25	Separate/Free	1.0884	1.2922	1.0988	8.6811
1	All	0.0000	0.0000	7.1229	0.0000
2	Shared	1.3204	1.1090	13.1887	0.0000
2	Separate/Free	0.8599	0.8788	13.1863	34.1471

On the matched Llama setting, TODO's strict-pair accuracy on held-out pairs exceeds DPO's by 3.35 points $[0.13, 6.59]$, so tie-aware training improves the ternary fit without losing strict-pair accuracy; the Qwen transfer comparison and remaining paired intervals are reported in Table 10.

7 Conclusion

We establish the exact universal compatibility boundary for uniform tie-breaking in the Davidson model from the sign and magnitude of its cycle residuals. Balanced adjacent gaps attain the tight fixed-span bound on residual magnitude. When the ordered cycle is the graph's only independent cycle, its residual determines the exact weighted squared score-projection error. The inverse link reduces parameter recovery to a cycle-closure problem that characterizes feasible tie strengths and local identification. These results explain DPO's population projection and separate fitting mismatch from changes in reward scale. Temperature calibration recovers the source-reward policy exactly when fitted scores are a positive rescaling of the rewards, as the experiments confirm.

AI USE STATEMENT

Generative AI tools assisted language editing and formatting.

REFERENCES

Siavash Ameli, Siyuan Zhuang, Ion Stoica, and Michael W. Mahoney. A statistical framework for ranking LLM-based chatbots. In *The Thirteenth International Conference on Learning Representations*, pp. 9505–9540, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/1a4cd257678d986ba545b5d1aa9b5923-Abstract-Conference.html.

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: 10.2307/2334029.

Daniele Calandriello, Zhaohan Daniel Guo, Remi Munos, Mark Rowland, Yunhao Tang, Bernardo Avila Pires, Pierre Harvey Richemond, Charline Le Lan, Michal Valko, Tianqi Liu, Rishabh Joshi, Zeyu Zheng, and Bilal Piot. Human alignment of large language models through online preference optimisation. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 5409–5435. PMLR, 2024. URL https://proceedings.mlr.press/v235/calandriello24a.html.

Angelica Chen, Sadhika Malladi, Lily H. Zhang, Xinyi Chen, Qiuyi Zhang, Rajesh Ranganath, and Kyunghyun Cho. Preference learning algorithms do not learn preference rankings, 2024. URL https://arxiv.org/abs/2405.19534.

Jinghong Chen, Guangyu Yang, Weizhe Lin, Jingbiao Mei, Chenxu Lyu, and Bill Byrne. On extending direct preference optimization to accommodate ties. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/4dbf3488a8da908a8177d525ede1a43a-Abstract-Conference.html.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios N. Angelopoulos, Tianle Li, Dacheng Li, Banghua Zhu, Hao Zhang, Michael I. 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*, 2024. URL https://arxiv.org/abs/2403.04132.

Ganqu Cui, Lifan Yuan, Ning Ding, Guanming Yao, Bingxiang He, Wei Zhu, Yuan Ni, Guotong Xie, Ruobing Xie, Yankai Lin, Zhiyuan Liu, and Maosong Sun. UltraFeedback: Boosting language models with scaled AI feedback. *arXiv preprint arXiv:2310.01377*, 2023. URL https://arxiv.org/abs/2310.01377.

Roger R. Davidson. On extending the Bradley–Terry model to accommodate ties in paired comparison experiments. *Journal of the American Statistical Association*, 65(329):317–328, 1970. doi: 10.1080/01621459.1970.10481082.

Kawin Ethayarajh, Winnie Xu, Niklas Muennighoff, Dan Jurafsky, and Douwe Kiela. Model alignment as prospect theoretic optimization. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 12634–12651. PMLR, 2024. URL https://proceedings.mlr.press/v235/ethayarajh24a.html.

Yassir Fathullah and Mark Gales. Generalised probabilistic modelling and improved uncertainty estimation in comparative LLM-as-a-judge. In *Proceedings of the Forty-First Conference on Uncertainty in Artificial Intelligence*, volume 286 of *Proceedings of Machine Learning Research*, pp. 1266–1288, 2025. URL https://proceedings.mlr.press/v286/fathullah25a.html.

Mohammad Gheshlaghi Azar, Zhaohan Daniel Guo, Bilal Piot, Remi Munos, Mark Rowland, Michal Valko, and Daniele Calandriello. A general theoretical paradigm to understand learning from human preferences. In *Proceedings of the 27th International Conference on Artificial Intelligence and*

Statistics, volume 238 of *Proceedings of Machine Learning Research*, pp. 4447–4455. PMLR, 2024. URL https://proceedings.mlr.press/v238/gheshlaghi-azar24a.html.

Yuxiang Guo, Lu Yin, Bo Jiang, and Jiaqi Zhang. TODO: Enhancing LLM alignment with ternary preferences. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/fef5607a9b826f1845533f923e308435-Abstract-Conference.html.

Xiaoye Jiang, Lek-Heng Lim, Yuan Yao, and Yinyu Ye. Statistical ranking and combinatorial Hodge theory. *Mathematical Programming*, 127(1):203–244, 2011. doi: 10.1007/s10107-010-0419-x.

Shang Liu, Yu Pan, Guanting Chen, and Xiaocheng Li. Reward modeling with ordinal feedback: Wisdom of the crowd. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 39190–39218. PMLR, 2025. URL https://proceedings.mlr.press/v267/liu25az.html.

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.

Yu Meng, Mengzhou Xia, and Danqi Chen. SimPO: Simple preference optimization with a reference-free reward. In *Advances in Neural Information Processing Systems*, volume 37, pp. 124198–124235, 2024. doi: 10.52202/079017-3946. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/e099c1c9699814af0be873a175361713-Abstract-Conference.html.

Remi Munos, Michal Valko, Daniele Calandriello, Mohammad Gheshlaghi Azar, Mark Rowland, Zhaohan Daniel Guo, Yunhao Tang, Matthieu Geist, Thomas Mesnard, Côme Fiegel, Andrea Michi, Marco Selvi, Sertan Girgin, Nikola Momchev, Olivier Bachem, Daniel J Mankowitz, Doina Precup, and Bilal Piot. Nash learning from human feedback. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 36743–36768. PMLR, 2024. URL https://proceedings.mlr.press/v235/munos24a.html.

Rattana Pukdee, Maria-Florina Balcan, and Pradeep Ravikumar. What does preference learning recover from pairwise comparison data? In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. URL https://icml.cc/virtual/2026/poster/63498.

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, 2023. doi: 10.52202/075280-2338. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/a85b405ed65c6477a4fe8302b5e06ce7-Abstract-Conference.html.

P. V. Rao and Lawrence L. Kupper. Ties in paired-comparison experiments: A generalization of the Bradley–Terry model. *Journal of the American Statistical Association*, 62(317):194–204, 1967. doi: 10.1080/01621459.1967.10482901.

Sayak Ray Chowdhury, Anush Kini, and Nagarajan Natarajan. Provably robust DPO: Aligning language models with noisy feedback. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 42258–42274. PMLR, 2024. URL https://proceedings.mlr.press/v235/ray-chowdhury24a.html.

Hao Sun, Yunyi Shen, and Jean-Francois Ton. Rethinking reward modeling in preference-based large language model alignment. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 46885–46922, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/7423902b5534e2b267438c85444a54b1-Paper-Conference.pdf.

Leszek Szczecinski and Aymen Djebbi. Understanding draws in Elo rating algorithm. *Journal of Quantitative Analysis in Sports*, 16(3):211–220, 2020. doi: 10.1515/jqas-2019-0102. URL https://doi.org/10.1515/jqas-2019-0102.

Yunhao Tang, Zhaohan Daniel Guo, Zeyu Zheng, Daniele Calandriello, Remi Munos, Mark Rowland, Pierre Harvey Richemond, Michal Valko, Bernardo Avila Pires, and Bilal Piot. Generalized preference optimization: A unified approach to offline alignment. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 47725–47742. PMLR, 2024. URL https://proceedings.mlr.press/v235/tang24b.html.

Zihao Wang, Chirag Nagpal, Jonathan Berant, Jacob Eisenstein, Alexander Nicholas D'Amour, Sanmi Koyejo, and Victor Veitch. Transforming and combining rewards for aligning large language models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 51161–51176. PMLR, 2024. URL https://proceedings.mlr.press/v235/wang24ay.html.

Yi Xu, Laura Ruis, Tim Rocktäschel, and Robert Kirk. Investigating non-transitivity in LLM-as-a-judge. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 69583–69612. PMLR, 2025. URL https://proceedings.mlr.press/v267/xu25w.html.

An Yang et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. URL https://arxiv.org/abs/2505.09388.

A Compatibility Proofs

Proof of Theorem 1. Uniform tie-breaking gives $q_\nu = p^+_\nu + p^0_\nu/2$. Its odds are $(e^{d/2}+\nu)/(e^{-d/2}+\nu)$, proving Equation (6) and oddness of h_ν. For $A = e^{x/2}$, $B = e^{y/2}$, direct subtraction yields

$$e^{\Delta_\nu(x,y)} - 1 = \frac{\nu(1-\nu)(A-1)(B-1)(AB-1)}{(1+\nu A)(1+\nu B)(AB+\nu)}. \tag{12}$$

The denominator is positive. For $x, y > 0$, all three gap factors are positive, giving the claimed sign. A linear link has zero triangle sum, so this identity allows only $\nu = 0, 1$; substitution gives $h_0(d) = d$ and $h_1(d) = d/2$. Conversely, these linear links integrate on every graph. A strict triangle with nonzero sum cannot be a difference field, by Proposition 7 below.

