MANUFACTURED CURL IN PREFERENCE LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Scalar reward models underpin reinforcement learning from human feedback (RLHF) for large language models (LLMs). But do cycles in human preferences imply that scalar rewards are inadequate? We examine how preference aggregation, prediction targets, and judgment dependence shape the answer. Averaging choice probabilities can create nonadditive log-odds even when every annotator follows scalar utilities. Concentrating disagreement on one response can make this distortion substantially larger. Recovering mean utilities and predicting choices can favor different estimators; calibration and smoothing affect their comparison. Observation design matters independently: shared rankings and independent judgments can have identical Bradley–Terry probabilities but threefold different asymptotic variances in estimated curl. Reassembling human summarization judgments across 363 contexts into disjoint rather than shared worker groups increases mean absolute curl by 24.4% at matched counts. Simulations at the observed comparison counts quantify detection limits, while held-out prediction tests sensitivity to regularization. These results show why measured preference cycles require an audit of aggregation and observation design before they can justify replacing scalar reward models.

1 A CYCLE CAN DESCRIBE DIFFERENT THINGS

Aligning large language models (LLMs) with human preferences often relies on a scalar reward model. In the Bradley–Terry formulation, the probability that response i beats j is $p_{ij} = \sigma(\theta_i - \theta_j)$, where σ is the logistic function (Bradley & Terry, 1952; Stiennon et al., 2020). Three questions must be distinguished. An *ordinal majority cycle* means $p_{ab}, p_{bc}, p_{ca} > 1/2$. A *log-odds cycle* means $\operatorname{logit} p_{ab} + \operatorname{logit} p_{bc} + \operatorname{logit} p_{ca} \neq 0$, which rules out a Bradley–Terry representation on that triangle but does not require a majority cycle. Finally, an *estimated cycle* uses finite-sample estimates of these probabilities. It may be nonzero even when the population log-odds are exactly additive.

A concrete example. Suppose three summaries answer the same prompt. Readers choose A over B with probability 0.75, and B over C with probability 0.75. A Bradley–Terry model multiplies their odds: $3 \times 3 = 9$, predicting probability 0.90 for A over C. If the actual probability is instead 0.75, the majority ordering is still $A > B > C$, but its strengths cannot be represented by additive logistic scores. A triangle is simply these three responses and their three pairwise comparisons. Its log-odds curl measures the failure of the strengths to add up around a closed path; it is not a count of people who contradict themselves.

Why this matters for learning. Scalar reward models support learning from human feedback (Christiano et al., 2017) and direct preference optimization (Rafailov et al., 2023). General pairwise models can instead support Nash learning when preferences do not admit a scalar representation (Munos et al., 2023). A measured cycle might appear to justify that change. But a representation decision based on the wrong sampling reference can confuse collection protocol with population structure. Conversely, replacing heterogeneous preferences by one score can erase meaningful differences. The practical question is which discrepancy is supported by the data and whether modeling it improves the intended task.

The distinction matters in both directions. A large estimated cycle may reflect sampling variation. A small cycle may reflect dependence between observations rather than stronger evidence for a population ranking. To understand the latter possibility, imagine obtaining all three comparisons from

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

the same person instead of assigning a different person to each pair. If a person's judgments tend to agree with a stable ordering, errors around the triangle can cancel. Independent-edge resampling does not preserve that cancellation.

Dependence in paired comparisons is a longstanding statistical topic (Dittrich et al., 2006), as is decomposition of ranking data into gradient and residual components (Jiang et al., 2011; Xu et al., 2016). Tests for lack of fit and limits of detection also precede this study (Singh & Davidov, 2026). We connect these ideas in an observation-design audit with three contributions: a population calculation showing how the distribution of disagreement controls pooling-induced curl; an identical-marginal counterexample and matched-count human reconstruction showing how observation dependence changes measured curl; and prediction comparisons that separate representation from calibration and regularization. Together, they explain what measured cycles can justify about reward models.

One argument across three levels. We ask which preferences to represent, how observations estimate them, and whether a representation helps prediction. Appendix A provides derivations, detection simulations, and the secondary-corpus analysis.

2 Aggregation can create population log-odds curl

Let annotator a have scalar utility $\phi^a = \bar{\phi} + \varepsilon^a$, and make conditionally independent choices with probability $\sigma(\phi_i^a - \phi_j^a)$. Write $\Delta_{ij} = \bar{\phi}_i - \bar{\phi}_j$ and $v_{ij} = \mathrm{Var}(\varepsilon_i^a - \varepsilon_j^a)$. Compare

$$Y_{ij}^{\mathrm{prob}} = \mathrm{logit}\,\mathbb{E}_a[\sigma(\phi_i^a - \phi_j^a)], \qquad Y_{ij}^{\mathrm{logit}} = \mathbb{E}_a[\phi_i^a - \phi_j^a] = \Delta_{ij}. \tag{1}$$

The second operator preserves a gradient when the same population weights apply to every edge. The first can break this property: it averages choice probabilities, a different target. This distinction is related to established marginal-versus-conditional effects of logistic models (Zeger et al., 1988; Neuhaus et al., 1991). Neither operator observes individual utilities directly when the actual data are binary choices.

For centered Gaussian perturbations and small variances, expansion gives

$$Y_{ij}^{\mathrm{prob}} = \Delta_{ij} - \tfrac{1}{2} v_{ij} \tanh(\Delta_{ij}/2) + O(v_{ij}^2). \tag{2}$$

With common pairwise variance V and circular gaps $x + y + z = 0$, the leading small-variance curl is $-(V/2)\sum_{d\in\{x,y,z\}} \tanh(d/2)$. Expanding also in the gaps yields a cubic term $Vxyz/16$. The cubic *form* persists at fixed variance, but that coefficient does not:

$$K_{\mathrm{pop}} = c(V)xyz + O(\|(x,y,z)\|^5), \qquad c(V) = V/16 + O(V^2). \tag{3}$$

Here $c(V) = 2t(V) + 16u(V)^3$, where $u(V) = \mathbb{E}\sigma'(\sqrt{V}Z)$ and $t(V) = \mathbb{E}\sigma^{(3)}(\sqrt{V}Z)$ for standard Gaussian Z. Appendix A.1 derives and numerically checks this coefficient. Figure 1 separates the pooling residual, the coefficient, and the approximation's limits.

Interpreting curl per unit xyz. Shrinking all three gaps by a factor s makes their product shrink as s^3. The limiting ratio $K_{\mathrm{pop}}/(xyz) = c(V)$ separates triangle geometry from heterogeneity. Figure 1b shows why extrapolating $V/16$ to large heterogeneity fails. The corpus does not reveal these latent utility gaps; dividing noisy empirical gaps into observed curl would not reliably estimate the coefficient.

