Beyond Elo: Low-Rank Trade-Off Structure in Multi-Axis LLM Preferences

Anonymous ACL submission

Abstract

LLM evaluation, alignment, and inference-time scaling routinely collapse pairwise preferences into one scalar (Elo, Bradley–Terry, reward). We find a sharp regime split: single-signal preference graphs (MT-Bench, Chatbot Arena) fit Bradley–Terry cleanly, but multi-axis preferences do not. On UltraFeedback the pooled graph sits 20σ below the BT null, and on PKU-SafeRLHF human helpful/safer labels disagree on 24% of pairs. The discarded structure is low-rank and interpretable: SVD of the per-axis BT-strength tensor recovers overall quality plus a single named trade-off axis (helpfulness $\leftrightarrow$ truthfulness on UltraFeedback, verbosity vs. quality on HelpSteer2), bootstrap-stable across split-halves and clustering models by RLHF lineage (LLaMA-2-chat vs. GPT). The structure is actionable: under strong-LLM-judge evaluation, a per-axis-BT plurality reranker nearly doubles scalar BT's Arena-Hard pass@1 ($0.20 \rightarrow 0.37$, $+17$pp), with the gain localized to axis-divergent prompts; on verifiable math, where axes agree, it ties scalar BT. The regime split predicts when scalar Elo discards usable signal and provides a simple replacement when multi-axis judgments are available.

1 Introduction

Pairwise comparisons are the workhorse signal of modern LLM development: RLHF (Ouyang et al., 2022) and DPO (Rafailov et al., 2023) train on preference pairs; best-of-N reranking and inference-time scaling pipelines (Cobbe et al., 2021; Snell et al., 2024) rely on pairwise judges; and leaderboards like Chatbot Arena (Chiang et al., 2024), MT-Bench (Zheng et al., 2024), and AlpacaEval aggregate pairs into single Elo or Bradley–Terry (BT) scores. The implicit modeling assumption throughout is a single coherent latent quality dimension on which candidates can be totally ordered.

We test this assumption across five datasets spanning four regimes: single-judge sparse (MT-Bench), aggregated single-signal at scale (Chatbot Arena), multi-axis LLM-judge (UltraFeedback), multi-axis human (HelpSteer2), and binary human safety (PKU-SafeRLHF). Our practical message is simple: *if the rating axes agree, scalar BT is fine; if the axes diverge, report the trade-off and use non-linear per-axis aggregation for reranking.* The findings break cleanly into three empirical claims.

Single coherent signal: BT suffices. On MT-Bench per-prompt 6-model tournaments and on the aggregated 17-model Chatbot Arena leaderboard, the observed HodgeRank transitivity ratio matches the BT null. Standard scalar BT fits this data well.

Multi-axis: scalar-inconsistent structure. On UltraFeedback (1.34M comparisons, four GPT-4 axes) the observed transitivity ratio is 20σ below the BT null: zero of 200 simulated graphs reach the observed value. On PKU-SafeRLHF, 24.1% of $73{,}907$ human-annotated pairs have the "more helpful" response disagreeing with the "safer" response, with a clean conditional structure across stakes (27% at moderate severity, 15% at severe). The phenomenon is not an artifact of GPT-4-as-judge; it is present in independently-annotated human preferences at large scale.

Precise low-rank shape, bootstrap-stable, separates model families. SVD of the per-axis BT-strength tensor $\Theta \in \mathbb{R}^{K \times A}$ (where Θ_{ia} is the BT log-strength of model i on axis a) yields PC1 = overall quality plus a single named PC2 trade-off axis: helpfulness $\leftrightarrow$ truthfulness on UltraFeedback (1.7% of variance), verbosity/complexity vs. quality on HelpSteer2 (14.6%). PC2 is bootstrap-stable within an evaluation distribution and clusters models by RLHF lineage: all three LLaMA-2-chat sizes in the helpfulness quadrant, both GPT models in the truthfulness quadrant.

The structure is actionable: on Arena-Hard open-ended prompts ($N = 200$, $K = 8$), a per-axis-BT plus axis-top-1 plurality reranker reaches pass@1 0.365 vs. scalar BT 0.199 ($\approx 3.3\sigma$), with a $7.2\times$ lift on factuality. On verifiable math (GSM8K, MATH 3–4) the same method ties scalar BT: multi-axis selection helps exactly where rating axes genuinely diverge. Cheap per-prompt cross-axis-disagreement statistics predict held-out ranking instability at $R^2 = 0.73$, giving a one-shot diagnostic for when to switch.

Contributions. **(1)** *Regime map.* An empirical characterization of *when* scalar BT suffices and when it fails on aggregated LLM preference data: clean fit for single-signal regimes, definitive failure for multi-axis, across both LLM and human raters. **(2)** *Named structure.* The residual is a low-rank perturbation on the per-axis BT-strength tensor Θ, with a bootstrap-stable, interpretable, family-clustering PC2 axis; cheap per-prompt cross-axis-disagreement statistics predict held-out ranking instability at $R^2 = 0.73$. **(3)** *Actionable reranker.* A regime-characterized inference-time method: per-axis BT + plurality reranking beats scalar BT by $\approx 3.3\sigma$ on Arena-Hard and ties on verifiable math, with aggregation ablations and category breakdowns. A complementary spectral active-comparison-selection method (saves $\sim$30% budget) is in Appendix A.