Proof of Theorem 2. Differentiation gives

$$h''_\nu(d) = \frac{\nu(\nu^2-1)\sinh(d/2)}{2\{1+\nu^2+2\nu\cosh(d/2)\}^2}. \tag{13}$$

For $0 < \nu < 1$, strict concavity on positive gaps and $h_\nu(0) = 0$ give $h_\nu(d_i) > (d_i/S)h_\nu(S)$. Summing proves positivity. Jensen's inequality gives $\sum_i h_\nu(d_i) \le k h_\nu(S/k)$, with equality exactly when all gaps coincide. For $\nu > 1$, convexity reverses both inequalities; at $0, 1$, linearity gives zero residual.

For a graph with exactly one independent cycle, write $c = c_\mathcal{C}$ and $R = c^\top \ell$. Every residual $u = \ell - Bs$ satisfies $c^\top u = R$. Weighted Cauchy–Schwarz gives $u^\top W u \ge R^2/(c^\top W^{-1} c)$. Equality is attained at $u = W^{-1}cR/(c^\top W^{-1}c)$: $\ell - u$ has zero cycle sum and therefore lies in $\mathrm{range}(B)$. Since $c^\top W^{-1} c = \sum_{e\in\mathcal{C}} w_e^{-1}$, this proves the exact projection identity and its balanced-gap equality case.

Small reward gaps. Oddness and Taylor expansion of Equation (6) give $h_\nu(d) = d/(1+\nu) + \nu(\nu-1)d^3/\{24(1+\nu)^3\} + O(d^5)$. The linear terms cancel in the cycle sum, and $x^3+y^3-(x+y)^3 = -3xy(x+y)$, so

$$\Delta_\nu(tx, ty) = \frac{\nu(1-\nu)}{8(1+\nu)^3} xy(x+y)t^3 + O(t^5). \tag{14}$$

Degenerate cases. A zero adjacent gap makes Equation (12) vanish. On a tree, any edge logits integrate along unique paths, so compatibility holds for that instance at every ν. On general longer cycles, cancellation is possible unless the gaps follow the strict order assumed in Theorem 2. Reversing a cycle reverses its residual sign, not the incompatibility. Finite rewards and finite ν keep both binary probabilities positive; full policy and reference support keep the DPO log ratios finite.

Proposition 7 (Fixed-graph compatibility)**.** *Let B be the oriented incidence matrix of a connected graph with m edges, and let C span $\ker(B^\top)$. For any edge probabilities in $(0,1)$, including heterogeneous comparison laws,*

$$\ell \in \operatorname{range}(B) \quad \Longleftrightarrow \quad C^\top \ell = 0, \qquad \ell_e = \operatorname{logit} q_e. \tag{15}$$

There are $m - |V| + 1$ independent cycle equations.

Proof. The identity $\operatorname{range}(B) = \ker(B^\top)^\perp$ proves the equivalence. Connectedness gives $\operatorname{rank} B = |V| - 1$, hence the dimension. The argument uses only marginal edge probabilities and applies separately to each prompt.

Pairwise preference models also support game-theoretic alignment without requiring a scalar reward representation (Munos et al., 2024; Calandriello et al., 2024). Other alignment objectives change the preference loss or the transformation of rewards (Gheshlaghi Azar et al., 2024; Ethayarajh et al., 2024; Wang et al., 2024). The following result characterizes the observation channels that remain compatible with global scores.

Theorem 8 (Scalar tie-resolution channels)**.** *Let continuous ternary probabilities satisfy $p^+(-d) = p^-(d)$, $p^0(-d) = p^0(d)$, and let a tie be assigned to the first endpoint with probability $a(d)$, where $a(-d) = 1 - a(d)$. Suppose $q_a = p^+ + ap^0 \in (0,1)$ has a continuous logit. Universal BT compatibility is equivalent to $q_a(d) = \sigma(cd)$ for a constant c; strict increase requires $c > 0$. A prescribed c is feasible precisely when $p^+(d) \le \sigma(cd) \le p^+(d) + p^0(d)$ for every d.*

Proof. Reversal gives $h_a(-d) = -h_a(d)$. The triangle with rewards $(x + y, y, 0)$ then forces $h_a(x) + h_a(y) = h_a(x + y)$. Additivity gives $h_a(d) = dh_a(1)$ first on rationals and then, by continuity, on reals. Conversely, $h_a(d) = cd$ integrates as the score cr. When $p^0(d) > 0$, the unique channel is $a(d) = \{\sigma(cd) - p^+(d)\}/p^0(d)$; the stated inequalities are exactly $0 \le a \le 1$, and reversal gives $a(-d) = 1 - a(d)$. Where $p^0 = 0$, they require $p^+ = \sigma(cd)$; set the unused channel to $1/2$. This completes the construction.

B Identification and Perturbations

The inverse link separates a common tie strength from the reward gaps. Throughout this section, the comparison graph is connected, the logits are finite, and rewards are anchored by $r \perp \mathbf{1}$.

B.1 The identified set

Lemma 9. *For every finite $\nu \ge 0$, h_ν is an odd, strictly increasing bijection with inverse*

$$D_\nu(z) = h_\nu^{-1}(z) = z + 2\operatorname{asinh}\{\nu \sinh(z/2)\}. \tag{16}$$

Proof. The derivative in Equation (17) is positive. For $\nu > 0$, $h_\nu(d) = d/2 - \log \nu + o(1)$ as $d \to \infty$; oddness gives the other endpoint, and $h_0(d) = d$. Writing $t = e^{d/2}$, the odds equation becomes $t^2 + \nu(1 - e^z)t - e^z = 0$. Its roots have negative product, so exactly one is positive; twice its logarithm gives the displayed inverse.

Theorem 10. *Under a common Davidson source, the parameters generating ℓ are exactly*

$$\nu \in \mathcal{I}_G(\ell) = \{\nu \ge 0 : C^\top D_\nu(\ell) = 0\}, \qquad r = (B^\top B)^\dagger B^\top D_\nu(\ell).$$

The set is all of $[0, \infty)$ on a tree or when the source rewards take at most two values. If $\ell \in \operatorname{range}(B)$, it contains both 0 and 1. Without a restriction on the ternary source law, each edge tie probability ranges over $[0, 2\min(q_e, 1 - q_e)]$ while preserving the binary marginal.

Proof. For each candidate ν, inverse-linking fixes the edge gaps uniquely. They integrate to a reward vector precisely when they are orthogonal to $\ker(B^\top)$. The displayed pseudoinverse gives that vector; connectedness makes the anchored solution unique. On a tree, $\operatorname{range}(B) = \mathbb{R}^{|E|}$. For two reward levels with positive gap d_0, replace that gap by $D_\nu(h_{\nu_0}(d_0))$; oddness gives the required inverse-linked gaps on every edge. The identities $D_0(z) = z$ and $D_1(z) = 2z$ prove the endpoint alias. Finally, for any tie mass in the stated interval, the ternary probabilities $(q_e - t_e/2, t_e, 1 - q_e - t_e/2)$ are nonnegative, sum to one, and induce q_e.

B.2 LOCAL IDENTIFICATION

At a true (ν_0, r), put $d = Br$. Implicit differentiation gives

$$\alpha_\nu(d) = \partial_d h_\nu(d) = \frac{1 + \nu\cosh(d/2)}{1 + \nu^2 + 2\nu\cosh(d/2)}, \tag{17}$$

$$j_\nu(d) = \partial_\mu D_\mu(z)|_{\mu=\nu, z=h_\nu(d)} = \frac{2\sinh(d/2)}{1 + \nu\cosh(d/2)}. \tag{18}$$

Write $j = j_{\nu_0}(Br)$ and $q_e = q_{\nu_0}(d_e)$. Let n_e be the edge count and $N = \sum_e n_e$, with $n_e/N \to \lambda_e > 0$ on every edge. Define

$$W_* = \mathrm{diag}\{\lambda_e q_e(1 - q_e)\alpha_{\nu_0}(d_e)^2\}, \qquad \|v\|^2_{W_*} = v^\top W_* v.$$

The profiled Fisher information removes from j the edge changes explained by reward perturbations.