Uneven disagreement changes the leading effect. The cubic cancellation above depends on equal edge variances. At fixed v, define $f_v(d) = \mathrm{logit}\,\mathbb{E}\sigma(d + \sqrt{v}Z)$ and $a(v) = 4\mathbb{E}\sigma'(\sqrt{v}Z)$. Symmetry gives

$$K_{\mathrm{pop}} = \sum_e a(v_e)\Delta_e + O(\|\Delta\|^3). \tag{4}$$

Equal v_e cancels the linear term because $\sum_e \Delta_e = 0$; unequal variances can leave it intact. For example, take mean utilities $(g, -g, 0)$ and let only the third response vary across annotators, with variance S. The curl is $2g - 2f_S(g)$. At $g = 0.5, S = 1$ it is 0.1722, versus 0.0037 when the same total item variance is spread equally across the three responses. In plain terms, disagreement concentrated on one response can distort relative preference strengths much more than equally distributed disagreement. Appendix A.2 derives the result and illustrates the scaling; these are specified populations, not estimates of human utility variance.

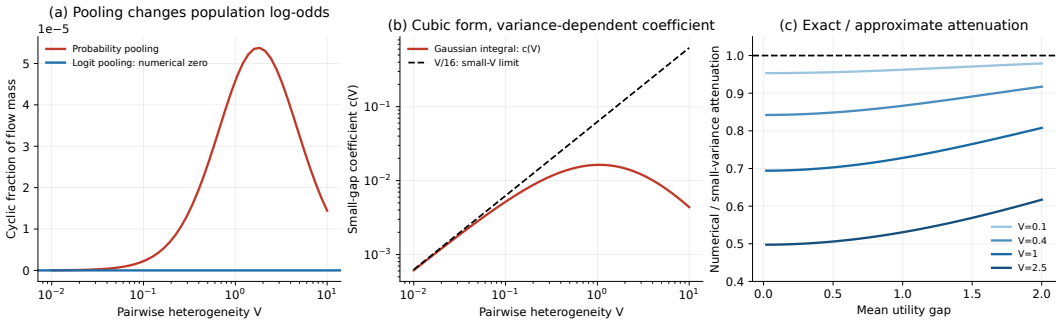


Figure 1: **Population pooling, with its approximation limits explicit.** (a) A fixed ten-item utility vector is held constant as Gaussian heterogeneity changes. Probability pooling has a small nonzero Hodge residual; true-logit pooling is a gradient to numerical precision. (b) The fixed-variance cubic coefficient is obtained from derivatives under the integral, with the small-variance reference shown separately. (c) Ratio of exact log-odds attenuation to its small-variance approximation; one means agreement. Reducing the gap does not remove the approximation error at fixed V. Darker blue indicates larger V, using the same variance-to-color mapping as Figure 2c. These are population calculations, not finite-label estimates.

Log-odds nonadditivity is not a majority cycle. For centered symmetric noise, $p(d) = \mathbb{E}\sigma(d + \varepsilon)$ is increasing and $p(0) = 1/2$. Thus $\operatorname{sign}(p(d) - 1/2) = \operatorname{sign}(d)$. Gaussian heterogeneity around a common ordering therefore preserves its majority direction even while the log-odds become nonadditive. This does not extend to every population mixture. For example, equally mixing utility vectors $(8, 0, -8)$, $(-8, 8, 0)$ and $(0, -8, 8)$ produces circular win probabilities approximately 0.6664 on all three edges despite each component having a scalar utility. Some population cycles are real; the task is to distinguish them from other sources.

3 Recovering mean utilities and predicting choices differ

The operator comparison suggests averaging individual utility estimates, but the prediction target determines whether this helps. We compare two estimators using identical binary training observations: one ridge Bradley–Terry fit to pooled counts, and the mean of separately fitted individual utility vectors. Both resulting predictors are gradients. Consequently, any difference in prediction is not simply a difference between a cyclic and an acyclic fitted model.

The simulation has ten items with centered mean utilities drawn from $\mathcal{N}(0, 0.8^2)$ and independent item perturbations $\mathcal{N}(0, V/2)$. Each annotator judges all 45 pairs four times. We cross six variances $V \in \{0.25, 0.5, 1, 2, 4, 8\}$ with six panel sizes $A \in \{5, 10, 20, 40, 80, 160\}$, twelve seeds, and ridge penalties 0.001 and 0.1. Appendix A.3 specifies the coupled random draws and convergence checks.

Two endpoints answer different questions. Utility RMSE compares the centered estimates with the simulated ϕ, which is known only in this synthetic experiment. Expected prediction log loss integrates the choices of a new independent annotator from the same Gaussian population by quadrature. It is an out-of-sample population risk calculation, not a measured human held-out score.

Figure 2 contrasts utility recovery (a,b) with prediction (c): pooling has lower seed-mean expected log loss in all 36 cells at each penalty. Individual averaging can recover the mean utilities better when heterogeneity and panel size are sufficiently large, but the crossover moves substantially with regularization. At $V = 4, A = 160$, the pooled-to-individual mean RMSE ratio is 1.36 at ridge 0.001 and 2.44 at ridge 0.1. Finite-data individual fits have both bias and variance; these ratios are not universal improvement factors. This comparison changes calibration and estimation error together; it does not isolate the effect of removing curl.

Calibration accounts for most of the gap in one population example. For the fixed ten-item utilities in Figure 1, we optimize one positive scale κ in $\sigma(\kappa\Delta_{ij})$ against the known marginal probabilities. Across the six variances above, this removes 99.66–99.998% of the unscaled utilities' excess log loss over those probabilities. Both scaled and unscaled logits are gradients. This oracle

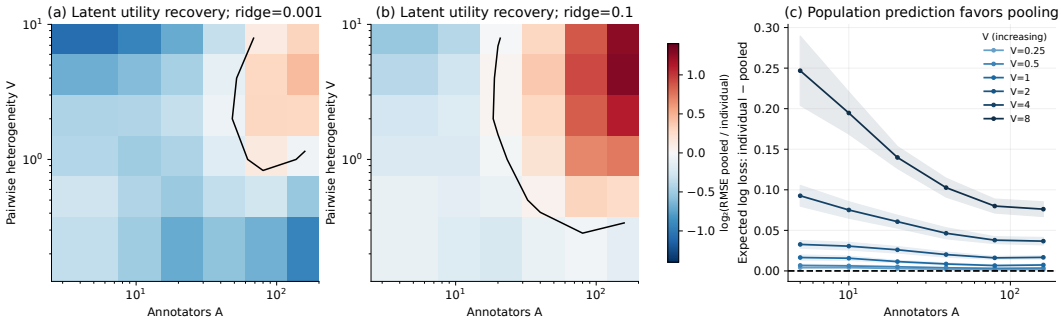


Figure 2: **Different targets favor different estimators.** (a,b) Red cells mean averaging individual fits better recovers mean latent utilities; blue favors a pooled fit. Black contours interpolate equality on the sampled grid, not a theoretical phase boundary. The change across ridge penalties shows estimator dependence. Each cell uses the ratio of mean RMSEs across twelve seeds. (c) Darker blue indicates larger V. Positive expected-log-loss differences favor pooling; lines are seed means with ± 1 standard-error bands, at ridge 0.001. Both predictors are scalar-utility models.

calculation isolates a calibration effect; it is not a fitted held-out baseline or a general bound on pooling distortion (Appendix A.4).