We use only classical machinery (BT MLE, HodgeRank, effective resistance, SVD); the contribution is the empirical finding of stable, named, low-rank structure in pairwise multi-axis preference data and its consequences for evaluation and selection. This distinguishes our work from latent-factor / IRT methods on task-accuracy benchmarks (Ilic et al., 2025; Lou et al., 2025; Liu et al., 2025), from prompt-conditional BT (Frick et al., 2025), and from multi-objective reward modeling that trains models rather than analyzing existing data (Wang et al., 2024a; Zhou et al., 2024; Munos et al., 2024).

2 Related Work

Pairwise comparison and ranking. Bradley–Terry (Bradley and Terry, 1952) and Elo (Elo, 1978) are the standard scalar-utility models for pairwise data; the BT Hessian at the MLE is a weighted graph Laplacian, central to Massey's method (Massey, 1997), Rank Centrality (Negahban et al., 2017), and Spectral MLE (Chen et al., 2019). Minimax rates in λ_2 of the comparison graph are due to Shah and Wainwright (2018).

Cycle decomposition and intransitive preferences. HodgeRank (Jiang et al., 2011) decomposes edge flows into gradient (transitive), curl, and harmonic components. Zhang et al. (2024) address intransitivity by embedding *responses* into a latent space rather than decomposing the comparison graph. We apply HodgeRank to LLM preference data, identify when the curl is statistically definitive (multi-axis), and characterize its shape via SVD of the per-axis BT-strength *model* tensor, a step beyond either prior approach.

Active comparison selection. The optimal-design literature (Karp and Kleinberg, 2007; Falahatgar et al., 2017; Heckel et al., 2018; Maystre and Grossglauser, 2017) provides effective-resistance and Fiedler-vector acquisition criteria; we apply these in LLM-judge cost regimes.

LLM-as-judge. MT-Bench (Zheng et al., 2024), Chatbot Arena (Chiang et al., 2024), AlpacaEval, and Arena-Hard (Li et al., 2024) use pairwise judges and aggregate to single Elo/BT scores. Chiang et al. (2024) report category-conditioned leaderboards via Kendall rank correlation but do not spectrally decompose the model $\times$ category matrix. Prompt-to-Leaderboard (Frick et al., 2025) trains an LLM to output prompt-conditional BT vectors on Arena, producing per-prompt leaderboards but not interpretable cross-prompt axes. UltraFeedback (Cui et al., 2024), HelpSteer2 (Wang et al., 2024b), and PKU-SafeRLHF (Ji et al., 2024) provide multi-axis or paired labels; we analyze structure rather than produce data.

Latent-factor / IRT methods. A recent line applies factor analysis or IRT to LLM evaluation (Ilic et al., 2025; Lou et al., 2025; Liu et al., 2025; Wang et al., 2025), but exclusively on *task-accuracy* benchmark data with binary correct/incorrect outcomes, producing latent ability vectors used for routing or benchmark quality. None of this literature analyses pairwise preference data, and none extract a named secondary axis with bootstrap-stable per-model coordinates from a per-axis BT-strength tensor.

DPO and listwise preference training. DPO (Rafailov et al., 2023) trains on (prompt, chosen, rejected) tuples; listwise variants LiPO (Liu et al., 2024), RRHF (Yuan et al., 2023), PRO (Song et al., 2024), and Constitutional

AI (Bai et al., 2022) extend the signal but still aggregate to a single utility. Our findings predict that scalar-aggregation training leaves structured signal on the table when underlying preferences are multi-axis.

Multi-objective RLHF and reward modeling. A growing line of work addresses the multi-objective preference setting at training time. Wang et al. (2024a) (ArmoRM) train an Anchored Multi-Objective Reward Model on multi-axis data including HelpSteer2 (the same dataset we analyze) with one reward head per axis and a learned mixture-of-experts gating. Zhou et al. (2024) (MODPO) extend DPO to multiple reward dimensions with per-objective margins. Ramé et al. (2023) (Rewarded Soups) propose Pareto-optimal alignment via interpolation of weights fine-tuned on different objectives. Munos et al. (2024) (Nash-LM / NLHF) drop the scalar-utility assumption entirely and formulate preference learning as a two-player game with a Nash equilibrium solution. Zhao et al. (2024) extend preference optimization to grouped user preferences. These works share our premise that single-utility aggregation discards structure; our contribution is complementary: we provide the empirical-evaluation side (a spectral map of the structure that exists in public preference data, with named axes and bootstrap-stable per-model coordinates) plus an inference-time selection method. ArmoRM in particular is the closest training-side analogue: it trains per-axis reward heads on HelpSteer2 axes that closely match our PC1/PC2 directions, but it does not analyze the geometry of those heads or test inference-time aggregation rules across axes.