Theorem 11. *The joint logit Jacobian has full column rank if and only if $C^\top j \neq 0$, equivalently $I_{\text{eff}} = \min_{a \in \mathbf{1}^\perp} \|j - Ba\|^2_{W_*} > 0$. This implies local identification, one-sided at $\nu_0 = 0$. On every cyclic graph the condition holds for almost every anchored reward vector at fixed finite ν_0. When it holds, ν_0 is isolated in $\mathcal{I}_G(\ell)$, and the identified set has finitely many elements on every bounded interval.*

Proof. Let U have orthonormal columns spanning $H = \mathbf{1}^\perp$, write $B_H = BU$, and set $A = \mathrm{diag}(\alpha_{\nu_0}(d_e))$. The joint Jacobian is $\mathcal{J} = A[-j, B_H]$. Since A is invertible and B_H has full column rank, $\mathcal{J}$ has full rank exactly when $j \notin \mathrm{range}(B)$, or $C^\top j \neq 0$. A nonsingular square row submatrix and the inverse function theorem give local injectivity. At zero, the finite-gap logit map extends smoothly to slightly negative ν; restriction back to $\nu \geq 0$ gives the one-sided conclusion. With $\Gamma = \mathrm{diag}(\lambda_e q_e(1 - q_e))$, profiling the reward coordinates in $\mathcal{J}^\top \Gamma \mathcal{J}$ gives the Schur complement

$$\min_{\theta \in \mathbb{R}^{n-1}} \|Aj - AB_H\theta\|^2_\Gamma = \min_{a \in \mathbf{1}^\perp} \|j - Ba\|^2_{A\Gamma A} = I_{\text{eff}}.$$

For genericity, fix a simple cycle with circulation c. The analytic function $r \mapsto c^\top j_{\nu_0}(Br)$ is not identically zero. To see this, $j_0(d) = 2\sinh(d/2)$ is nonlinear; for positive ν_0, j_{ν_0} is bounded, odd, and nonconstant, hence also nonlinear. Continuity then gives positive x, y with $j_{\nu_0}(x) + j_{\nu_0}(y) \neq j_{\nu_0}(x + y)$. Assign consecutive cycle vertices rewards $x + y, y, 0, \ldots, 0$, assign the other vertices arbitrarily, and subtract the overall mean. The resulting cycle sum is exactly this nonzero difference. The zero set of a nontrivial real-analytic function has measure zero. For completeness, at each zero some finite directional derivative is nonzero; on a small box it remains nonzero, so Rolle's theorem bounds the zeros on parallel lines. Fubini's theorem and a countable cover give the assertion.

Finally, some coordinate of $C^\top D_\nu(\ell)$ has nonzero derivative at ν_0. It is a nonconstant analytic function of ν, so its zeros are isolated. The inverse formula is analytic on a neighborhood of each finite compact interval, making the number of zeros there finite. The identified set is a subset of these zeros.

For a circulation c, a compact interval $\mathcal{N}$ containing an admissible ν_0 has a unique admissible value if, throughout it,

$$c^\top j_\nu(D_\nu(\ell)) \geq \kappa > 0 \quad \text{or} \quad c^\top j_\nu(D_\nu(\ell)) \leq -\kappa < 0. \tag{19}$$

These are the two possible strict monotonicity conditions for $c^\top D_\nu(\ell)$. For a strict triangle generated at $\nu_0 \in (0, 1)$, the reward-ordered profile $F(\nu) = c^\top D_\nu(\ell)$ has $F(0) = \Delta > 0$ and $F(1) = 2\Delta > 0$. A simple interior root changes sign and therefore forces another root in $(0, 1)$. The two branches and additional comparisons that separate them are given in Appendix F.5.

Information at small gaps. For fixed $d = Br$, ν, and positive sampling proportions λ_e, scale rewards to tr as $t \downarrow 0$. Write $\Lambda = \mathrm{diag}(\lambda_e)$ and $Q_\Lambda(v) = \min_a \|v - Ba\|^2_\Lambda$, with powers taken componentwise. Then

$$I_{\text{eff}}(t) = \frac{(1 - 2\nu)^2}{2304(1 + \nu)^6} t^6 Q_\Lambda(d^{\odot 3}) + O(t^8). \tag{20}$$

Indeed, Equation (18) gives $j_\nu(td) = td/(1+\nu) + (1-2\nu)t^3 d^{\odot 3}/\{24(1+\nu)^2\} + O(t^5)$, and $W_*(t) = \Lambda/\{4(1+\nu)^2\} + O(t^2)$. The linear term lies in $\mathrm{range}(B)$. Projecting the remainder and using smooth dependence on positive weights proves the formula. Thus $I_{\mathrm{eff}} = \Theta(t^6)$ when $\nu \neq 1/2$ and $C^\top d^{\odot 3} \neq 0$. At $\nu = 1/2$, $j_{1/2}(td) = 2td/3 - t^5 d^{\odot 5}/4320 + O(t^7)$ and $W_*(t) = \Lambda/9 + O(t^2)$. The same projection argument gives $I_{\mathrm{eff}}(t) = t^{10} Q_\Lambda(d^{\odot 5})/12960^2 + O(t^{12})$. Hence $I_{\mathrm{eff}} = \Theta(t^{10})$ if $C^\top d^{\odot 5} \neq 0$, as on every strictly ordered triangle, since $x^5 + y^5 - (x+y)^5 \neq 0$ for $x, y > 0$. These are information statements along small-gap paths; the estimation limit below assumes a fixed locally identified parameter.

B.3 Estimation and perturbations

Uncertainty across branches. For fixed independent binomial samples on each edge, take the simultaneous intervals $[L_e, U_e]$ used by Exact-Cycle in Appendix E. Define

$$\widehat{\mathcal{I}}_{1-\delta} = \{\nu \geq 0 : \exists r \perp \mathbf{1}, \quad D_\nu(\mathrm{logit}\, L_e) \leq (Br)_e \leq D_\nu(\mathrm{logit}\, U_e) \text{ for all } e\}. \tag{21}$$

Endpoints $0, 1$ give extended bounds $-\infty, +\infty$. For each candidate ν, this is linear feasibility in r. All feasible components, including unbounded ones, are retained. On the joint edge-coverage event, every parameter generating the true q satisfies these inequalities; hence $\mathbb{P}\{\mathcal{I}_G(\ell) \subseteq \widehat{\mathcal{I}}_{1-\delta}\} \geq 1 - \delta$. No near-truth interval or choice of a single root is needed.

Two candidate solutions (ν_a, r^a), (ν_b, r^b) can be separated by an additional pair (u, v) when their predicted probabilities $q_{\nu_a}(r^a_u - r^a_v)$ and $q_{\nu_b}(r^b_u - r^b_v)$ differ. An added edge set separates a finite collection of candidates exactly when every two candidates differ on at least one added edge. If all existing-response predictions agree, a new response needs connections forming a new cycle; the example in Appendix F.5 uses two such comparisons.

Let $\check{\ell}$ be the logits of empirical edge proportions clipped to $[a_{n_{\min}}, 1 - a_{n_{\min}}]$, where $0 < a_j < 1/2$, $a_j \downarrow 0$, and $n_{\min} = \min_e n_e \to \infty$. On a compact interval $\mathcal{N}$, define

$$\widehat{\nu} \in \operatorname*{arg\,min}_{\nu \in \mathcal{N}} \|C^\top D_\nu(\check{\ell})\|_2^2, \qquad \widehat{r} = (B^\top B)^\dagger B^\top D_{\widehat{\nu}}(\check{\ell}). \tag{22}$$

Corollary 12. *If observations on each edge are i.i.d. Bernoulli and $\mathcal{I}_G(\ell) \cap \mathcal{N} = \{\nu_0\}$, every measurable choice of these minimizers converges almost surely to (ν_0, r).*

Proof. The strong law gives $\check{\ell} \to \ell$. Joint continuity of $D_\nu(z)$, a finite edge set, and compactness give uniform convergence of the profile criterion. Its unique zero is separated from every closed set avoiding ν_0, proving argmin consistency. Continuity of the displayed reward map completes the argument.

Theorem 13. *Suppose all edge records are independent, $n_e/N \to \lambda_e > 0$, and the true finite-gap parameter is an interior point of a compact anchored parameter set. Assume it is the unique parameter in that set generating q, and $I_{\mathrm{eff}} > 0$. Every measurable global MLE of the correctly specified binary Davidson likelihood over this compact parameter set is consistent, and*

$$\sqrt{N}(\widehat{\nu}_{\mathrm{ML}} - \nu_0) \Longrightarrow \mathcal{N}(0, I_{\mathrm{eff}}^{-1}).$$

Its full asymptotic covariance is $(\mathcal{J}^\top \Gamma \mathcal{J})^{-1}$. The ν-coordinate attains the efficient asymptotic covariance. The local asymptotic minimax squared-error lower bound has the same constant I_{eff}^{-1}.

Proof. After normalization by N, the likelihood converges uniformly on the compact parameter set. Its population difference from the truth is the negative sum of Bernoulli KL divergences with positive weights, so uniqueness gives consistency. The score central limit theorem and uniform Hessian convergence give asymptotic covariance $(\mathcal{J}^\top \Gamma \mathcal{J})^{-1}$; its first diagonal entry is the inverse Schur complement I_{eff}^{-1}. Bounded derivatives on an interior neighborhood also give the local log-likelihood expansion $h^\top Z_N - \frac{1}{2} h^\top \mathcal{I} h + o_P(1)$ under $N^{-1/2}$ perturbations, where $\mathcal{I} = \mathcal{J}^\top \Gamma \mathcal{J}$ and $Z_N \Rightarrow \mathcal{N}(0, \mathcal{I})$. The Gaussian shift bound, applied first to truncated squared loss on bounded local neighborhoods and then taking both limits, gives the stated minimax constant.

An isolating interval satisfying Equation (19) verifies uniqueness. Over a larger search set, additional comparisons must separate the feasible branches through their cycle equations; the number of independent cycles alone does not ensure uniqueness.