Why reward scale matters. For reward r, reference policy π_0, global scale $\kappa > 0$, and penalty $\beta > 0$, maximizing $\mathbb{E}_\pi[\kappa r] - \beta\,\mathrm{KL}(\pi\|\pi_0)$ is equivalent to maximizing $\mathbb{E}_\pi[r] - (\beta/\kappa)\mathrm{KL}(\pi\|\pi_0)$. The population example motivates scale calibration as a baseline; small preference-loss error does not establish equivalent learned policies. Observation dependence requires a separate audit.

4 Identical pairwise probabilities can produce different noise

Consider three equally valued items and n observations per pair. Protocol I draws an independent Bernoulli$(1/2)$ observation on every edge on every repetition. Protocol S draws one uniform random ranking on each repetition and extracts its three pairwise comparisons. Drawing independent equal-location Gumbel item utilities produces precisely this uniform ranking. Both protocols have binomial$(n, 1/2)$ marginal counts on every edge and the same population Bradley–Terry probabilities.

Orient the edges as $(a, b), (b, c), (c, a)$. A single ranking produces a binary vector X. Each coordinate has variance $1/4$, but each pair of oriented coordinates has covariance $1/6 - 1/4 = -1/12$. Consequently,

$$\mathrm{Var}\left(\sum_e X_e\right) = \begin{cases} 3/4 & \text{independent edges,} \\ 1/4 & \text{shared ranking.} \end{cases} \tag{5}$$

Use the finite estimator $\hat{y}_e = \log[(w_e + 1/2)/(n - w_e + 1/2)]$ and $K = \sum_e \hat{y}_e$. The derivative of the logit at $1/2$ is four, so the delta method gives

$$n\,\mathrm{Var}(K) \longrightarrow 12 \quad \text{under I,} \qquad n\,\mathrm{Var}(K) \longrightarrow 4 \quad \text{under S.} \tag{6}$$

Thus graph, counts, and marginal probabilities do not determine the noise in estimated curl. At $n = 1$ the variance ratio is also exactly $1/3$; at intermediate counts the variance can differ from the asymptotic prediction. Figure 3 verifies the limiting behavior with 20,000 simulated triangles per condition and count.

The same point has a graph formulation. For a fixed observed graph with incidence matrix B and edge weights W, let $P = B(B^\top W B)^+ B^\top W$ be the gradient projection and $A = I - P$. If an estimated edge flow y has mean μ and covariance C, then

$$\mathbb{E}[(Ay)^\top W(Ay)] = (A\mu)^\top W(A\mu) + \mathrm{tr}(A^\top W A C). \tag{7}$$

This standard quadratic expectation identity exposes the role of cross-edge covariance. Moreover, smoothing means $\mathbb{E}[y]$ can differ from the population log-odds. Appendix A.7 derives the identity and distinguishes energy from a normalized residual fraction. Missing edges must have zero weight; inserting unobserved pairs as ties changes the problem.

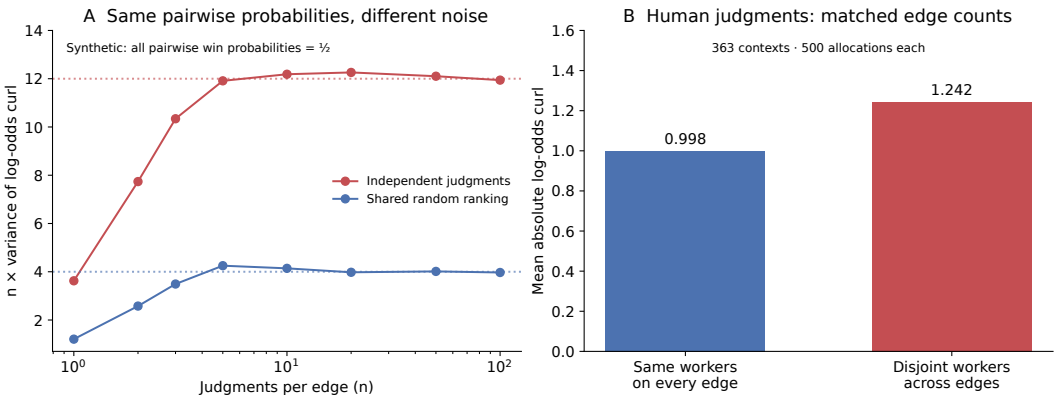


Figure 3: **Observation dependence changes measured cycles.** Left: simulated protocols have identical pairwise probabilities and counts. Dotted lines are asymptotic predictions, not fitted curves. Right: retrospective reallocation of existing judgments on 363 contexts, averaged over 500 allocations per context. Shared and disjoint constructions use the same number of votes per edge. The bars are allocation averages, not independent experimental replications.

5 A CONDITIONAL DIAGNOSTIC AND ITS LIMITS

A fitted-null subtraction can have a nonzero expectation under the very null it is intended to represent. We therefore use classical exponential-family conditioning to remove fitted utility parameters from a triangle-level diagnostic.

For this calculation orient counts as $w = (w_{ab}, w_{bc}, w_{ac})$, with fixed totals $n = (n_1, n_2, n_3)$. Under independent-binomial Bradley–Terry comparisons, conditioning on each item's total wins leaves only states

$$w(k) = w + k(1, 1, -1), \qquad \Pr(w(k) \mid \text{item wins}, n) \propto \prod_{e=1}^{3} \binom{n_e}{w_e(k)}, \tag{8}$$

where $\max(-w_1, -w_2, w_3 - n_3) \le k \le \min(n_1 - w_1, n_2 - w_2, w_3)$. Utility-dependent likelihood terms are constant on this set (Appendix A.6). We enumerate the distribution exactly and use absolute smoothed log-odds curl as the statistic. For aggregate diagnostics we sum across one selected triangle per context and draw 9,999 conditional Monte Carlo replicates, with inclusive tails and a plus-one correction.

Selecting one triangle prevents overlapping triangles from being counted as independent within a context. It does not make repeated judgments by a worker independent, nor remove dependence between contexts. Our diagnostic explicitly tests an *independent-binomial* model. Its lower tail must not be read as proof of transitivity. Thresholds and analyses were developed after examining the data and are exploratory.

There is also a precise sparsity limitation. With one observation on every edge, conditioning fixes a transitive triple completely; for a cyclic triple it allows only the two cycle orientations. Those orientations have identical absolute curl. The conditional absolute-curl statistic therefore has no power at this design. Figure 4b shows how power changes with counts; this limitation concerns this statistic and conditioning choice. In larger graphs, triangles also miss harmonic residuals on chordless longer cycles.