Novelty boundary. To our knowledge, this is the first application of spectral decomposition with named interpretable axes to a (model $\times$ rating-axis) BT-strength tensor from pairwise preferences, with bootstrap-stable per-model coordinates, connected to inference-time selection.

3 Background and Notation

We use K candidates compared pairwise on A rating axes. For each unordered pair (i, j) and axis a, $w_{ij}^{a,+}$ / $w_{ij}^{a,-}$ are the win counts.

Bradley–Terry MLE. With pooled counts $w_{ij}^{\pm} = \sum_a w_{ij}^{a,\pm}$ the weighted BT MLE finds $\hat{\theta} \in \mathbb{R}^K$ with $P(i \succ j) = \sigma(\hat{\theta}_i - \hat{\theta}_j)$ by maximizing the standard log-likelihood; we use Zermelo iteration with Laplace smoothing. Let $B \in \mathbb{R}^{|E| \times K}$ be the signed incidence matrix ($B_{e,i} = +1$, $B_{e,j} = -1$ for $e = (i, j)$, $i < j$); $L = B^\top B$ is the unweighted comparison-graph Laplacian. The BT Hessian at $\hat{\theta}$ is a weighted Laplacian:

$$\nabla^2 \mathcal{L}_{\mathrm{BT}}(\hat{\theta}) = B^\top D B, \quad D = \mathrm{diag}\big(w_{ij}\, \sigma'(\hat{\theta}_i - \hat{\theta}_j)\big) \tag{1}$$

(Massey, 1997; Negahban et al., 2017). Effective resistance on this Laplacian gives the asymptotic variance of $\hat{\theta}_i - \hat{\theta}_j$ (Shah and Wainwright, 2018); we use it for calibrated abstention (section 7) and active comparison selection (Appendix A).

HodgeRank decomposition. Edge log-odds $Y_e = \log((w_e^+ + 0.5)/(w_e^- + 0.5))$ decompose orthogonally as $Y = B\theta^{\mathrm{grad}} + C^\top \phi + h$ (gradient, curl, harmonic) (Jiang et al., 2011), where $\theta^{\mathrm{grad}} = L^+ \mathrm{div}(Y)$ and C is the triangle curl operator. The transitivity ratio $\rho_{\mathrm{grad}} = \|B\theta^{\mathrm{grad}}\|^2 / \|Y\|^2$ measures the fraction of edge-flow energy that a single scalar potential fits.

BT-null calibration. To test whether observed ρ_{grad} is distinct from BT noise: fit BT, simulate n_{sim} replicate graphs by drawing Bernoulli outcomes from the fitted probabilities (preserving per-pair mass), recompute ρ_{grad} on each. The p-value is the fraction of simulated values $\leq$ observed. Substantially-below-null observed ρ_{grad} indicates structural non-transitivity beyond BT noise.

The central object: per-axis BT-strength tensor Θ. For each axis a, fit BT separately to obtain $\hat{\theta}^{(a)} \in \mathbb{R}^K$ (mean-zero). Stack:

$$\Theta \in \mathbb{R}^{K \times A}, \qquad \Theta_{ia} = \hat{\theta}_i^{(a)}. \tag{2}$$

The SVD $\Theta = U \Sigma V^\top$ gives PC1 = $U_{:,1}\sigma_1$ as the dominant per-model direction and $V_{:,1}$ as its axis loading. Our headline finding is that Θ on multi-axis LLM preference data has a single dominant quality direction (PC1, 70–98% of variance) plus a stable leading non-quality direction (PC2, 1.7–14.6%) encoding a named, interpretable trade-off axis.

4 When Scalar Bradley–Terry Suffices, and When It Fails

We ask whether the aggregated preference graph factors through a single scalar utility, using the diagnostic best matched to each dataset's regime: HodgeRank ρ_{grad} + BT-null simulation when only

pooled pairwise verdicts are available (MT-Bench, Chatbot Arena), direct label disagreement when two parallel labels exist (PKU-SafeRLHF), cross-axis correlation when per-candidate multi-axis ratings exist (UltraFeedback, HelpSteer2).

The three diagnostics are distinct but complementary observables of one phenomenon. HodgeRank tests *scalar inconsistency* in a pooled pairwise comparison graph (curl on triangles). Direct label disagreement tests *multi-objective conflict* between parallel labels on the same pair; this is not the same as a HodgeRank cycle, but it is the cleanest possible evidence that the data does not factor through a single utility. Cross-axis correlation tests whether per-candidate axis ratings collapse to one direction. We treat all three as evidence for the broader claim that scalar aggregation is valid only when the rating axes are effectively aligned, and we report whichever is most informative on each dataset.

