RUBRIC-RESIDUAL PREFERENCE GEOMETRY FOR RLHF SIGNAL SELECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Preference collection for RLHF can conflate stable disagreement with choices induced by the instructions given to raters. We introduce Rubric-Residual Graph Reward (RRG-Reward), a preference model that separates these two roles before using local residual agreement as a graph regularizer. The method cross-fits an arm-adjusted annotation model, builds edges only from pre-specified repeated co-labeling, and regularizes a latent preference mixture with the resulting graph. Its policy interface uses a reversal-antisymmetric, variance-normalized margin, so the same uncertainty term cannot reverse a pairwise preference when response order changes. We give matched equations for graph-free mixture and global-factor comparators, a target construction for independent binary labels, a single clustered inference estimand, and a direct human-evaluation aggregation rule. Together, these definitions make rubric-residual preference geometry a precise and testable signal-selection primitive for disagreement-aware RLHF.

1 INTRODUCTION

RLHF selects a preference signal that determines which policy updates are reinforced. Aggregate preference models such as Bradley–Terry represent this signal with a single ordering (Bradley & Terry, 1952); direct preference optimization (DPO) then turns pairwise supervision into policy updates (Rafailov et al., 2023). Both operations become delicate when a rater's choice reflects a durable tradeoff—for example safety versus helpfulness—and when it also responds to the instruction supplied for that task. Treating these sources of variation as the same object can make an annotator graph encode shared instruction exposure rather than persistent preference structure.

RRG-Reward treats the randomized rubric as a nuisance for graph construction and a separately identified intervention for annotation analysis. It first estimates an arm-adjusted label probability out of fold. The residuals from that model are compared only on pairs of annotators for which the collection design supplies sufficient shared items. A sparse positive residual-agreement graph regularizes a latent mixture but is not assigned a causal interpretation. This construction is distinct from a graph-free annotator mixture and from a global low-rank rater factor: each can receive the same arm-adjusted labels without using local residual neighborhoods.

The paper makes three technical contributions.

- It defines a residual-graph construction with a cross-fitting boundary, a minimum co-labeling rule, positive-edge selection, and a prediction rule for unseen prompts and annotators.
- It gives a reversal-antisymmetric policy score, including DPO label conversion, confidence weights, ties, and frozen decision parameters.
- It specifies matched comparators and a complete evaluation protocol spanning binary-label calibration, randomized-arm contrasts, two-way clustered uncertainty, and direct blinded human comparison.

These definitions isolate the scientific question: whether a local graph estimated after instruction adjustment adds predictive or policy value beyond models that receive the same preference information without that graph. The protocol makes that question answerable without treating instruction randomization as a proof of a graph covariance or treating a public dataset's native fields as a different kind of label.

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.

2 RUBRIC-RESIDUAL GRAPH REWARD

2.1 COLLECTION INTERFACE AND RESIDUAL GRAPH

Let $q = (x, y_1, y_2)$ be a prompt and two candidate responses, u an annotator, and $r \in \mathcal{R}$ a rubric arm. The observed binary label $A_{uqr} = 1$ means that u selects y_1. An RRG collection fixes its arms, allocation probabilities, candidate annotator pairs, and split assignment before model fitting. It uses a balanced incomplete-block allocation for every candidate graph edge: a pair (u, v) is eligible only when it has at least h shared prompts, where h is fixed before fitting. Pairs below h are not imputed and never enter the graph. This incidence condition is separate from the randomization of the rubric arm.

For each cross-fitting fold k, fit the arm-adjusted model on the complementary folds,

$$\operatorname{logit} \widehat{p}_{uqr}^{(-k)} = a_q + \eta_u + \beta_r + g(q)^\top \kappa_r, \tag{1}$$

where $g(q)$ is a pre-specified vector of prompt and response-pair features, a_q is an item effect, η_u is an annotator intercept, and (β_r, κ_r) are arm effects. The only residual used for a held-out observation is

$$e_{uqr} = A_{uqr} - \widehat{p}_{uqr}^{(-k)}. \tag{2}$$

Thus graph construction does not reuse a label to fit and evaluate its own residual probability.