5.1 ACYCLICITY CAN ALSO BE IMPOSED BY THE ENCODING

Suppose every participant chooses one winner from the same fixed menu, and the analyst expands each choice into a win against every other item. Let c_i be the number of times item i is selected. The induced edge has c_i wins against c_j losses, so its smoothed log-odds are

$$\hat{y}_{ij} = \log \frac{c_i + 1/2}{c_j + 1/2} = s_i - s_j, \qquad s_i = \log(c_i + 1/2). \tag{9}$$

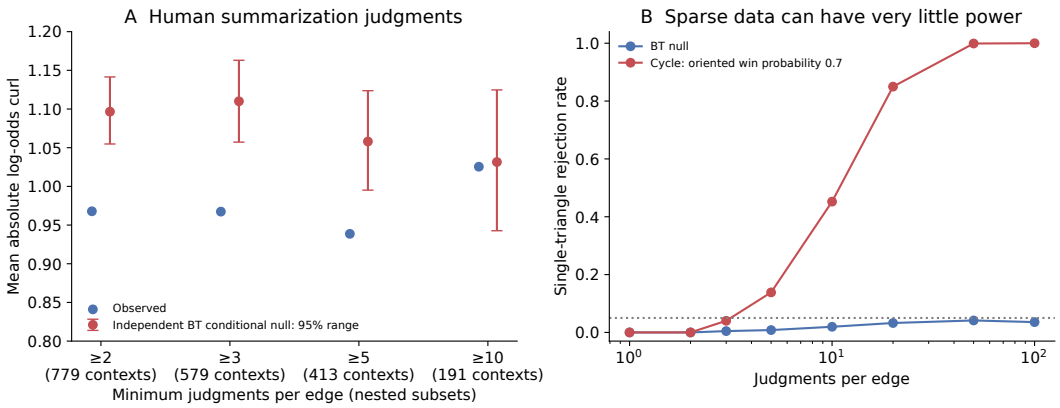


Figure 4: **Calibration and power qualify empirical conclusions.** (a) Observed absolute curl and the conditional independent-BT null on nested summarization subsets. Bars show 95% conditional null ranges, not uncertainty intervals for a population effect. (b) Rejection rates at a nominal 5% upper-tail threshold for independent BT and a cyclic alternative with oriented win probability 0.7; 20,000 simulated triangles per point. At one observation per edge, this conditioned statistic has no power.

Table 1: Primary conditional diagnostic. Count thresholds define nested subsets. The null assumes independent judgments, an assumption questioned by the worker audit. Informative contexts have nonconstant conditional absolute curl. Monte Carlo lower-tail probabilities are exploratory model checks.

Min. count	Contexts	Informative	Observed mean	Conditional mean	Lower-tail p
2	779	384	0.968	1.097	0.0001
3	579	349	0.967	1.110	0.0001
5	413	287	0.939	1.058	0.0001
10	191	154	1.025	1.031	0.4563

Every observable cycle is zero by construction. This gives a useful lesson from best-of-four feedback: an acyclic pairwise expansion cannot independently validate human transitivity. The identity requires identical menus and winner-versus-all expansion. It does not apply to pooling full ranked ballots, variable menus, or arbitrary edge-specific weights; those can retain or generate majority cycles. The structural identity is established here without treating a historical corpus's zero-curl result as an additional independent statistical test.

6 A REPRODUCIBLE AUDIT OF HUMAN COMPARISONS

The local release of Stiennon et al. (2020)'s comparison data contains 178,944 records. We preserve source files and SHA-256 hashes. Twelve rows compare identical stripped texts, 33 are exact duplicate records, and 27 post IDs have multiple context-metadata hashes. Our primary construction deduplicates records, excludes identical responses, and chooses the lexicographically first context hash for each post. It retains 178,531 rows. Sensitivities retain all records, use the earlier post-ID grouping, or retain one vote per worker per pair.

We construct response nodes from SHA-256 hashes of stripped text. For each context, we select the triangle maximizing its minimum edge count, break ties by total count, then by hashed node IDs. Selection uses no outcomes. There are 12,478 contexts with at least one triangle, but only 415 selected triangles have a nonconstant conditional absolute-curl statistic. At least two observations per edge leave 779 contexts; every one has a worker who judges multiple selected edges.

Table 1 and Figure 4a report the conditional comparison. The modest-count deficit survives deduplication, alternative grouping, and one vote per worker per pair. It appears in both local train and validation files. It is absent in the subset with at least ten observations per edge; these subsets differ

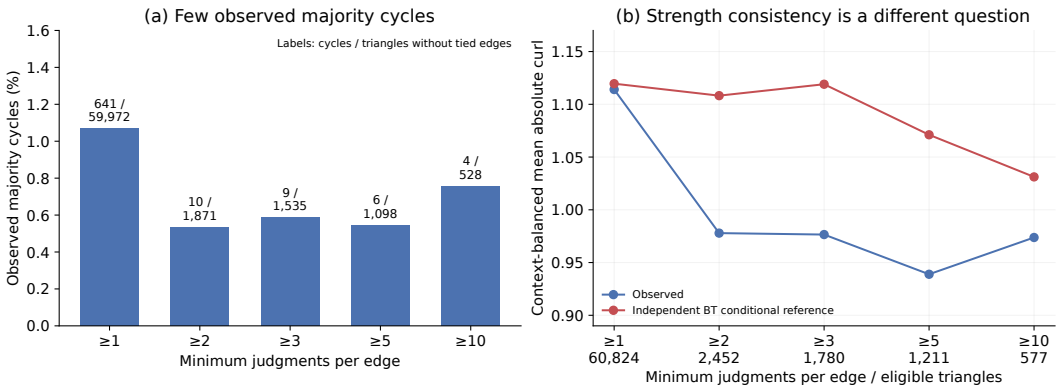


Figure 5: **Real-data triangle census separates direction from strength.** (a) Observed majority-cycle frequencies exclude triangles with any tied edge; labels give numerators and denominators. (b) Absolute log-odds curl is averaged over eligible triangles within each context, then equally over contexts. The conditional reference assumes independent BT judgments. Thresholds change the included contexts and triangles, so connecting lines are descriptive, not learning curves. No independent-triangle confidence intervals are assigned to overlapping observations.

in composition, so this is not evidence of a causal effect of increasing sample size. Retaining only the first record per worker per context leaves just two contexts with triangles and none with two observations per edge. This shows why simply deleting repeated participation cannot recover a useful independent-edge audit from these data.

Can these counts reveal population curl? At the 779 observed count vectors, independent simulations of the unequal-disagreement population give rejection rates of 8.0% at curl 0.172 and 29.5% at 0.538. At curl 0.782, detection falls from 89.5% when all triangles are affected to 20.5% when each is affected with probability one quarter. Appendix A.8 reports methods and simulation uncertainty. Detectability depends on prevalence and marginal probabilities as well as curl magnitude; these simulations neither calibrate the dependent corpus nor bound its latent cyclicity.

6.1 What all observed triangles add

Selecting one triangle per context is useful for the diagnostic, but discards a descriptive view of the graph. We therefore separately enumerate all 60,824 triangles in the primary construction. Of the 59,972 triangles with no tied edge majority, 641 (1.07%) have circular majority directions. Requiring at least two votes per edge leaves 2,452 triangles, of which 1,871 are decisive and ten (0.53%) are cyclic. Figure 5 reports ties through the denominator rather than silently assigning them a direction. These are overlapping observed triangles, not independent samples of population cycles.

The census also shows that curl summaries depend on triangle selection and weighting (Appendix A.9). Majority directions can look orderly while consistency of preference strengths remains a separate question. Few observed cycles cannot establish few population cycles when observations are sparse and dependent.