Single-axis preferences fit BT. MT-Bench gpt4_pair provides 80 prompts, 6 models, K(K-1)/2 = 15 pairs judged by GPT-4 per prompt, with each pair judged in both orderings. We build 160 per-prompt graphs (80 prompts $\times$ 2 conversation turns). Observed ρ_{grad} has mean 0.841; the BT-null mean is 0.717. Only 0.6% of graphs have $p \leq 0.05$. Observed transitivity is, if anything, *higher* than BT noise: at $K = 6$ with one comparison per pair, single-comparison Bernoulli variance dominates and there is no detectable structural non-transitivity.

Chatbot Arena provides 57,477 human pairwise battles across 64 models. We restrict to the top-K most-active models for $K \in \{10, 15, 20\}$ and aggregate all battles between each model pair into a single graph with mean mass $\sim$100–1000 per edge. At all three values of K, observed ρ_{grad} matches the BT-null mean to within ± 0.020: $K = 10$: 0.925 vs 0.904, $K = 15$: 0.807 vs 0.799, $K = 20$: 0.719 vs 0.710. Aggregated human-preference data with a single underlying preference signal is BT-explainable.

Multi-axis preferences carry scalar-inconsistent structure. UltraFeedback (Cui et al., 2024) provides 55,684 prompts, 17 models, 4 GPT-4 axes (helpfulness, honesty, instruction-following, truthfulness), and 1.34M pairwise verdicts when we pool per-axis votes. Observed $\rho_{\text{grad}} = 0.9795$. BT-null over 200 simulations has mean 0.9976 with std 0.0009. **Zero of 200 simulations** produce a value as low as the observed; the gap is approximately 20σ. The observed graph is unambiguously less transitive than a single-utility model can fit (fig. 1).

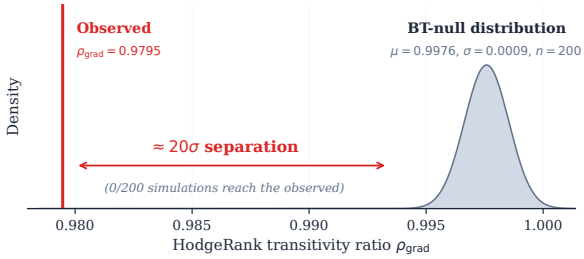

Figure 1: UltraFeedback aggregate. Observed $\rho_{\text{grad}} = 0.9795$ (dashed) sits in complete isolation from the BT-null distribution (mean 0.9976, std 0.0009, $n = 200$ simulations); $\approx 20\sigma$.

PKU-SafeRLHF (Ji et al., 2024) provides 73,907 human-annotated preference pairs with **two independent labels** per pair: a `better_response_id` (helpfulness) and a `safer_response_id` (safety). The scalar-inconsistency question reduces here to a direct count of multi-objective conflict: how often do the two labels disagree? **24.1% of all 73,907 pairs** have `better_id` $\neq$ `safer_id`. The disagreement is structured by stakes:

Max severity	n	Disagreement
0 (no harm)	30,437	26.0%
1 (mild)	4,594	**27.3%**
2 (moderate)	31,573	23.8%
3 (severe)	7,303	**15.1%**

Table 1: PKU-SafeRLHF helpfulness–safety label disagreement by maximum harm severity in the pair. The trade-off is sharpest in the moderate-stakes regime and weakest when stakes are unambiguous.

DPO training on `better_response_id` alone collapses this 24% of label disagreement into a single-axis signal that ignores safety on $\approx 24\%$ of pairs.

HelpSteer2 (Wang et al., 2024b) provides 20,324 human-rated responses on five axes. Helpfulness–correctness Spearman $r = 0.92$ (near-identical), helpfulness–verbosity $r = 0.04$, helpfulness–complexity $r = 0.14$ (verbosity and complexity are nearly orthogonal to quality). In per-prompt $K = 2$ analysis, 17.6% of HelpSteer2 prompt pairs have helpfulness and verbosity preferring different responses. Combined with PKU's 24.1% rate, multi-axis disagreement of 15–30% is

the typical regime in human-annotated preference data.

5 Structure: A Stable, Low-Rank, Named Trade-Off Axis

Having shown that scalar BT fails in the multi-axis regime (section 4), we now ask what structure the scalar model discards. We characterize the shape via SVD of Θ on UltraFeedback and HelpSteer2.

5.1 Spectral analysis: the curl is a single named axis

UltraFeedback ($K = 17$, $A = 4$). Column-centered SVD gives singular values $(5.024, 0.659, 0.334, 0.141)$ with variance explained $(97.8\%, 1.7\%, 0.4\%, 0.1\%)$. PC1 has near-equal positive loadings on all axes, capturing overall quality. PC2 axis loadings:

$$V_2^{\mathrm{UF}} = \begin{pmatrix} +0.779 \\ -0.089 \\ -0.257 \\ -0.565 \end{pmatrix} \begin{matrix} \text{(help.)} \\ \text{(hon.)} \\ \text{(IF)} \\ \text{(truth.)} \end{matrix}$$

define a single named axis: helpfulness vs. truthfulness, the other two near zero. PC2's 1.7% matches the HodgeRank curl mass (1.8%); rank-2 reconstruction accounts for 99.5% of total variance.