Noise-corrected DPO also studies random preference flips (Ray Chowdhury et al., 2024). The following bound describes how perturbed edge probabilities change the cycle signal.

Theorem 14. *Let $q_e^0 = q_{\nu_0}((Br)_e)$, with $q_e, q_e^0 \in [\eta, 1-\eta]$, $0 < \eta \le 1/2$, and $|q_e - q_e^0| \le \rho_e$. For every circulation c,*

$$|c^\top \operatorname{logit}(q) - c^\top \operatorname{logit}(q^0)| \le \frac{\sum_e |c_e| \rho_e}{\eta(1-\eta)}. \tag{23}$$

The baseline sign is preserved if its magnitude exceeds this bound.

Proof. The derivative of logit is at most $\{\eta(1-\eta)\}^{-1}$ on the stated interval. Apply the mean-value theorem edgewise and sum with weights $|c_e|$. For a ternary law P_e within total variation ζ_e of the baseline and tie orientation satisfying $|a_e - \frac{1}{2}| \le b_e$, one may take $\rho_e = \zeta_e + b_e P_e(\text{tie})$: the uniform encoding lies in $[0,1]$, and changing the orientation adds at most $b_e P_e(\text{tie})$. For a mixture $q_e = \sum_j \lambda_{ej} q_{ej}$, convexity gives $\rho_e = \sum_j \lambda_{ej} |q_{ej} - q_e^0|$.

C Projection, Risk, and Ranking

The population projection separates the fit to processed comparisons from the recovery of source rewards. Throughout this section the graph is finite and connected, $w_e > 0$, $q_e = \sigma(\ell_e) \in (0,1)$, and scores are centered, including the source rewards r. Write $g(z) = \log(1+e^z)$, $L_W = B^\top W B$, $W_+ = \sum_e w_e$, and $\mathcal{R}_{\text{edge}} = \sum_e w_e\{g(\ell_e) - q_e \ell_e\}$. The population logistic risk is

$$\mathcal{R}(s) = \sum_{e \in E} w_e \{g((Bs)_e) - q_e (Bs)_e\}. \tag{24}$$

Theorem 15. *The unique minimizer of the population risk in Proposition 5 satisfies Equation (11), with exact edgewise fit precisely when $C^\top \ell = 0$.*

Proof. The Hessian is $B^\top W \operatorname{diag}\{\sigma(Bs)(1-\sigma(Bs))\}B$, positive definite on $\mathbf{1}^\perp$. Put $\eta = \min_e \min(q_e, 1-q_e) > 0$. Since $g(z) - q_e z \ge \eta|z|$ and B is injective on $\mathbf{1}^\perp$, the risk is coercive there. Existence, uniqueness, and the balance equation follow. Each scalar loss is uniquely minimized at ℓ_e; all edges attain their minima exactly when $Bs = \ell$, equivalent to cycle closure. Finally, $s_u = \beta \log\{\pi(u)/\pi_{\text{ref}}(u)\}$ modulo its mean is a bijection between full-support policies and centered scores.

For completeness, if the desired logits equal cBr, connectedness forces $\beta \log\{\pi(u)/\pi_{\text{ref}}(u)\} = cr_u + A$. Normalization gives $\pi_c(u) \propto \pi_{\text{ref}}(u) e^{cr_u/\beta}$, proving Corollary 3. A temperature change preserves the score-difference subspace; it cannot fit a nonzero cycle sum. For a compatible scale c, the oracle policy is recovered at $\beta_{\text{train}} = c\beta_{\text{obj}}$.

C.1 Cycle residuals and fit error

Let s^Q minimize $\|\ell - Bs\|_W^2$, with value $Q_W(\ell)$. A circulation is an edge vector γ with $B^\top \gamma = 0$. When the graph contains a cycle, the Hodge dual is

$$Q_W(\ell) = \sup_{0 \ne \gamma \in \ker(B^\top)} \frac{(\gamma^\top \ell)^2}{\gamma^\top W^{-1} \gamma}. \tag{25}$$

The normal equation gives $s^Q = L_W^\dagger B^\top W \ell$. For every circulation γ, weighted Cauchy–Schwarz yields

$$(\gamma^\top \ell)^2 = (\gamma^\top (\ell - Bs^Q))^2 \le (\gamma^\top W^{-1} \gamma) Q_W(\ell).$$

If $Q_W > 0$, equality is attained by $\gamma = W(\ell - Bs^Q)$, which is a circulation by the normal equation. If $Q_W = 0$, every numerator is zero. On a tree B is surjective, so $Q_W = 0$.

Theorem 16. *Let $\psi(t) = t - 1 + e^{-t}$, $a_{\min} = \min_e w_e q_e(1-q_e)$, and $\mathcal{E}^\dagger = \mathcal{R}(s^\dagger) - \mathcal{R}_{\text{edge}}$. Then*

$$a_{\min}\psi\Big(\sqrt{Q_W(\ell)/W_+}\Big) \le \mathcal{E}^\dagger \le Q_W(\ell)/8. \tag{26}$$

For a simple cycle $\mathcal{C}$ of length k, put $\Delta_\mathcal{C} = c_\mathcal{C}^\top \ell$ and $a_\mathcal{C} = \min_{e\in\mathcal{C}} w_e q_e(1-q_e)$. Every score s satisfies

$$\mathcal{R}(s) - \mathcal{R}_{\text{edge}} \ge k a_\mathcal{C} \psi(|\Delta_\mathcal{C}|/k). \tag{27}$$

If ℓ and Bs lie in $[-K, K]^m$, their excess risk lies between $a_K\|\ell - Bs\|_W^2/2$ and $\|\ell - Bs\|_W^2/8$, where $a_K = \sigma(K)\{1-\sigma(K)\}$.

Proof. Since $g'' \le 1/4$ and $g''(\ell + t) \ge e^{-|t|} g''(\ell)$, integral Taylor expansion gives

$$q(1-q)\psi(|z-\ell|) \le g(z) - g(\ell) - q(z-\ell) \le (z-\ell)^2/8.$$

The excess risk is the weighted sum of these scalar divergences. Evaluating the upper bound at s^Q proves the upper sandwich bound. For every s, some edge has $|(Bs)_e - \ell_e| \ge \sqrt{Q_W/W_+}$, which proves its lower bound. On a cycle, telescoping gives $\sum_{e\in\mathcal{C}} |(Bs)_e - \ell_e| \ge |\Delta_\mathcal{C}|$; convexity and monotonicity of ψ prove the cycle bound by Jensen's inequality. On $[-K, K]$, $g'' \ge a_K$, giving the quadratic bound.

The lower bound in Equation (26) uses one edge and its weight; Equation (27) instead aggregates a whole cycle. The bounded-logit bound retains the full weighted squared error.

Corollary 17. *For finitely distributed prompts with probabilities p_x, consider any shared score-difference class $\mathcal{Z}$. If prompt x has a cycle $\mathcal{C}_x$ of length k_x, then its approximation error above the conditional edgewise oracle satisfies*

$$\inf_{z\in\mathcal{Z}} \{\mathcal{R}(z) - \mathcal{R}_{\text{edge}}\} \ge \sum_x p_x k_x a_{x,\mathcal{C}_x} \psi(|\Delta_{x,\mathcal{C}_x}|/k_x).$$

Proof. The cycle bound holds for every class member separately at each prompt. Average these inequalities and then take the common infimum.

For a policy, $s = \beta \log(\pi/\pi_{\text{ref}})$ modulo constants. Therefore its weighted squared error in differences of policy/reference log ratios, relative to ℓ/β, is at least $Q_W(\ell)/\beta^2$.

C.2 Source-reward regret and ranking

Cycle closure concerns the processed target. Its relation to source rewards additionally depends on the transformed link and comparison design.

Theorem 18. *For the common-parameter Davidson model, let $g_r = B^\top W\{\sigma(Br) - q_\nu(Br)\}$ and $\alpha_* = \min_u \min\{\pi^*(u), \pi^\dagger(u)\} > 0$. At a common training and objective temperature β,*

$$\mathcal{J}_r(\pi^*) - \mathcal{J}_r(\pi^\dagger) = \beta D_{\text{KL}}(\pi^\dagger\|\pi^*) \ge \frac{8\alpha_*\|g_r\|_2^2}{\beta\lambda_{\max}(L_W)^2}. \tag{28}$$

The gap is positive for $\nu > 0$ and nonconstant r. At $\nu = 1$, setting $\beta_{\text{train}} = \beta_{\text{obj}}/2$ recovers π^*.