6.2 Matched-count panel reconstruction

We retain the first observed vote per worker per selected pair. A context qualifies if at least six workers have observations on all three selected edges. For m qualifying workers, set $k = \lfloor m/3 \rfloor$ and randomly permute the worker list. The shared construction takes the first k workers on all edges. The disjoint construction takes the first k on edge one, the next k on edge two, and the next k on edge three. Both use exactly k votes per edge from the same original pool. They intentionally differ in the number of distinct contributing workers.

Across 363 contexts, k ranges from two to seven. Figure 3b summarizes the effect; Figure 6a shows its variation across contexts. Averaging 500 allocations per context gives absolute curl 0.9984 under shared workers and 1.2421 under disjoint workers, a difference of 0.2437 (24.4%). The mean

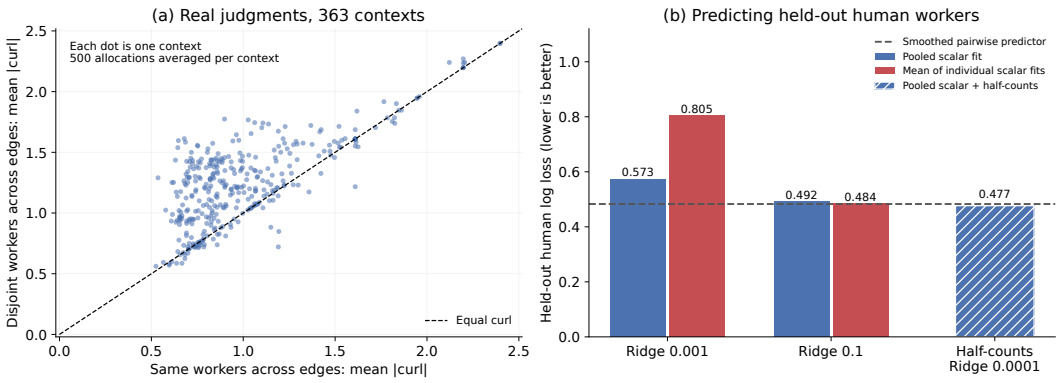


Figure 6: **Real human data: effects across contexts and held-out prediction.** (a) Each point is one of 363 summarization contexts, using the same recorded judgments and equal votes per edge in the shared/disjoint constructions. (b) Mean log loss on actual held-out human workers, averaged equally over contexts and 100 splits per context. The original ridge fits and the added half-count BT sensitivity are shown together; none is tuned on test labels. The dashed baseline estimates each pair separately with the same half-count correction used in the rightmost bar. Repeated splits reuse observations; the bars are not independent trial replications.

observed-minus-conditional-null statistic changes from -0.1503 to $+0.0235$. The 2.5th–97.5th percentile spread of the corpus-mean difference across allocations is $[0.1230, 0.3582]$. This is *allocation variability*, not a population confidence interval. Original judgments are reused, and the construction does not randomize the original elicitation protocol. Nevertheless, it directly demonstrates sensitivity of the statistic to which workers' votes are assembled together.

6.3 Prediction on held-out human workers

We test whether the synthetic prediction ordering transfers to actual human judgments. On the same 363 contexts with at least six complete workers, we randomly hold out $\lfloor m/3 \rfloor$ workers and train on the remaining workers, repeating this 100 times per context. Training and test workers are disjoint *within a context*; workers may recur in other contexts. The comparison is exploratory, with no hyperparameter selected on test judgments. Each worker contributes their first observed binary choice on each selected edge. Per-worker fits use only that worker's records, and only training workers' fitted utilities enter the averaged predictor.

Figure 6b compares held-out prediction. With ridge 0.001, pooled and individual-averaged scalar fits have mean held-out log losses 0.5730 and 0.8048, respectively. With ridge 0.1 they have losses 0.4918 and 0.4837, a small reversal. A pairwise predictor $(w_{ij} + 1/2)/(n_{ij} + 1)$, which allows nonadditive logits, has loss 0.4829; the matched-smoothing comparison below qualifies this gap. All fits pass the same 10^{-5} gradient tolerance used in the synthetic experiment. The average individual-minus-pooled difference across random splits has a 95% allocation spread of $[0.1986, 0.2668]$ at ridge 0.001 and $[-0.0170, -0.0013]$ at ridge 0.1. These are allocation spreads on the observed corpus, not population confidence intervals.

Matching the half-count smoothing reverses the small baseline gap. The original scalar fits use a parameter penalty, whereas the pairwise baseline adds half a win and half a loss per training pair. Applying these same pseudocounts to pooled BT gives loss 0.4769 at ridge 0.0001: BT minus pairwise is -0.0060, with allocation spread $[-0.0096, -0.0022]$. Ridge 0.01 gives loss 0.4768 and the same rounded difference (Appendix A.5). All smoothing uses training counts only. Matching pseudocounts does not equate every aspect of regularization; the reversal shows why predictive differences alone cannot identify a need for nonadditive probabilities.

Neither the synthetic ordering nor near-equal human losses establish universal superiority or statistical equivalence. Sparse fits, smoothing, regularization, and model constraints all affect prediction. This small-menu experiment measures recorded choices, not downstream policy quality or unobserved mean human utilities.

Secondary corpus. Content-matched MT-Bench comparisons (Zheng et al., 2023) provide limited additional evidence: selected GPT-4 triangles have no conditional absolute-curl information. Appendix A.10 reports this audit; apparent consistency in such sparse data cannot establish a tight bound on population cyclicity.

7 Relationship to prior work

Population aggregation and hidden preferences. The marginal-versus-conditional distinction in correlated binary models predates preference learning (Zeger et al., 1988; Neuhaus et al., 1991). In modern RLHF, Siththaranjan et al. (2024) study hidden context and propose learning utility distributions, while Zhang et al. (2024) document sources of annotator disagreement and limitations of standard reward models in representing that disagreement. Our calculation makes a local log-odds distortion explicit and separates mean-utility recovery from choice prediction. Neither target is automatically a normative definition of the best response.

Ranking geometry and observation design. Hodge ranking decomposes pairwise data into score-compatible and residual components (Jiang et al., 2011); applications connect graph sampling to subjective-quality assessment (Xu et al., 2012), and mixed-effects HodgeRank models individual variation (Xu et al., 2016). Dependence in paired comparisons is also established (Dittrich et al., 2006). We show its diagnostic consequence: identical edge probabilities and counts can produce different curl noise, with corresponding sensitivity in matched-count human reconstructions.

Representation and policy learning. Reward-learning theory studies when comparisons support policy learning (Zhu et al., 2023). Direct preference optimization uses a reward–policy relationship (Rafailov et al., 2023), while Nash learning permits general pairwise preferences (Munos et al., 2023). Recent work examines representation–rationalizability tradeoffs (Dong et al., 2026), separates transitive and cyclic components (Huang et al., 2026), and studies majority consistency in RLHF (Xiao et al., 2025). Our complementary question is whether the discrepancy motivating a richer representation survives an appropriate observation model and helps prediction.