Let $\mathcal{C}_{uv}$ be the pre-specified shared-prompt set for an eligible pair, and align the two residual vectors by prompt. We define

$$s_{uv} = \frac{|\mathcal{C}_{uv}|}{|\mathcal{C}_{uv}| + \lambda_n} \operatorname{corr}\left(\{e_{uqr}\}_{q \in \mathcal{C}_{uv}}, \{e_{vqr'}\}_{q \in \mathcal{C}_{uv}}\right), \qquad |\mathcal{C}_{uv}| \geq h. \tag{3}$$

The graph retains a validation-frozen fraction ρ of positive eligible similarities, sets $w_{uv} = \max\{s_{uv}, 0\}$ for retained edges, and sets all other weights to zero. Isolated annotators remain in the likelihood but have no Laplacian penalty. Equations 1–3 define a regularizer; they do not assert a universal covariance identity or a latent-cluster recovery guarantee.

2.2 LATENT REWARD MIXTURE

Let $d_c(q) = H_c(x, y_1) - H_c(x, y_2)$ be the reward margin of latent component $c \in \{1, \ldots, C\}$, and let $z_u \in \Delta^{C-1}$ be an annotator's membership vector. Conditional on component c, the arm-aware annotation likelihood is

$$P(A_{uqr} = 1 \mid q, r, c) = \operatorname{sigmoid}\big(d_c(q) + \beta_r + g(q)^\top \kappa_r\big). \tag{4}$$

With Z stacking the z_u and L_G the Laplacian of the graph above, RRG minimizes

$$\mathcal{L}_{\mathrm{RRG}} = -\sum_{(u,q,r)} \log\left[\sum_{c=1}^{C} z_{uc} P(A_{uqr} \mid q, r, c)\right] + \lambda_G \operatorname{tr}(Z^\top L_G Z) + \lambda_{\mathrm{KL}} \sum_u \mathrm{KL}(z_u \| \pi). \tag{5}$$

The deployment prediction marginalizes the rater membership,

$$p_{\mathrm{RRG}}(y_1 \succ y_2 \mid q) = \sum_{c=1}^{C} \pi_c \operatorname{sigmoid}(d_c(q)). \tag{6}$$

Consequently a new prompt is scored by H_c, and a new annotator is represented by π rather than by an inferred graph edge. Rubric effects appear in the annotation likelihood but not in the neutral deployment prediction.

2.3 REVERSAL-ANTISYMMETRIC POLICY SUPERVISION

The raw form mean margin minus variance is not suitable as a pairwise policy score: swapping y_1 and y_2 reverses the mean but not the variance. RRG instead defines

$$m(q) = \sum_c \pi_c d_c(q), \qquad v(q) = \sum_c \pi_c \big(d_c(q) - m(q)\big)^2, \tag{7}$$

$$a(q) = \frac{m(q)}{\sqrt{1 + \lambda_{\mathrm{var}} v(q)}}, \qquad p(q) = \operatorname{sigmoid}(a(q)), \tag{8}$$

where $\lambda_{\text{var}} \geq 0$ is selected on a development split and frozen before policy training. Since $m(q^{\text{rev}}) = -m(q)$ and $v(q^{\text{rev}}) = v(q)$, $a(q^{\text{rev}}) = -a(q)$. This is the required response-order antisymmetry.

The threshold $\epsilon \geq 0$ is selected once on the development split and frozen before policy training. Form a DPO training example only when $|p(q) - 1/2| > \epsilon$. The preferred response is y_1 when $p(q) > 1/2 + \epsilon$ and y_2 otherwise; the example weight is $w(q) = 2|p(q) - 1/2|$. With y^+ and y^- so defined, the weighted objective is

$$-w(q) \log \text{sigmoid}\left(\alpha \left[\log \frac{\pi_\theta(y^+ \mid x)}{\pi_{\text{ref}}(y^+ \mid x)} - \log \frac{\pi_\theta(y^- \mid x)}{\pi_{\text{ref}}(y^- \mid x)} \right] \right), \tag{9}$$

with DPO temperature α fixed across methods. Ties are excluded rather than assigned an arbitrary direction. Every compared signal uses the same pair pool, development-selected and frozen ϵ, α, reference policy, optimization budget, and seed set; a method without variance gating sets $\lambda_{\text{var}} = 0$ in Equation 8.