Proof. Substitution of π^* gives $\mathcal{J}_r(\pi) = \beta \log \sum_u \pi_{\text{ref}}(u) e^{r_u/\beta} - \beta D_{\text{KL}}(\pi\|\pi^*)$. For $v = s^\dagger - r \perp \mathbf{1}$, all policy coordinates along the score segment are at least α_*: their logarithms are concave along that segment. Consequently $\text{Var}_{\pi_s}(v) = \min_a \sum_u \pi_s(u)(v_u - a)^2 \ge \alpha_*\|v\|_2^2$. The log-partition Hessian is $\beta^{-2}\,\text{Cov}_{\pi_s}$, so its integral Taylor formula gives $D_{\text{KL}}(\pi^\dagger\|\pi^*) \ge \alpha_*\|v\|_2^2/(2\beta^2)$. The risk Hessian is at most $L_W/4$, hence $\|g_r\|_2 \le \lambda_{\max}(L_W)\|v\|_2/4$, proving the bound. For $\nu > 0$, $t > 0$, and $a = e^{t/2} > 1$, $1 < (a+\nu)/(a^{-1}+\nu) < a^2$, so $0 < h_\nu(t) < t$. Oddness therefore gives $r^\top g_r = \sum_e w_e(Br)_e\{\sigma((Br)_e) - q_\nu((Br)_e)\} > 0$ whenever r is nonconstant on a connected graph. Thus $g_r \ne 0$. The last assertion follows from $h_1(t) = t/2$.

Theorem 19. *Under the common-parameter Davidson model, on a complete graph with equal positive edge weights, $s_i^\dagger > s_j^\dagger$ exactly when $r_i > r_j$, and score ties coincide with reward ties.*

Proof. Define $F_i(s) = \sum_{k \neq i} \sigma(s_i - s_k)$ and $G_i(r) = \sum_{k \neq i} q_\nu(r_i - r_k)$. Balance gives $F_i(s^\dagger) = G_i(r)$. In $F_i - F_j$, every common-neighbor term and the direct comparison have the sign of $s_i - s_j$; equality holds exactly when that gap vanishes. Strict monotonicity of q_ν gives the same property for $G_i - G_j$ and $r_i - r_j$. Their signs therefore agree.

Proposition 20. *For each finite $\nu > 0$, $\nu \neq 1$, a four-response path with any positive edge weights admits rewards with $r_0 > r_3$ but $s_0^\dagger < s_3^\dagger$. The reversal concerns the unobserved endpoint pair; all observed edges fit exactly. At $\nu = 0, 1$, every connected design preserves reward order. Three edges are minimal for endpoint reversal on a path, and a reversal requires gaps of both signs.*

Proof. Write $h = h_\nu$. If $0 < \nu < 1$, then $h(2) - 2h(1) < 0$. Choose $\varepsilon > 0$ small enough that $h(2+\varepsilon) - 2h(1) < 0$, and use successive path gaps $(2+\varepsilon, -1, -1)$. Their reward sum is $\varepsilon > 0$; their transformed sum is negative. If $\nu > 1$, choose small $0 < \varepsilon < 2$ with $2h(1) - h(2-\varepsilon) < 0$, and use $(1, 1, -2+\varepsilon)$. A tree integrates all edge logits, so these sums are exactly the fitted endpoint gaps regardless of positive weights. At $\nu = 0, 1$, the fitted scores are r and $r/2$ modulo constants. On a two-edge path, $x + y > 0$ implies $x > -y$, so strict increase and oddness give $h(x) + h(y) > 0$. Gaps of one sign also retain that sign after transformation, precluding an endpoint reversal.

Temperature calibration. Here temperature varies while the population score and objective temperature remain fixed. The score minimizer is independent of training temperature because full-support policies parameterize every centered score.

Proposition 21. *Fix a full-support reference, centered nonconstant r, and $\beta_{\mathrm{obj}} > 0$. Write $\pi^*(u) \propto \pi_{\mathrm{ref}}(u) e^{r_u/\beta_{\mathrm{obj}}}$ and $\pi_\tau(u) \propto \pi_{\mathrm{ref}}(u) e^{s_u^\dagger/\tau}$, $\tau > 0$. Then*

$$\inf_{\tau > 0} D_{\mathrm{KL}}(\pi_\tau \| \pi^*) = 0 \quad \Longleftrightarrow \quad s^\dagger = ar \text{ for some } a > 0.$$

In that case, $\tau = a\beta_{\mathrm{obj}}$ recovers π^*. Otherwise even the optimally calibrated source-reward regret is positive.

Proof. Equality of policies is equivalent to $s^\dagger/\tau = r/\beta_{\mathrm{obj}}$, after centering log ratios to the reference. If $D_{\mathrm{KL}}(\pi_{\tau_k} \| \pi^*) \to 0$, Pinsker's inequality gives $\pi_{\tau_k} \to \pi^*$. Full support then gives $s^\dagger/\tau_k \to r/\beta_{\mathrm{obj}} \neq 0$. Thus $s^\dagger \neq 0$ and $1/\tau_k \to b > 0$, forcing $s^\dagger = r/(b\beta_{\mathrm{obj}})$. The regret identity in Theorem 18, at fixed β_{obj}, completes the claim.

For finitely many prompts of positive mass and a common β_{obj}, one finite temperature recovers every oracle policy exactly when $s_x^\dagger = ar_x$ for all prompts with the same $a > 0$. This follows by equating centered log ratios promptwise. Order preservation alone does not imply this proportionality.

For the equal-weight triangle $r = (1, 0, -1)$, reflection symmetry and uniqueness force $s^\dagger = (a, 0, -a)$, where $a > 0$ solves $\sigma(a) + \sigma(2a) = q_\nu(1) + q_\nu(2)$. Calibration therefore recovers the oracle although $\Delta_\nu(1, 1) \neq 0$ for $\nu \notin \{0, 1\}$. For a contrasting cyclic design, add the endpoint edge to a reversing path from Proposition 20, with sufficiently small positive relative weight. The anchored Hessian at zero added weight is positive definite, so the implicit function theorem preserves the strict reversal. Writing $T < 0$ for the transformed path sum, the new cycle has residual $T - h_\nu(r_0 - r_3) < 0$. The fitted score is not a positive multiple of r, so calibration leaves positive regret.

Theorem 22. *On a common oriented edge universe, let (W^0, q^0) and (W, q) have connected positive-weight supports and centered projections $s^0, s^\dagger$; zero weights encode missing edges. Suppose the logits along their score segment, on the support of W, lie in $[-K, K]$. Define $a_K = \sigma(K)\{1 - \sigma(K)\}$ and*

$$\varepsilon_{\mathrm{cov}} = \frac{\left\| B^\top [(W - W^0)\{\sigma(Bs^0) - q^0\} + W(q^0 - q)] \right\|_2}{a_K \lambda_2(L_W)}.$$

Then $\|s^\dagger - s^0\|_2 \leq \varepsilon_{\mathrm{cov}}$ and $|(s_i^\dagger - s_j^\dagger) - (s_i^0 - s_j^0)| \leq \sqrt{2}\, \varepsilon_{\mathrm{cov}}$. Every baseline gap exceeding $\sqrt{2}\, \varepsilon_{\mathrm{cov}}$ retains its sign. The segment condition holds a priori if $K_0 := \max_{e : w_e > 0} |(Bs^0)_e| < K$ and $\sqrt{2}\, \varepsilon_{\mathrm{cov}} \leq K - K_0$.

Proof. Subtract the baseline balance equation from $\nabla\mathcal{R}(s^0)$; this gives the numerator vector above. Along the segment, $\nabla^2\mathcal{R} \succeq a_K L_W$. With $v = s^\dagger - s^0 \perp \mathbf{1}$, strong monotonicity and $\nabla\mathcal{R}(s^\dagger) = 0$ imply $a_K\lambda_2(L_W)\|v\|_2^2 \le \|v\|_2\|\nabla\mathcal{R}(s^0)\|_2$. Division proves the score bound (the zero case is immediate); $|v_i - v_j| \le \sqrt{2}\|v\|_2$ proves the pairwise bound.

For the a priori condition, put $R = (K - K_0)/\sqrt{2}$ and $\mu = a_K\lambda_2(L_W)$. Every centered displacement with $\|v\|_2 \le R$ satisfies $|(Bv)_e| \le \sqrt{2}R$, so the entire ball has logits in $[-K, K]$ on the target support. On its boundary, $v^\top\nabla\mathcal{R}(s^0 + v) \ge \mu R^2 - R\|\nabla\mathcal{R}(s^0)\|_2 \ge 0$. If the minimizer lay outside, strict convexity along the ray from s^0 to that minimizer would make the directional derivative strictly positive there, contradicting stationarity. The minimizer therefore lies in the ball, establishing the segment condition.

D Tie Handling and Information

Table 4 records the population targets on a connected graph. The retained-outcome likelihood uses the correctly specified Davidson model; Dav-MLE in the experiments uses the generating ν. Explicit tie models are also used for LLM-based ranking (Ameli et al., 2025).

Table 4: Population targets of tie handling. Here z is the fitted binary logit and p^+, p^0, p^- are the original outcome probabilities.

Rule	Observation	Population target
Discard ties	Strict win/loss	$z = d$, since $p^+/p^- = e^d$
Retain ties	Davidson outcome	Source reward, from the same odds ratio
Logit-Half	$X \in \{0, 1/2, 1\}$	Logistic projection of $q_\nu(d)$
Logit-Coin	$\widetilde{X} \in \{0, 1\}$	Same projection as Logit-Half

Population targets. Conditioning on a strict outcome gives probability $\sigma(d)$. Its retention probability $s = (e^d + 1)/(e^d + 1 + 2\nu e^{d/2})$ is positive on every edge, so discarding ties preserves the connected population graph and its oracle target. For ternary fitting, excess cross-entropy is the KL divergence from the true three-outcome distribution. It vanishes only at that distribution, whose ratio $p^+/p^- = e^d$ determines the reward gaps.