8 Implications and limitations

Before choosing a richer preference model, specify the target population, retain worker and menu identities, distinguish majority cycles from strength nonadditivity, and check diagnostic power at the available counts. Held-out prediction then answers a different question: whether the extra structure helps.

Average fit can miss localized disagreement. The prevalence experiment (Table 4) shows why large log-odds cycles confined to a subset of contexts can be hard to detect with a corpus-wide statistic. This motivates prespecified analyses of response categories relevant to deployment, with enough repeated judgments to assess power within each category. Such analyses should test where a scalar model succeeds or fails, rather than treating low aggregate curl or good average prediction as evidence of adequacy everywhere.

A prospective study should randomize shared versus independent participants across edges at matched counts and fixed responses, assess power against known cyclic alternatives, and compare scalar and general pairwise predictors. Policy-level evaluation could then test whether better preference prediction improves downstream outcomes. Retrospective reallocation motivates this sequence but does not substitute for it.

The wider connection is to hierarchy in networks, including animal dominance (Shev et al., 2012): cycles can describe populations, encounters, or their summarization. The graph covariance identity provides a common language, but cross-domain universality requires predicting the right quantities in independently collected domains.

Limitations. The empirical evidence comes from one substantial human corpus, retrospective worker allocation, and small response menus. Generalization requires an adequately repeated second corpus and prospective validation. The results do not establish universal human transitivity.

REPRODUCIBILITY STATEMENT

Source and derived hashes, selected triangle counts, simulation seeds, per-context resampling summaries, and plotting scripts document the analyses. Exact enumeration checks validate conditional probabilities against an independent binomial likelihood and verify finite-sample type-I control in small designs. Graph tests check weighted orthogonality, missing-edge handling, and a longer cycle without triangles. Appendix derivations state the limits and assumptions used by the numerical calculations.

AI ASSISTANCE

LLMs assisted literature retrieval, mathematical derivations and proofs, experiment design and execution (including synthetic simulations and real-data preprocessing), and editing.

REFERENCES

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.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30, 2017.

Regina Dittrich, Brian Francis, Reinhold Hatzinger, and Walter Katzenbeisser. Modelling dependency in multivariate paired comparisons: A log-linear approach. *Mathematical Social Sciences*, 52(2):197–209, 2006.

Jing Dong, Yaoliang Yu, and Pascal Pourpart. The representation-rationalizability tradeoff in reward learning. *arXiv preprint arXiv:2606.00291*, 2026.

Yucong Huang, Xiucheng Li, Kaiqi Zhao, and Jing Li. Transitivity meets cyclicity: Explicit preference decomposition for dynamic large language model alignment. *arXiv preprint arXiv:2605.17342*, 2026.

Xiaoye Jiang, Lek-Heng Lim, Yuan Yao, and Yinyu Ye. Statistical ranking and combinatorial hodge theory. *Mathematical Programming*, 127(1):203–244, 2011.

Rémi Munos, Michal Valko, Daniele Calandriello, Mohammad Gheshlaghi Azar, Mark Rowland, Zhaohan Daniel Guo, Yunhao Tang, Matthieu Geist, Thomas Mesnard, Andrea Michi, et al. Nash learning from human feedback. *arXiv preprint arXiv:2312.00886*, 2023.

John M Neuhaus, John D Kalbfleisch, and Walter W Hauck. A comparison of cluster-specific and population-averaged approaches for analyzing correlated binary data. *International Statistical Review/Revue Internationale de Statistique*, pp. 25–35, 1991.

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. *Advances in neural information processing systems*, 36:53728–53741, 2023.

Aaron Shev, Fushing Hsieh, Brianne Beisner, and Brenda McCowan. Using markov chain monte carlo (mcmc) to visualize and test the linearity assumption of the bradley–terry class of models. *Animal Behaviour*, 84(6):1523–1531, 2012.

Rahul Singh and Ori Davidov. Testing for lack of fit in paired comparison data. *arXiv preprint arXiv:2604.00426*, 2026.

Anand Siththaranjan, Cassidy Laidlaw, and Dylan Hadfield-Menell. Distributional preference learning: Understanding and accounting for hidden context in rlhf. In *International Conference on Learning Representations*, volume 2024, pp. 27448–27471, 2024.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. Learning to summarize with human feedback. *Advances in neural information processing systems*, 33:3008–3021, 2020.

Jiancong Xiao, Zhekun Shi, Kaizhao Liu, Qi Long, and Weijie J Su. Theoretical tensions in rlhf: Reconciling empirical success with inconsistencies in social choice theory. *arXiv preprint arXiv:2506.12350*, 2025.

Qianqian Xu, Qingming Huang, Tingting Jiang, Bowei Yan, Weisi Lin, and Yuan Yao. Hodgerank on random graphs for subjective video quality assessment. *IEEE Transactions on Multimedia*, 14 (3):844–857, 2012.

Qianqian Xu, Jiechao Xiong, Xiaochun Cao, and Yuan Yao. Parsimonious mixed-effects hodgerank for crowdsourced preference aggregation. In *Proceedings of the 24th ACM international conference on Multimedia*, pp. 841–850, 2016.

Scott L Zeger, Kung-Yee Liang, and Paul S Albert. Models for longitudinal data: a generalized estimating equation approach. *Biometrics*, pp. 1049–1060, 1988.

Michael JQ Zhang, Zhilin Wang, Jena D Hwang, Yi Dong, Olivier Delalleau, Yejin Choi, Eunsol Choi, Xiang Ren, and Valentina Pyatkin. Diverging preferences: When do annotators disagree and do models know? *arXiv preprint arXiv:2410.14632*, 2024.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623, 2023.

Banghua Zhu, Michael Jordan, and Jiantao Jiao. Principled reinforcement learning with human feedback from pairwise or k-wise comparisons. In *International Conference on Machine Learning*, pp. 43037–43067. PMLR, 2023.

A Derivations and interpretation details

A.1 The Gaussian small-gap coefficient

Let $p_V(d) = \mathbb{E}\sigma(d + \sqrt{V}Z)$ and $f_V(d) = \operatorname{logit} p_V(d)$. Symmetry gives $p_V(-d) = 1 - p_V(d)$, so f_V is odd. Logistic derivatives are bounded, allowing differentiation under the Gaussian integral. With $u = p_V'(0)$ and $t = p_V^{(3)}(0)$,

$$p_V(d) = \tfrac{1}{2} + ud + \tfrac{1}{6}td^3 + O(d^5). \tag{10}$$

Using $\operatorname{logit}(1/2 + q) = 4q + (16/3)q^3 + O(q^5)$ yields

$$f_V(d) = 4ud + \left(\tfrac{2}{3}t + \tfrac{16}{3}u^3\right) d^3 + O(d^5). \tag{11}$$

For $x + y + z = 0$, the linear terms cancel and $x^3 + y^3 + z^3 = 3xyz$, giving $c(V) = 2t + 16u^3$. The remainder is at fixed V as the gaps shrink. Expanding instead in small V gives $u = 1/4 - V/16 + O(V^2)$ and $t = -1/8 + V/8 + O(V^2)$, hence $c(V) = V/16 + O(V^2)$. These are different limits. On the displayed grid $V \in [0.01, 10]$, the coefficient peaks near 0.01636, a grid maximum rather than a global bound. The 161- and 321-node quadratures differ by at most 2.6×10^{-8} on that grid. Differentiating under the integral avoids subtracting nearly equal numbers to recover a tiny cubic term. For numerical evaluation, with $q = \sigma(\sqrt{V}Z)$, the integrated derivatives are $q(1 - q)$ and $q(1 - q)(1 - 6q + 6q^2)$.