HelpSteer2 ($N = 20{,}324$ responses, $A = 5$). SVD of the response$\times$axis rating matrix gives variance explained $(69.5\%, 14.6\%, 7.5\%, 6.5\%, 1.9\%)$. PC1 = helpfulness/correctness-weighted quality. PC2 axis loadings:

$$V_2^{\mathrm{HS2}} = \begin{pmatrix} +0.045 \\ +0.033 \\ +0.116 \\ -0.583 \\ -0.802 \end{pmatrix} \begin{matrix} \text{(help.)} \\ \text{(corr.)} \\ \text{(coh.)} \\ \text{(compl.)} \\ \text{(verb.)} \end{matrix}$$

verbosity/complexity vs. quality. PC2's named direction differs from UltraFeedback (different axis set); the universal pattern is a single dominant non-quality direction that is interpretable and stable. The biplots (fig. 2) visualize this.

5.2 Stability and family clustering

Split-half on UltraFeedback. Random 50/50 prompt split (seed fixed); fit per-axis BT on each half independently; SVD; sign-fix PC2 against full-data V_2; correlate per-model coordinates across halves. PC1: Pearson $r = +0.9997$. PC2: Pearson $r = +\mathbf{0.9952}$. Even at PC2 magnitudes of ± 0.3 fit with half the data, per-model coordinates lie on the $y = x$ diagonal (fig. 3).

Bootstrap on UltraFeedback. $n_{\mathrm{boot}} = 100$ prompt resamples; rebuild Θ, SVD, sign-fix, record per-model PC2. Typical std $\sigma \approx 0.008$ vs. PC2 magnitudes up to ± 0.32 (signal-to-noise $\approx$ 40:1); 15 of 17 models have $|\mathrm{mean}|/\sigma > 2$ (fig. 4). The two exceptions (Bard, Vicuna-33B) are axis-neutral. All three LLaMA-2-chat sizes (7B/13B/70B) sit at PC2 $\in [+0.16, +0.32]$; both GPT models at ≈ -0.17. Per-model bootstrap std is itself predicted by effective resistance on the comparison-graph Laplacian (Pearson $r = +0.93$, $p < 10^{-7}$; Appendix M): graph-isolated models such as Pythia-12B have noisier PC2 estimates for a classical, closed-form reason.

Bootstrap on HelpSteer2. HS2 lacks per-response model identity; we instead test PC2 *axis direction*. Across $n_{\mathrm{boot}} = 200$ prompt resamples, cosine similarity between bootstrap V_2 and full-data V_2 has mean 0.9999, min 0.9994 (per-axis loadings in Appendix G). As an auxiliary cross-check, an independently-trained 8B multi-objective reward model (ArmoRM, Wang et al. 2024a) recovers the paper's V_1, V_2 at cosine $+0.87, +0.90$ on $n = 300$ HelpSteer2 responses (Appendix I); we treat this as corroborative rather than headline evidence.

Family clustering, with a caveat. Within UltraFeedback, fine-tuning lineage predicts trade-off-axis position regardless of capability: LLaMA-2-chat 7B/13B/70B cluster in the helpfulness quadrant; both GPT models in the truthfulness quadrant; effect size $\sim 40\sigma$ of bootstrap noise. We frame this as evidence that PC2 captures a stable per-model trade-off direction *within a given evaluation distribution and axis set*, with partial transfer across related prompt domains: PC2 transfers strongly between similar prompt sources (sharegpt $\leftrightarrow$ ultrachat $r = +0.92$) but only moderately between distant ones (flan_v2_niv2 $\leftrightarrow$ ultrachat $r = +0.24$); see Appendix H. The model-clustering observation is robust within UltraFeedback. As a directional robustness check on a small frontier-2026 cohort ($K = 6$ models from six different labs), the leading non-quality residual remains substantial but its named direction shifts away