For $\ell(z; X) = \log(1 + e^z) - Xz$, conditioning on the original label gives $\mathbb{E}[\widetilde{X} \mid X] = X$ and $\mathrm{Var}(\widetilde{X} \mid X) = \mathbf{1}_{\{X=1/2\}}/4$. Thus both encodings have mean $q_\nu(d)$ and identical expected logistic loss. Total variance yields

$$\mathrm{Var}(\widetilde{X}) = \mathrm{Var}(X) + p^0/4.$$

The same increment holds for the gradient $\sigma(z) - X$; the loss variance increment is $z^2 p^0/4$. Each increment is divided by n for an average of n independent records.

Fixed-edge information. For the Davidson model, assume ν is known and only d is estimated. Write $q = q_\nu(d)$, $a = \sigma(d)$, and $v = \mathrm{Var}(X)$. The ternary probabilities form an exponential family with sufficient statistic X, hence score $X - q$ and $q'_\nu(d) = v$. Consequently,

$$I_{\mathrm{full}} = v = sa(1-a) + s(1-s)(a - \tfrac{1}{2})^2, \qquad I_{\mathrm{hard}} = \frac{v^2}{q(1-q)} = \frac{v^2}{v + p^0/4}.$$

Here the variance decomposition conditions on whether the result is strict: its conditional variance is $a(1-a)$, and the two conditional means are a and $1/2$. Conditional binary fitting uses $I_{\mathrm{drop}} = sa(1-a)$ per original comparison. Modeling the retention indicator adds $s(1-s)(a-\frac{1}{2})^2$, recovering I_{full}. Since $v > 0$, randomization strictly reduces information whenever $p^0 > 0$. A three-valued label fitted with the Davidson likelihood retains this full information; fitting it with ordinary logistic loss instead has the processed target in Equation (11).

E Cycle Tests and Detection Rates

Recent studies examine non-transitive judgments (Xu et al., 2025), richer ordinal feedback (Liu et al., 2025), and uncertainty in comparisons and rankings (Fathullah & Gales, 2025). Here the target is

additive logit consistency. A finite-sample certificate compares a signed cycle sum with uncertainty in its edge logits. Fix an orientation of a graph with m edges, write $\ell_e = \mathrm{logit}(q_e)$, and let $\Delta_{\mathcal{C}} = c_{\mathcal{C}}^\top \ell$, where $c_{\mathcal{C},e} \in \{-1, 0, 1\}$ records each simple cycle's orientation.

Theorem 23 (Simultaneous cycle confidence intervals)**.** *For each edge, take a fixed number $n_e \geq 1$ of observations $Y_{e,j} \in [0,1]$ whose conditional mean, given the common history before that observation, is $q_e \in [\eta, 1-\eta]$, with $0 < \eta \leq 1/2$. Set $\widehat{q}_e = n_e^{-1} \sum_j Y_{e,j}$, $\widehat{\ell}_e = \mathrm{logit}\{\Pi_{[\eta/2, 1-\eta/2]}(\widehat{q}_e)\}$, and $\kappa_\eta = \{(\eta/2)(1-\eta/2)\}^{-1}$. For $0 < \delta < 1$, with probability at least $1-\delta$, simultaneously for all simple cycles,*

$$\left| c_{\mathcal{C}}^\top \widehat{\ell} - \Delta_{\mathcal{C}} \right| \leq \tau_{\mathcal{C}} := \kappa_\eta \sum_{e \in \mathcal{C}} \sqrt{\frac{\log(2m/\delta)}{2n_e}}. \tag{29}$$

Rejecting whenever any resulting interval excludes zero controls family-wise type-I error at δ, including for data-selected cycles. On the same event, a cycle is detected if $|\Delta_{\mathcal{C}}| > 2\tau_{\mathcal{C}}$.

Proof. Conditional Hoeffding concentration gives $\mathbb{P}(|\widehat{q}_e - q_e| > \epsilon) \leq 2e^{-2n_e\epsilon^2}$. A union bound yields the stated edgewise bounds for all m edges. Clipping cannot increase the error because q_e lies in the clipping interval; the logit derivative there is at most κ_η. Summing these bounds proves simultaneous coverage of every cycle. BT-compatible logits have zero cycle sums, and $|c_{\mathcal{C}}^\top \widehat{\ell}| \geq |\Delta_{\mathcal{C}}| - \tau_{\mathcal{C}}$ proves detection. No independence across edges is needed.

For a k-edge cycle with $n_e = n$, the sufficient budget is

$$n > \frac{2k^2\kappa_\eta^2}{\Delta_{\mathcal{C}}^2} \log \frac{2m}{\delta}. \tag{30}$$

The same proof covers predictable edge selection up to a fixed horizon T, with current counts $N_e(t) \geq 1$: replace n_e by $N_e(t)$ and $\log(2m/\delta)$ by $\log(2mT/\delta)$. Indeed, each edge's successive observations are a bounded martingale difference sequence. Union-bounding its Hoeffding events over the mT edge/count pairs gives coverage at every $t \leq T$, including bounded stopping times.

Exact-Cycle. For binomial counts $S_e \sim \mathrm{Bin}(n_e, q_e)$, take Clopper–Pearson intervals $[L_e, U_e]$ of confidence $1 - \delta/m$. For $S_e > 0$, L_e is the $\delta/(2m)$-quantile of $\mathrm{Beta}(S_e, n_e - S_e + 1)$, otherwise $L_e = 0$. For $S_e < n_e$, U_e is the $1 - \delta/(2m)$-quantile of $\mathrm{Beta}(S_e + 1, n_e - S_e)$, otherwise $U_e = 1$. Transform the endpoints by logit and add them with the cycle signs, reversing endpoints for negative terms. Binomial coverage, a union bound, and monotonicity of logit prove simultaneous coverage for all cycles, with extended endpoints $\mathrm{logit}(0) = -\infty$ and $\mathrm{logit}(1) = +\infty$. This construction requires no specified η.

Theorem 24 (Cycle-Wald inference)**.** *Assume independent Bernoulli observations within and across edges. For a fixed cycle c, let $n_e/N \to \lambda_e > 0$, and define*

$$\bar{q}_e = \frac{S_e + 1/2}{n_e + 1}, \qquad \bar{\Delta} = c^\top \mathrm{logit}(\bar{q}), \qquad \widehat{v}_N = \sum_{e : c_e \neq 0} \frac{c_e^2}{n_e \bar{q}_e (1 - \bar{q}_e)}. \tag{31}$$

If $q_{e,N} \to q_e^0 \in (0,1)$, then $\sqrt{N}(\bar{\Delta} - \Delta_N) \Rightarrow \mathcal{N}(0, V_0)$ and $N\widehat{v}_N \to V_0$ in probability, where $V_0 = \sum_e c_e^2/\{\lambda_e q_e^0(1-q_e^0)\}$. Thus $\bar{\Delta}/\sqrt{\widehat{v}_N}$ is asymptotically standard normal under the cycle null. If $\sqrt{N}\Delta_N \to \gamma$, its limit is $\mathcal{N}(\gamma/\sqrt{V_0}, 1)$.

Proof. Smoothing changes each sample proportion by $O(n_e^{-1})$. The independent Bernoulli central limit theorem and the delta method, with logit derivative $1/\{q_e(1-q_e)\}$, give the stated variance; consistency of $\widehat{v}_N$ gives the standardized limits. The Fisher information for free edge probabilities is diagonal with entries $\lambda_e/\{q_e(1-q_e)\}$. Applying its inverse to the gradient of the cycle functional gives exactly V_0, proving efficiency in this product-Bernoulli model.

For a fixed total budget, Cauchy–Schwarz minimizes this variance at $\lambda_e \propto |c_e|/\sqrt{q_e(1-q_e)}$. These are the oracle cycle-information (OCI) allocation weights used in the experiments.

Theorem 25 (Compatible-null lower bound)**.** *Let P_ℓ be a three-edge product-Bernoulli law with finite logits ℓ, cycle sum $\Delta = c^\top \ell \neq 0$, and n independent records per edge. There is a BT-compatible law P_0 such that every test ϕ satisfies*

$$P_0(\phi = 1) + P_\ell(\phi = 0) \geq 1 - |\Delta|\sqrt{n/48}. \tag{32}$$

Both errors can be at most $0 < \alpha < 1/2$ only if $n \geq 48(1-2\alpha)^2/\Delta^2$.

Proof. Set $\ell^0 = \ell - (\Delta/3)c$. Since $c^\top c = 3$, $c^\top \ell^0 = 0$, so these logits are BT-compatible on the triangle. The Bernoulli log-partition function $A(u) = \log(1 + e^u)$ satisfies $A''(u) \leq 1/4$. Its Taylor remainder bounds the single-record KL divergence by $(\ell_e - \ell_e^0)^2/8$; hence $D_{\mathrm{KL}}(P_\ell \| P_0) \leq n\Delta^2/24$. Pinsker's inequality bounds total variation by $|\Delta|\sqrt{n/48}$. The error sum is at least one minus total variation, proving the claim.