3 Matched Comparators

All comparators use the same encoder family, arm features $g(q)$, train/development/test partitions, and neutral deployment target. They differ only in the structure assigned to rater heterogeneity.

AE-MoE+R. This model uses the likelihood in Equation 4 and annotator memberships z_u, but removes the graph term by setting $\lambda_G = 0$. Its prediction is Equation 6. It therefore receives the same residualized rubric information as RRG without local coupling.

Residualized Shared Mixture (RSM). RSM uses the same component functions and arm likelihood but fixes $z_u = \pi$ for all annotators. It tests whether a shared mixture explains the data without individual membership or a graph.

RLF-BT. The global-factor comparator replaces local memberships with a rank-d rater–item interaction,

$$P(A_{uqr} = 1 \mid q, r) = \text{sigmoid}\big(d_0(q) + a_u^\top B\phi(q) + \beta_r + g(q)^\top \kappa_r\big), \tag{10}$$

where $a_u \in \mathbb{R}^d$, $B\phi(q) \in \mathbb{R}^d$, and $d_0(q)$ is a global Bradley–Terry margin. Its unseen-annotator prediction sets $a_u = 0$, and its deployment prediction omits rubric terms. Thus RLF-BT models global factors rather than graph neighborhoods.

For a matched comparison, $C \in \{2, 3, 4\}$, component or factor width $d \in \{128, 256, 512\}$, and ℓ_2 penalty in $\{10^{-5}, 10^{-4}, 10^{-3}\}$ are selected only on the development split. RRG additionally selects $\lambda_G \in \{0.1, 0.3, 1.0\}$ and $\rho \in \{0.05, 0.10, 0.20, 0.40\}$ there. Hidden widths are adjusted so that each method's non-encoder trainable-parameter count is within one percent of RRG's selected count. An eligible result reports the chosen configuration, parameter count, regularization values, development criterion, and treatment of every unseen unit.

4 Evaluation Protocol

4.1 Calibration target and uncertainty

Calibration uses only a held-out panel of independently collected binary comparisons under a declared target rubric. For pair q, let $A_{qj} \in \{0, 1\}$ denote the jth independent label and m_q its label count. The descriptive target is $\bar{A}_q = m_q^{-1} \sum_j A_{qj}$. A finite-label correction for a pair-level Brier display is

$$\ell_q^{\text{db}}(p) = (p - \bar{A}_q)^2 - \frac{\bar{A}_q(1 - \bar{A}_q)}{m_q - 1}, \qquad m_q > 1, \tag{11}$$

whose expectation removes the binomial sampling term under independent labels. Primary comparison can equivalently average the label-level loss difference $(p_{\text{RRG}} - A_{qj})^2 - (p_b - A_{qj})^2$, which preserves the evaluator identity for uncertainty estimation. ECE is reported with a fixed, pre-declared bin partition $\{B_b\}_{b=1}^B$:

$$\text{ECE} = \sum_{b=1}^{B} \frac{|B_b|}{|\mathcal{Q}|} \left| \frac{1}{|B_b|} \sum_{q \in B_b} p_q - \frac{1}{|B_b|} \sum_{q \in B_b} \bar{A}_q \right|. \tag{12}$$

Rubric arms are not averaged into this target. If an arm-specific target is studied, it is a separate estimand with its own held-out label layer.

For a predictive comparison, regress the label-level paired loss difference on an intercept. Compute the CR2 covariance of that intercept separately after clustering by prompt, evaluator, and the prompt–evaluator intersection, and use the two-way inclusion–exclusion covariance $\widehat{V} = \widehat{V}_q^{\mathrm{CR2}} + \widehat{V}_j^{\mathrm{CR2}} - \widehat{V}_{qj}^{\mathrm{CR2}}$. The primary Wald test uses the conservative reference degrees of freedom $\min\{G_q, G_j\} - 1$, where G_q and G_j are the observed prompt and evaluator cluster counts; the report gives both counts and the number of nonempty intersection clusters. This is the sole primary interval and test procedure. A declared family of K primary paired comparisons receives Holm adjustment over exactly those K tests. Resampling, jackknife, and alternative clustering may be reported as sensitivity analyses but are not substituted for this estimand.