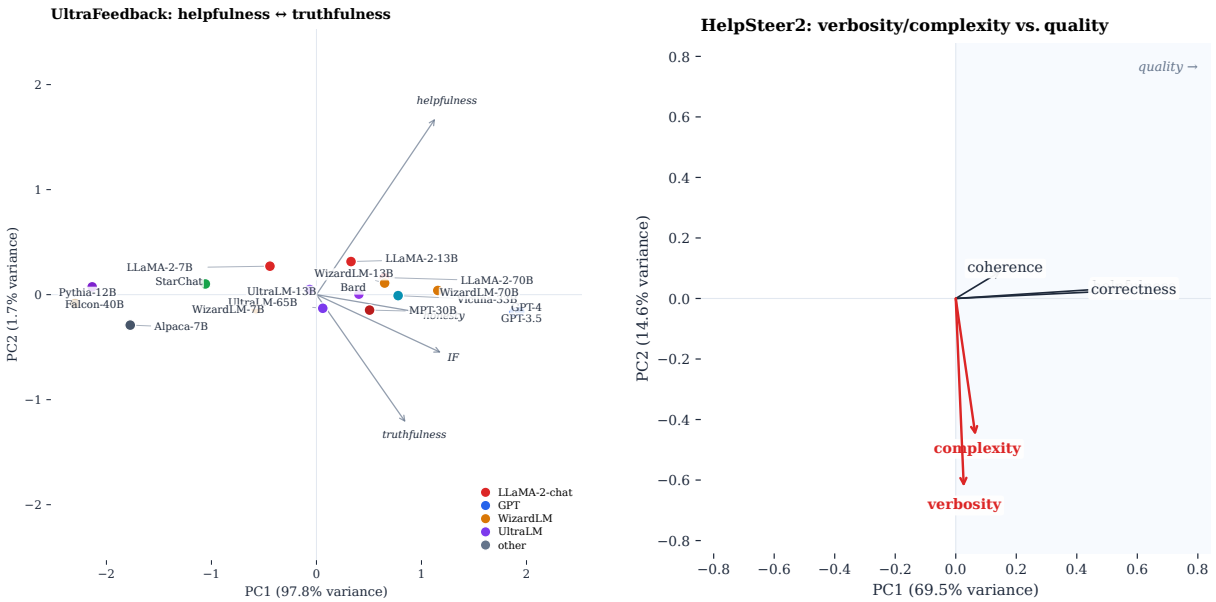

Figure 2: Per-axis BT-strength tensor biplots. **Left (UltraFeedback):** variance explained (PC1 $= 97.8\%$, PC2 $= 1.7\%$) and 17 models on (PC1, PC2) with axis-loading arrows. LLaMA-2-chat in the helpfulness quadrant; GPT models in the truthfulness quadrant. **Right (HelpSteer2):** PC1 (69.5%) = helpfulness/correctness/coherence; PC2 (14.6%) = verbosity/complexity orthogonal to quality. Direction is dataset-specific but structural shape matches.

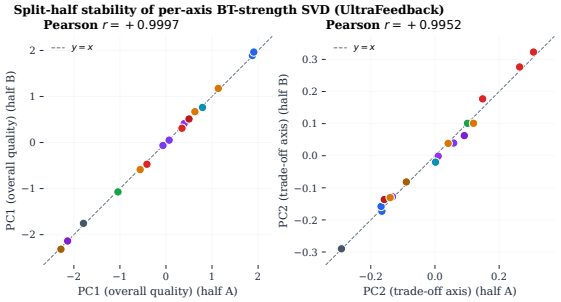

Figure 3: Split-half stability of per-axis BT-strength SVD on UltraFeedback. Left: PC1 (overall quality) per-model coordinates across two random halves of prompts, Pearson $r = +0.9997$. Right: PC2 (trade-off axis) coordinates, $r = +0.9952$. Both correlations sit on the $y = x$ diagonal.

from the 2023 helpfulness ↔ truthfulness axis, suggesting that the specific 2023 trade-off is less dominant in this modern, capability-compressed cohort (Appendix J).

6 Prediction: Per-Prompt Diagnostics Forecast Ranking Instability

If the scalar approximation fails only when axes diverge, we should be able to predict prompt-level instability from cheap cross-axis disagreement statistics. We test this directly: can we detect at the per-prompt scale whether a given prompt's ranking will be unstable to changes in the rating-axis set?

Setup. On UltraFeedback ($K = 4$, $A = 4$, 55,538 prompts) we compute per-prompt diagnostics: mean/min Kendall τ across the $\binom{4}{2} = 6$ axis-pair rank correlations; cross-axis pair disagreement rate; per-prompt HodgeRank ρ_{grad}; rating spread. Held-out target: leave-one-axis-out disagreement; sum the other three axes as training score, predict pairwise winners, compare to the held-out axis's verdicts (averaged across the four held-outs).

Result. Simple per-prompt statistics predict held-out instability with high accuracy:

Binned-mean relationships are monotone and tight for the two strong predictors and nearly flat for per-prompt ρ_{grad}; the visualization is in Appendix F.

Scale-dependence note. Aggregate-scale HodgeRank curl carries the structural signal (20σ on UltraFeedback, section 4), but per-prompt ρ_{grad} at $K = 4$ is uninformative because per-prompt graphs have only 6 edges. Simple rank statistics dominate at small K; spectral methods dominate at aggregate K.

Implication. Per-prompt min Kendall τ across axes is a 10-line-of-code diagnostic predicting $\approx 73\%$ of variance in held-out ranking instability, a

Diagnostic	min Kendall τ	cross-axis disagr.	mean Kendall τ	rating range	n_3cycles (K=4)	ρ_{grad} (K=4)
Spearman ρ	$\mathbf{-0.817}$	$\mathbf{+0.815}$	-0.732	-0.278	$+0.174$	-0.058
p	$< 10^{-30}$	$< 10^{-30}$	$< 10^{-30}$	$< 10^{-30}$	$< 10^{-9}$	$< 10^{-5}$
Combined OLS $R^2 = \mathbf{0.728}$						

Table 2: Per-prompt diagnostics vs. leave-one-axis-out disagreement on UltraFeedback ($n = 55{,}538$). Two simple statistics (min Kendall τ between axis rankings and cross-axis disagreement rate) carry essentially all of the predictive signal.