Proof of Corollary 6. For fixed $x, y > 0$ and $\nu \in (0, \infty) \setminus \{1\}$, the expansion in Equation (14) gives $\Delta_t = Ct^3 + O(t^5)$, with $C = \nu(1-\nu)xy(x+y)/\{8(1+\nu)^3\} \neq 0$. If $nt^6 \to 0$, Theorem 25 makes the paired-null error-sum limit inferior at least one. If $nt^6 \to \infty$, the edge probabilities tend to $1/2$, $\bar{\Delta} - \Delta_t = O_{\mathbb{P}}(n^{-1/2})$, and $\widehat{v}_n \asymp_{\mathbb{P}} n^{-1}$. The standardized signal therefore diverges, so any fixed-level two-sided Cycle-Wald test has power tending to one.

At $\nu = 1/2$, this forward cycle signal remains cubic. The t^{10} information in Appendix B concerns estimating ν after profiling out rewards: the cubic term of j_ν, rather than that of the cycle residual, vanishes there.

Effect size and comparison budget. For a fixed k-edge cycle and total equal-allocation budget $N = kn$, Equation (29) detects $|\Delta_{\mathcal{C}}| > \kappa_\eta \sqrt{2k^3 \log(2m/\delta)/N}$ on its coverage event. For the independent triangle model, Theorem 25 requires $N \geq 144(1-2\alpha)^2/\Delta^2$ to make both errors at most α. These bounds specify budgets for a target logit effect, rather than a fixed requirement for every dataset. The audit first needs same-prompt cycles and repeated judgments of their fixed response pairs, as counted in Table 5. Acyclic components supply no cycle restrictions; repetition controls the precision on existing cycles.

F Experimental Details

F.1 Data and comparison graphs

The model comparisons use `irisxx/ultrafeedback_tied` and `irisxx/chatarena_tied`, derived from UltraFeedback (Cui et al., 2023) and Chatbot Arena (Chiang et al., 2024). The graph study uses `openbmb/UltraFeedback`. All 145,500 raw records passed schema checks; removing empty records, duplicates, and training rows with test prompts leaves 142,114 records without changing labels or scores.

The semi-synthetic graph study assigns rewards from response scores and sets the reward scale and tie strength. Its 63,687 response sets give 250,185 triangles at each of three scales and seven tie strengths. Among them, 51,712 have at least three distinct scores; incompatible settings have the predicted residual sign and complete equal-weight projections preserve strict order. Table 5 counts same-prompt cycles and fixed-pair repetitions using exact response text, excluding self-pairs and collapsing repeated edges.

The cycle test uses the observed cycles and edge counts, with the detection threshold stated in Equation (30); it is defined whenever the comparison graph contains at least one cycle.

F.2 Controlled designs and cycle tests

The controlled matrix uses $\nu \in \{0, 0.1, 0.25, 0.5, 1, 2, 4\}$, gap scales $\{0.125, 0.25, 0.5, 1, 2\}$, and complete, cycle, Erdős–Rényi, and path graphs. Stochastic summaries average three seeds. The phase grid has 82 tie strengths and 60 values for each adjacent gap, totaling 295,200 cells. Small-gap slopes use the smallest gap scales at five incompatible tie strengths. Complete-graph ranking uses 630 runs; five tie-specific paths give 15 reversal runs.

Table 5: **Comparison graph coverage.** Records are retained dataset rows. Graphs use exact prompt and response text and exclude self-pairs; repeated pairs have at least two records within one prompt.

Dataset / split	Records	Pairs	Cyclic prompts	Repeated pairs	Max. rows per pair
UltraFeedback / train	19,384	19,245	32	13	2
UltraFeedback / test	1,998	1,977	0	0	1
UltraFeedback / 5k subset	5,000	4,970	0	4	2
ChatArena / train	19,698	19,547	1	15	3
ChatArena / test	1,497	1,491	0	0	1

Cycle tests use adjacent gaps $(0.7t, 1.3t)$, four tie strengths, five scales, and 12 per-edge sample sizes: 240 cells with 900 trials per aggregate cell. The allocation study fixes $\nu = 0.25$ and gaps $(0.2, 4.0, 4.2)$. EA assigns one third of comparisons to each edge; OCI uses the generating probabilities. Table 6 reports rejection rates and the matched-budget comparison. Figure 4 shows the residual and power curves; intervals are 1.96 standard errors across replicate-level rates.

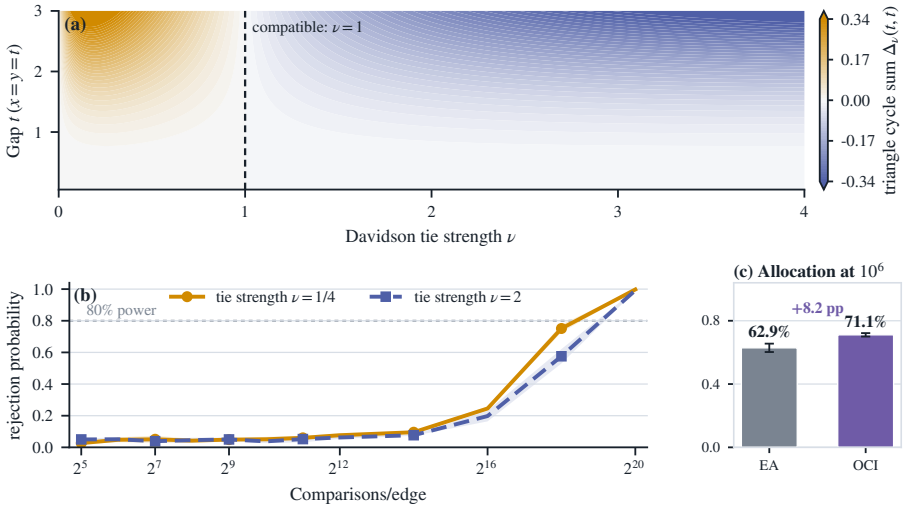


Figure 4: **Cycle diagnostics.** (a) Residual sign; (b) Cycle-Wald power; (c) OCI and EA at a total budget of 10^6. Bands and bars are replicate-level 95% intervals.

Table 6: **Cycle tests across budgets.** Panel (a) reports rejection rates (%) at $t = 1$, each n_e column averaging $\nu \in \{1/4, 2\}$. Exact-Cycle is simultaneous, whereas Cycle-Wald targets a fixed cycle. Panel (b) reports Cycle-Wald power across the EA/OCI budget sweep on the heterogeneous triangle; OCI uses generating probabilities. Bold marks higher power at each displayed budget.

(a) Cycle-test rejection rate

Method	$n_e = 2^{18}$	$n_e = 2^{20}$
Exact-Cycle	5.06	71.78
Cycle-Wald	**66.33**	**99.50**

(b) Cycle-Wald power by total comparison budget

Allocation	30k	100k	300k	1m
EA	**7.44**	11.67	22.33	62.89
OCI	6.11	**14.44**	**26.89**	**71.11**

Temperature-calibrated regret. We calibrate positive temperatures on complete, cycle, path, and Erdős–Rényi graphs with 3, 6, or 8 responses. Centered Gaussian rewards have scale 0.8; edge weights are equal and references uniform or normalized log-normal. With $\beta_{\text{obj}} = 1$, numerical minimization of $D_{\text{KL}}(\pi_\tau \| \pi^*)$ either shares one temperature across 30 instances of each design or calibrates each instance separately. At $\nu = 0, 1$, we use the exact scores $r, r/2$. Table 2 uses complete six-response graphs with uniform references. Table 7 pools $\nu \in \{0.1, 0.25, 0.5, 2, 4\}$. Common calibration is fitted within each graph-size, graph-type, tie-strength, and reference cell.

The individual-regret 95th percentile is 1.96×10^{-5} on complete graphs, with median 1.86×10^{-7}. Cycle graphs retain mean individually calibrated regret 6.26×10^{-5}. Path regrets occur within the BT-compatible class and reflect distortion of source-score direction (Proposition 20).

Table 7: **Calibrated regret across comparison graphs.** Means over the full nonlinear grid: 900 policies and 450 score instances per graph. P95 summarizes individual-calibrated regret.

Graph	Fixed mean $\times 10^{-2}$	Common mean $\times 10^{-4}$	Individual mean $\times 10^{-4}$	Individual P95 $\times 10^{-4}$
Complete	5.6651	0.6202	0.0444	0.1964
Cycle	5.6027	1.7983	0.6256	1.7114
ER	5.3103	0.6735	0.2004	0.9647
Path	5.6554	2.9881	1.4880	6.0252

F.3 Tie-handling estimates

Dav-MLE, BT-Drop, Logit-Half, and Logit-Coin are fitted to the same Davidson samples as defined in Section 6. Dav-MLE uses the generating ν; Logit-Coin uses one fixed tie randomization per dataset. The 4,320 paired datasets yield 17,280 method evaluations. Table 8 disaggregates Table 1. At $\nu = 1$, Logit-Half has source RMSE 0.3546 but only 0.0441 around the half-scale target, versus 0.0631 for Logit-Coin. Its smaller target error at every sampled positive ν agrees with the coin-variance calculation in Appendix D.

Table 8: **Tie-handling errors by tie strength.** Mean source and population-target RMSEs. Dav-MLE uses known ν; source and target coincide for Dav-MLE and BT-Drop. Entries are three-seed means.