4.2 Randomized-arm and graph diagnostics

For arm r, define $D_q(r)$ as the disagreement probability between two independently selected eligible labels for prompt q under arm r. The randomized-arm estimand is the prompt-average contrast

$$\tau_r = \frac{1}{|\mathcal{Q}|} \sum_{q \in \mathcal{Q}} \left[D_q(r) - D_q(r_0) \right], \tag{13}$$

where r_0 is neutral. Its estimator first computes a within-prompt arm disagreement estimate from the pre-specified repeated labels, then averages prompt contrasts with inverse allocation weights. Prompts lacking the required arm replication are outside this estimand rather than silently pooled into a different one. Random arm assignment identifies Equation 13 under consistency, positivity, and no interference for the assigned task. It does not confer a causal interpretation on graph edges, latent components, or a policy trained from them.

Graph reporting records the allocation incidence matrix, the number of eligible and retained edges, the isolate rule, h, λ_n, ρ, and the cross-fitting folds. A graph diagnostic must state whether it uses a fitted empirical model or a simulation; simulations use their own declared data-generating process and are not presented as fitted-study evidence.

4.3 Direct human evaluation and routing

For a direct comparison between policy M and baseline b, each blinded, response-order-randomized judgment produces $Y_{qjs}^{M,b} \in \{0, 1/2, 1\}$ for loss, tie, or win of M on prompt q, evaluator j, and policy seed s. The reported effect is

$$\Delta_{M,b} = \frac{1}{S} \sum_{s=1}^{S} \frac{1}{|\mathcal{Q}|} \sum_{q \in \mathcal{Q}} \left(\frac{1}{J_{qs}} \sum_{j=1}^{J_{qs}} Y_{qjs}^{M,b} - \frac{1}{2} \right). \tag{14}$$

The comparison is direct: an effect against a best baseline is not inferred by subtracting two separate scalar-reference effects. Prompt and evaluator identifiers are retained for the same two-way CR2 inference rule, and seed-level effects are reported alongside the aggregate.

A traffic router, if evaluated, is trained only on a designated training partition with a fixed feature map and threshold selected on development data. Its test report gives the independent test partition, calibration treatment, score distribution, and propagated uncertainty. Policy value is established by direct evaluation under the target traffic distribution, not by substituting an arithmetic mix extrapolation for that evaluation.

5 Discussion and Conclusion

RRG-Reward gives disagreement-aware signal selection a well-defined object: a graph built from out-of-fold, instruction-adjusted residual agreement under an allocation design that supports the retained edges. The graph is deliberately predictive rather than causal. A graph-free mixture and a global factor model can be compared at matched supervision, capacity, optimization, and deployment boundaries, while the policy score remains valid when response order is reversed.

The accompanying protocol links the method to the quantities that matter for assessment: independent binary preference targets, an explicit finite-label correction, prompt-average arm contrasts, dependency-aware uncertainty, and direct policy comparisons. This combination enables empirical work to distinguish a gain due to local residual geometry from a gain due to labels, arms, tuning, aggregation, or a mismatched target.

AI USE STATEMENT

Generative AI was used only for manuscript writing, editing, and presentation. The authors reviewed the edits and remain responsible for the final content.

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. doi: 10.1093/biomet/39.3-4.324. URL https://doi.org/10.1093/biomet/39.3-4.324.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea Finn. Direct Preference Optimization: Your Language Model is Secretly a Reward Model. In *Advances in Neural Information Processing Systems*, volume 36, pp. 53728–53741, 2023. doi: 10.52202/075280-2338. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/a85b405ed65c6477a4fe8302b5e06ce7-Paper-Conference.pdf.