A.2 Unequal disagreement: a linear contribution

The expansion above also gives $f_v(d) = a(v)d + O(d^3)$, where $a(v) = 4u(v)$. Summing around a triangle proves Equation 4 at fixed edge variances. Equal coefficients cancel the linear term, but unequal coefficients do not generally cancel. Since $a(v) = 1 - v/4 + O(v^2)$, a further small-variance expansion yields the leading term $-\frac{1}{4}\sum_e v_e\Delta_e$. This last approximation requires small variances as well as small gaps.

For mean utilities $(g, -g, 0)$ and independent item perturbations with variances $(0, 0, S)$, the circular gaps are $(2g, -g, -g)$ and their variances are $(0, S, S)$. Since $f_0(d) = d$ and f_S is odd,

$$K_{\text{unequal}} = 2g - 2f_S(g) = 2[1 - a(S)]g + O(g^3). \tag{12}$$

For $S > 0$, $a(S) < 1$ because $\sigma'(x) \le 1/4$, with equality only at zero. The linear coefficient is therefore positive. Giving each item variance $S/3$ instead preserves total item variance, makes every edge variance $2S/3$, and yields

$$K_{\text{equal}} = f_{2S/3}(2g) - 2f_{2S/3}(g) = 2c(2S/3)g^3 + O(g^5). \tag{13}$$

The comparison fixes the mean utilities and total item variance; it does not assert equivalence of all aspects of the populations. Figure 7a contrasts these orders for $S = 1$. Symmetric Gaussian perturbations still preserve majority directions in both examples.

A.3 Synthetic estimator comparison

The simulation has ten items, $\bar{\phi}_i \sim \mathcal{N}(0, 0.8^2)$ centered to zero, and individual perturbations $\varepsilon_i^a \sim \mathcal{N}(0, V/2)$. Every annotator judges all 45 pairs four times, independently conditional on their utilities. We cross six variances $V \in \{0.25, 0.5, 1, 2, 4, 8\}$ with six panel sizes $A \in \{5, 10, 20, 40, 80, 160\}$ and twelve seeds. Utility vectors and random-number streams are shared across conditions, with nested annotator sets within each variance. We repeat the grid with ridge penalties 0.001 and 0.1. Batched Newton optimization uses backtracking; every fit terminates with maximum absolute gradient below 10^{-5}.

A.4 A population calibration decomposition

For each $V \in \{0.25, 0.5, 1, 2, 4, 8\}$, let $p_{ij} = \mathbb{E}\sigma(\Delta_{ij} + \sqrt{V}Z)$ for the ten-item vector in Figure 1. Averaging over its 45 pairs, define

$$L(\kappa) = \frac{1}{45}\sum_{i<j}\left[\log(1 + \exp(\kappa\Delta_{ij})) - p_{ij}\kappa\Delta_{ij}\right]. \tag{14}$$

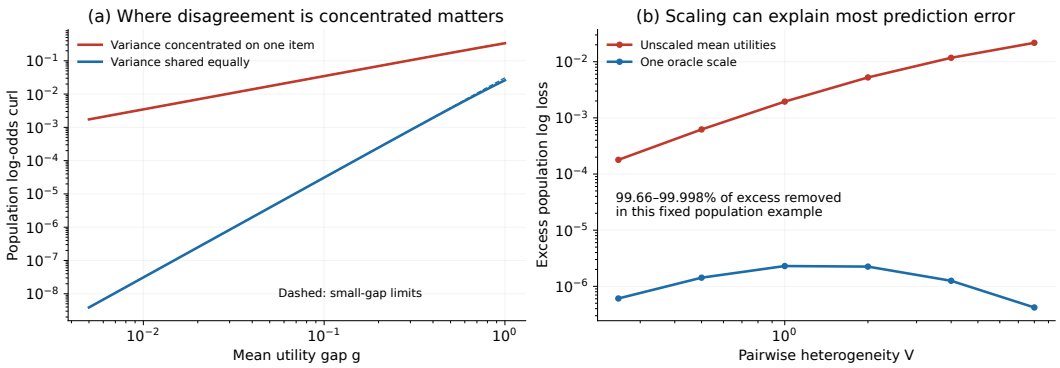


Figure 7: **Where pooling changes strengths, and where scaling suffices.** (a) Mean utilities $(g, -g, 0)$ with total item variance one. Concentrating variance on the third item gives a linear leading curl; distributing it equally gives a cubic leading curl. Dashed lines show the small-gap limits; the red limit nearly coincides with the exact curve at this scale. (b) For the fixed ten-item population of Figure 1, one oracle scale removes most excess log loss relative to true marginal probabilities. Both predictions have additive logits; the remaining excess is nonzero. The scale uses known population probabilities, not held-out labels.

Minimize this population loss over $\kappa \in [0.01, 2]$; all optima are interior, decreasing from 0.9488 to 0.4833 as V increases. Let $L_\star$ be the entropy loss of the true marginal probabilities. The removed fraction is $[L(1) - L(\kappa)]/[L(1) - L_\star]$, not a ratio of Hodge norms. It ranges from 0.99660 to 0.99998; remaining excess loss ranges from 4.2×10^{-7} to 2.3×10^{-6}. Figure 7b illustrates this example without attributing all pooling effects to temperature in general. We use 321-node Gaussian quadrature and check the unequal-disagreement curves against 161 nodes.

A.5 Matched-smoothing prediction sensitivity

We reuse exactly the 363 contexts, worker ordering, seed, and 100 within-context splits from Section 6.3. If a training pair has w wins out of n votes, the saturated estimate is $(w + 1/2)/(n + 1)$. The matched BT objective is

$$\sum_{i<j} \left[(n_{ij} + 1)\log(1 + e^{\theta_i - \theta_j}) - (w_{ij} + 1/2)(\theta_i - \theta_j)\right] + \frac{\lambda}{2}\|\theta\|^2. \tag{15}$$

Half-counts enter training only; test losses use the original binary judgments. Both penalties in Table 2 are reported without test-label selection. The original pairwise loss is reproduced to numerical precision, and all added BT fits satisfy the 10^{-5} gradient criterion. Equal pairwise pseudocounts leave different constraints and effective model complexity, so this sensitivity does not prove population transitivity.

Table 2: Same half-count observations, different model constraints. Differences are BT minus pairwise loss; negative favors BT. Ranges are 2.5th–97.5th percentiles across corpus-mean allocations, not population confidence intervals.