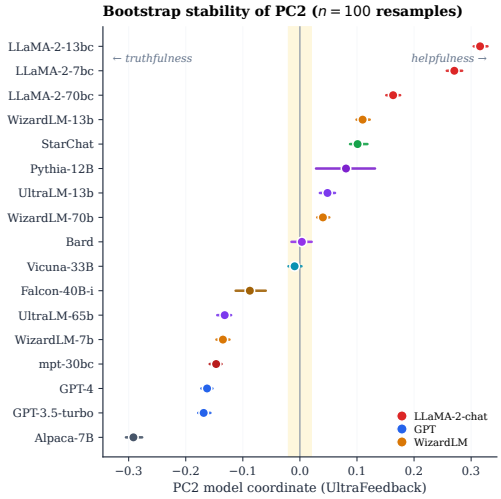

Figure 4: Bootstrap forest plot of PC2 model coordinates on UltraFeedback ($n = 100$ resamples). Models ordered by PC2 position. LLaMA-2-chat family clusters together in the helpfulness quadrant; both GPT models cluster together in the truthfulness quadrant. 15/17 CIs exclude zero; Bard and Vicuna-33B (near zero) are the "axis-neutral" models. Family lineage predicts PC2 position with effect size $\approx 40\sigma$ of bootstrap noise.

cheap "multi-objective ambiguity" signal for evaluation pipelines.

7 Inference-Time Reranking: A Regime-Dependent Win

Finally, if the discarded structure is real, it should change which candidate we select at inference time. Best-of-N reranking is the practical setting where multi-axis preferences would most plausibly help. We test multi-axis-aware rerankers against scalar BT across two regimes that behave very differently, and find that they do, exactly where the diagnostic of section 4 predicts.

7.1 Reranker variants and aggregation ablations

For each per-prompt graph (4 axes, K candidates, $\binom{K}{2}$ pairwise verdicts per axis) we test rerankers in three classes. **(i) Single-utility baselines**: *scalar BT* (pool axes, fit BT, argmax); *HodgeRank gradient* ($L^+\mathrm{div}(Y)$); *Rank Centrality* (Negahban et al., 2017) (Perron eigenvector); *Variant A* (per-axis BT, stack into Θ, return $\arg\max U_{:,1}\sigma_1$). All four collapse multi-axis info to a scalar. **(ii) Linear multi-axis aggregations**: *Borda* (sum of per-axis ranks), *average rank* (mean of per-axis 'position' rank), *weighted plurality* (per-axis top-1 vote, axis weight = discriminative range). These three use richer per-axis information but aggregate linearly. **(iii) Non-linear multi-axis aggregations**: *Variant E (plurality)* (per-axis top-1, plurality winner); *maximin* (Pareto-conservative: minimize worst-axis position). **Calibrated abstention** layers on top of any base reranker, falling back to majority vote when the BT margin is below $1.96\sqrt{\mathrm{Var}(\hat{\theta}_{(1)} - \hat{\theta}_{(2)})}$. Random / longest are baselines; majority vote applies only to math.

7.2 Verifiable math tasks: methods tie scalar BT

Setup (shared). $K = 8$ candidates from Llama-3.1-8B-Instruct (Dubey et al., 2024) at temperature 0.8; pairwise verdicts from GPT-4o-mini (OpenAI, 2024) on four axes (correctness, clarity, conciseness, completeness); 2 swap orderings per pair.

GSM8K (easy, $N = 50$). Base accuracy 85% (mean 6.77/8 candidates correct). All BT-based methods tie at pass@1 = 0.917; majority vote = 0.958; **scalar BT + L^+-uncertainty abstention reaches** 0.979 (+6.2pp), the one GSM8K regime where the comparison-graph Laplacian explicitly drives a reranker win.

MATH levels 3–4 (hard, $N = 200$). Problems from MATH (Hendrycks et al., 2021) where Llama-3.1-8B has meaningful but non-trivial failure (base accuracy 29%); 135/200 have ≥ 1 correct candidate. Pass@1 over the 135 discriminative prompts:

Method	pass@1	stderr
majority vote	0.674	0.040
scalar BT	0.674	0.040
per-axis PC1 (Variant A)	0.681	0.040
HodgeRank grad.	0.667	0.041
Rank Centrality	0.674	0.040
Variant E (plurality)	**0.696**	0.040

Table 3: Pass@1 on MATH levels 3–4 ($n = 135$ discriminative prompts). Random $= 0.437$, longest $= 0.400$. All BT-based methods cluster within ± 0.03. We initially observed at $N = 50$ a $+5.9$pp gap for Variant A that did not survive scaling; $N = 200$ numbers are authoritative.