ν	Dav-MLE	BT-Drop	Logit-Half		Logit-Coin	
	Source	Source	Source	Target	Source	Target
0.1	0.0737	0.0745	0.1039	0.0662	0.1058	0.0695
0.25	0.0757	0.0773	0.1624	0.0596	0.1653	0.0666
0.5	0.0794	0.0829	0.2491	0.0521	0.2517	0.0650
1	0.0879	0.0940	0.3546	0.0441	0.3573	0.0631
2	0.1078	0.1160	0.4571	0.0372	0.4595	0.0629
4	0.1351	0.1454	0.5445	0.0291	0.5481	0.0639

F.4 Preference prediction

Llama-3.2-3B-Instruct is fine-tuned on UltraFeedback-Tied and Qwen3-4B (Yang et al., 2025) on ChatArena-Tied, using DPO, TODO, and DPO-D from Section 6.

Ternary preference prediction. For each trained model, BF16 inference sums scored completion-token log probabilities at the 512/256 limits and subtracts the corresponding base-model sums. Multiplying their pairwise difference by a positive fitted scale gives the gap supplied to Equation (1). The scale and tie strength are fitted on 1,002 UltraFeedback pairs and evaluated on 996 others, including 250 ties. The two splits are prompt-disjoint and exclude both models' training prompts. Negative log-likelihood (NLL) averages the negative log probability of the observed outcome; Brier averages the sum of squared errors across its three probabilities. Sequence accuracy averages correctness over the 746 strict pairs, giving half credit to equal scores. Response-score RMS is the root mean square of both endpoints' policy-to-base log-ratios.

The no-training baseline uses the calibration tie rate and splits the remaining probability equally. Matched NLL rescales response-score RMS to DPO's on calibration data and holds the fitted probability mapping fixed. Table 10 compares TODO directly with DPO. We resample all 986 test prompts jointly for both methods, preserving every pair within a sampled prompt and keeping trained models and probability mappings fixed. Accuracy uses strict pairs; NLL and Brier use all outcomes.

F.5 Identification with multiple solutions

For a complete triangle with $\nu_0 = 0.25$ and $r = (1, 0, -1)$, write $a = h_{\nu_0}(1)$, $b = h_{\nu_0}(2)$. The inverse-link profile is $F(\nu) = 2D_\nu(a) - D_\nu(b)$. Both solutions in Table 11 generate logits

Table 9: **Held-out evaluation results.** Means on the same held-out UltraFeedback split; Qwen transfers from ChatArena. Sequence accuracy uses strict pairs. Matched NLL fixes the probability map after matching score RMS to DPO; No-training has zero policy-to-base log-ratios.

Objective	Seq. acc. (%)	NLL	Brier	Score RMS	Matched NLL
Llama-3.2-3B					
No-training	50.00	1.0826	0.6565	0.0000	1.0826
DPO	58.58	1.0590	0.6394	1.2375	1.0590
TODO	61.93	1.0587	0.6393	1.4600	1.0590
DPO-D	58.98	1.0607	0.6407	1.4842	1.0606
Qwen3-4B					
No-training	50.00	1.0826	0.6565	0.0000	1.0826
DPO	56.70	1.0768	0.6520	6.9116	1.0768
TODO	55.70	1.0757	0.6512	7.7262	1.0759
DPO-D	55.70	1.0765	0.6518	8.0494	1.0765

Table 10: **Paired TODO–DPO differences.** Llama is the matched UltraFeedback setting used for the main comparison; Qwen is a transfer-only comparison from ChatArena. Entries are mean differences with 95% intervals from 100,000 paired test-prompt bootstrap samples, with trained models and fitted probability mappings fixed. Accuracy differences are percentage points; NLL and Brier differences are scaled by 10^3.

Model	Accuracy difference	NLL difference	Brier difference
Llama-3.2-3B	$+3.35$ $[0.13, 6.59]$	-0.381 $[-8.508, 7.851]$	-0.154 $[-6.065, 5.816]$
Qwen3-4B	-1.01 $[-3.80, 1.80]$	-1.144 $[-3.691, 1.392]$	-0.801 $[-2.702, 1.087]$

$(0.796046, 0.796046, 1.569445)$ and have nonzero profile derivative. With $A = \sinh(a/2)$, $B = \sinh(b/2)$,

$$F'(\nu) = \frac{4A}{\sqrt{1+\nu^2A^2}} - \frac{2B}{\sqrt{1+\nu^2B^2}}.$$

This derivative changes sign once, at $\sqrt{(B^2-4A^2)/(3A^2B^2)} \simeq 0.47518$. The positive observed residual restricts candidates to $(0, 1)$. Since $F(0) > 0$, its minimum is negative, and $F(1) > 0$, there are exactly two roots. We solve both monotone branches.

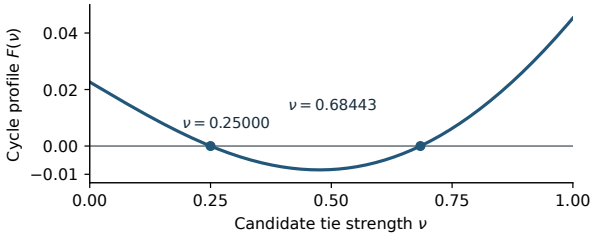


Figure 5: **Identification profile.** Both admissible roots are retained.

Table 11: **Two admissible branches.** Rewards are $r = (s, 0, -s)$; G is the added-cycle profile.

ν	s	$F'(\nu)$	$G(\nu)$
0.250000	1.000000	-0.07010	0
0.684429	1.348336	0.08186	0.01659

Adding response $r_4 = 2$ and edges $(4, 1), (4, 3)$ gives $G(\nu) = D_\nu(a) + D_\nu(b) - D_\nu(h_{\nu_0}(3))$. It is zero at the first solution and 0.01659 at the second, identifying the first uniquely in this augmented design.

Estimation without selecting a branch. We sample $r = t(1, 0, -1)$, $t \in \{0.25, 0.5, 1\}$, at $\nu \in \{0.25, 0.5, 2\}$ and near $\nu = 0.5$, then add the response $r_4 = 2t$ and its two comparisons above. Adaptive numerical outer covers evaluate Equation (21) over all $\nu \geq 0$ with simultaneous coverage. The reported augmented cover intersects the original and augmented constructions, each with error probability 0.025. Table 12 measures their length in $v = \nu/(1+\nu)$. Both designs retain the generating value in all 675 trials and the original triangle's alternate solution in all 540 trials where one exists; the added comparisons narrow these covers while preserving the simultaneous guarantee.

Table 12: **Information and uncertainty after adding comparisons.** I_{eff} is information per record; width is outer-cover length in $v = \nu/(1+\nu)$, averaged over three seeds. Each present edge has $n_e = 2^{18}$ records, so total counts increase from $3n_e$ to $5n_e$. Subscripts 3 and 5 denote the original and augmented edge sets.

ν	t	$I_{\text{eff},3}$	$I_{\text{eff},5}$	Width_3	Width_5
0.25	0.25	2.67×10^{-8}	2.08×10^{-7}	0.8492	0.7697
0.25	0.5	1.52×10^{-6}	1.04×10^{-5}	0.6951	0.6081
0.25	1	6.00×10^{-5}	2.53×10^{-4}	0.5428	0.4262
0.5	0.25	5.49×10^{-13}	1.95×10^{-11}	0.8828	0.8055
0.5	0.5	5.10×10^{-10}	1.61×10^{-8}	0.7365	0.6413
0.5	1	3.53×10^{-7}	7.14×10^{-6}	0.5733	0.4599
2	0.25	5.08×10^{-9}	3.98×10^{-8}	0.9612	0.9280
2	0.5	2.97×10^{-7}	2.12×10^{-6}	0.8792	0.8257
2	1	1.36×10^{-5}	7.14×10^{-5}	0.7758	0.1677

On the original triangle, halving t from 0.5 to 0.25 reduces profiled information per record by a factor of 928 at $\nu = 0.5$, versus 57 at $\nu = 0.25$, matching the small-gap orders in Equation (20). These factors are the predicted cost of separating nearly tied rewards; Table 12 reports the added comparisons that recover identification under that cost.

F.6 Shared-policy experiments

Three prompts have rewards $r_p = a_p(1, 0, -1)$, $a_p \in \{0.7, 1, 1.4\}$, and equal comparison weights. With a uniform reference and $\beta = 1$, the shared policy $\pi_\theta(i \mid p) \propto e^{\theta r_{p,i}}$ has DPO logits $\theta(r_{p,i} - r_{p,j})$, fitted jointly on all nine edges, so sharing reduces to one cross-prompt scale whose cost the Separate and Free controls isolate. The paired Half/Coin comparison uses 512 ternary records per edge and 100 trials per seed; parameter errors are three-seed means.

Separate scales replace θ by θ_p; Free uses arbitrary centered scores for each prompt, and in this design their optima coincide, so their common logistic excess is the cycle mismatch and the additional Shared excess is the cost of cross-prompt sharing (Table 3). A common temperature exactly corrects Shared because its scores are a positive reward scaling; Separate and Free need promptwise temperatures at $\nu = 0.25, 2$. The stored Half/Coin parameter RMSEs at $\nu = 0.25$ are $0.022/0.025$, around the same population target.