BT ridge	BT loss	BT minus pairwise	Allocation spread
0.0001	0.4769	−0.0060	$[-0.0096, -0.0022]$
0.01	0.4768	−0.0060	$[-0.0097, -0.0023]$

A.6 Why conditioning removes utilities

Under independent BT edges, the likelihood of counts is proportional to

$$\left[\prod_e \binom{n_e}{w_e}\right] \frac{\exp\{\sum_e w_e(\theta_{i(e)} - \theta_{j(e)})\}}{\prod_e [1 + \exp(\theta_{i(e)} - \theta_{j(e)})]^{n_e}}. \tag{16}$$

With totals fixed, the denominator does not vary over the conditional states. Fixing item wins fixes the coefficient of every θ_i in the numerator, leaving only the binomial coefficients. On three edges the admissible changes are multiples of $(1, 1, -1)$ in the (ab, bc, ac) orientation. This proves the conditional mass used in Section 5; it does not extend the likelihood to dependent observations. A small lower-tail value can therefore diagnose dependence or another mismatch, not certify a gradient population.

A.7 Graph residuals include observation covariance

For deterministic A and W, write $y = \mu + \eta$ with $\mathbb{E}\eta = 0$ and $\mathbb{E}\eta\eta^\top = C$. Expanding the quadratic form cancels the cross term in expectation and gives

$$\mathbb{E}[y^\top A^\top W A y] = \mu^\top A^\top W A \mu + \operatorname{tr}(A^\top W A C). \tag{17}$$

The fixed-graph assumption is essential: selection based on outcomes can make A random and dependent on y. Even for fixed A, replacing C by a diagonal matrix discards covariance terms. A normalized residual fraction introduces an additional random denominator, so this identity for residual energy alone is not an identity for the expectation of that fraction.

A.8 Detection at the observed comparison counts

We evaluate the upper-tail aggregate conditional test at level 0.05 using the selected count vectors, while replacing all outcomes by simulations. Each cell uses 200 independent simulated corpora and 999 conditional null draws per test, with inclusive tails and the plus-one correction. This is fewer conditional draws than the 9,999 used for the human audit. Tables 3 and 4 report Wilson intervals for Monte Carlo uncertainty across corpora. They are not confidence intervals for human population parameters. Counts are held fixed; neither outcomes nor human utility parameters are used to choose simulated probabilities.

Random baselines plus circulation. For each triangle, independently draw $\theta_i \sim \mathcal{N}(0, 0.8^2)$ and a sign $s \in \{-1, 1\}$. Set circular logits to $\theta_a - \theta_b + sK/3$, $\theta_b - \theta_c + sK/3$, and $\theta_c - \theta_a + sK/3$. Their curl is exactly sK. Draw independent binomial counts from the resulting probabilities, using the actual count vectors at minimum edge counts two, five, and ten. Every triangle carries the effect. Table 3 gives estimated power of about 10–33% for $K = 0.2$–0.4 and 99–100% for $K = 0.8$. The zero-curl rejection rates are 3–6%. This experiment uses seed 923 and does not simulate the Gaussian heterogeneity mechanism itself.

Direct unequal-disagreement populations. We next use mean utilities $(g, -g, 0)$ and independent item perturbations with variances $(0, 0, S)$, exactly as in Appendix A.2. Integrating the logistic choice probabilities by 321-node Gaussian quadrature gives the pairwise marginals. Independent binomial votes from these marginals correspond to drawing a fresh independent annotator for every comparison, not reusing one annotator across edges. Each context receives an independent uniformly random permutation of the three item identities, avoiding a fixed association between the polarizing response and an edge's count. We use the 779 count vectors with at least two observations per edge and seed 924.

For prevalence ρ, each context independently receives the affected population with probability ρ; otherwise it has the same mean utilities and zero heterogeneity. Table 4 reports the zero-heterogeneity control and all four alternative settings. The control's 7.0% rejection rate has a Monte Carlo interval containing 5%. Curl 0.172 is difficult to detect even at prevalence one; curl 0.782 is detected much less often when prevalence falls to one quarter. All these Gaussian populations preserve majority directions. Nonadditive strengths can therefore be hard to detect even without majority cycles or dependent observations.

These are specified-population design calculations. A fixed curl magnitude does not fix power: marginal probabilities determine which conditional states are observed, and prevalence determines how much signal is aggregated. Shared-worker dependence can change both calibration and power, so the independent simulations are not upper bounds for the human corpus. A non-rejection here cannot certify scalar population utilities there.

Table 3: Count-matched power under random scalar baselines plus circulation. Each cell is an estimated rejection rate with a 95% Wilson interval over 200 simulated corpora. Every triangle carries the stated curl magnitude. Comparisons and contexts are independent by construction.

Min. count (N)	$K = 0$	$\|K\| = 0.2$	$\|K\| = 0.4$	$\|K\| = 0.8$
2 (779)	0.04 [0.02, 0.07]	0.11 [0.07, 0.16]	0.28 [0.22, 0.35]	0.99 [0.96, 1.00]
5 (413)	0.06 [0.03, 0.10]	0.10 [0.06, 0.14]	0.33 [0.27, 0.40]	1.00 [0.98, 1.00]
10 (191)	0.03 [0.01, 0.06]	0.10 [0.07, 0.15]	0.30 [0.24, 0.37]	1.00 [0.98, 1.00]

Table 4: Direct unequal-disagreement mechanism at the 779 count vectors with at least two observations per edge. Fraction is the probability a context receives the affected population. Unaffected contexts have the same mean utilities and zero heterogeneity. Intervals describe simulation uncertainty, not uncertainty about the human population.

Gap g	Variance S	Fraction	Affected $\|K\|$	Rejection rate [95% interval]
1	0	1	0.000	0.070 [0.042, 0.114]
0.5	1	1	0.172	0.080 [0.050, 0.126]
1	2	1	0.538	0.295 [0.236, 0.362]
1	4	1	0.782	0.895 [0.845, 0.930]
1	4	0.25	0.782	0.205 [0.155, 0.266]

A.9 All-triangle weighting sensitivity

The all-triangle, context-balanced mean curl is 0.978 versus a conditional mean of 1.108 at minimum count two. At minimum count ten it is 0.974 versus 1.031, retaining a descriptive deficit where the one-triangle-per-context check is near its reference. Thus the high-count picture also depends on triangle selection and weighting. Neither view isolates a sample-size effect. The census strengthens the distinction in our opening example: majority directions can look orderly while consistency of preference strengths remains a separate modeling question. Moreover, few observed cycles cannot establish few population cycles when observations are sparse and dependent.

A.10 A secondary corpus illustrates limited information

For MT-Bench (Zheng et al., 2023), we match question, turn, and response conversation content through the evaluated turn before intersecting observed pairs. We exclude identical conversation responses (21 human and 15 GPT-4 records). Dropping ties leaves 1,221 common content-matched pairs. On that common graph, mean observed-edge Hodge residual fractions are 0.1323 for humans and 0.0877 for GPT-4. These are descriptive values with unequal counts, not causal judge effects. The selected GPT-4 triangles provide zero conditional absolute-curl information in this analysis. Thus their apparent consistency cannot support a tight upper bound on population cyclicity. Random tie breaking changes the descriptive fractions substantially and is only a sensitivity, not a preferred measurement model.