Method diversity persists. Majority vote and scalar BT pick the same candidate on only 17% of prompts (each correct on $72/112$ disagreements); the MV + scalar-BT oracle reaches 0.793 ($+12$pp) but no realizable ensembler captures this gap.

7.3 Open-ended tasks: per-axis plurality strongly outperforms scalar BT

The MATH result raises the question: is multi-axis structure ever actionable for direct reranking? We test on Arena-Hard (Li et al., 2024), where prompts are open-ended and rating axes can genuinely diverge across candidates.

Setup. 200 prompts from `lmsys/arena-hard-auto-v0.1`; same generator/judge stack as math, axes (helpfulness, factuality, conciseness, fluency). **Ground truth**: per-candidate per-axis 1–5 ratings from Llama-3.3-70B (Dubey et al., 2024) (independent stronger judge, strict rubric). Five oracles: sum-of-axes ("overall") and per-axis best. We filter to non-flat oracles ($n = 92$–156).

Headline result. Per-axis plurality (Variant E) wins on every oracle. Ground truth is LLM-rated; as a robustness check, Appendix E replicates under a second strong judge from a different model family (Claude Opus 4.6). The gap survives under each judge separately and *strengthens* to $+26$pp on the n=50 subset where both judges agree on the top-1 (0.480 vs. 0.220).

Why linear fails, non-linear wins. table 4 shows a clean dichotomy: every *linear* aggregation clusters within ± 0.02 of scalar BT; both *non-linear* ones substantially outperform (Variant E by $\approx 3.3\sigma$ overall and $7.2\times$ on factuality; maximin by $+19$pp). Mechanism: the gain is localized to axis-divergent prompts (n=141), where Variant E reaches 0.355 vs. scalar BT 0.170 ($+18.5$pp); single-dominator prompts (n=15) tie at 0.467. Linear rules collapse to the largest-discriminative-range axis; non-linear rules recover per-axis structure (Appendix B). On verifiable math, axes agree on the correct candidate and the distinction collapses (table 3). Within Arena-Hard, the regime split replicates across subject categories: Variant E gains $+37$pp on advice, $+23$pp on creative, but ± 0pp on math/reasoning (Appendix D).

Method	Overall (n=156)	Factuality (n=92)
(i) Single-utility baselines		
scalar BT	0.199	0.087
per-axis PC1 (A)	0.205	0.076
HodgeRank grad.	0.167	0.065
Rank Centrality	0.186	0.076
(ii) Linear multi-axis aggregations		
Borda	0.205	0.087
average rank	0.205	0.087
weighted plurality	0.205	0.087
(iii) Non-linear multi-axis aggregations		
Variant E (plurality)	**0.365**	**0.630**
maximin	**0.385**	0.446

Table 4: Pass@1 on Arena-Hard vs. Llama-3.3-70B per-axis ground truth ($N = 200$, $K = 8$). Only two columns shown for body compactness; Helpfulness, Conciseness, and Fluency columns in Appendix C. Linear aggregations cluster around scalar BT (≈ 0.20); non-linear ones (plurality, maximin) substantially outperform. Plurality factuality lift is $7.2\times$ (0.630 vs. 0.087); maximin edges plurality on the overall by $+2$pp.

8 Discussion

Scalar BT is regime-conditional. The central lesson is regime-conditionality. Scalar BT/Elo is a good approximation when rating axes agree (section 4: MT-Bench, Chatbot Arena), but in multi-axis regimes the residual is not noise: it is low-rank, stable, named, and consequential for selection ($+18.5$pp Variant-E lift on axis-divergent prompts, section 7). For leaderboards, we suggest reporting $(|\mathrm{PC}_2|, V_2)$ alongside scalar quality when bootstrap-stable; the LLaMA-2-chat $\leftrightarrow$ GPT PC2 separation is a property no single Elo can express.

Limitations

Cheap-judge saturation. Our pairwise judging on Arena-Hard had TIE rates of 50–70% per axis, partly because Llama-3.1-8B candidates at temper-

ature 0.8 are often hard to distinguish. We mitigated this with swap orderings and aggregation, but a stronger cheap judge could shift the results.

Strong-judge ground truth. Arena-Hard ground truth is LLM-rated, not human-annotated. We mitigate via (i) a strict rubric, (ii) per-axis evaluation, and (iii) replication with a second strong judge from a different model family (Claude Opus 4.6, Appendix E) where the Variant E advantage survives in the same direction with a smaller magnitude, and strengthens to $+26$pp on the subset where both judges agree on the top-1 candidate. A human-annotated multi-axis open-ended benchmark would further strengthen the result.

Sample size. Variant E vs. scalar BT is $\approx 3.3\sigma$ at $N = 200$ on Arena-Hard. The per-axis effects (factuality $7.2\times$) are large enough that direction is unambiguous; larger N would tighten error bars without changing the regime conclusion.

Model family coverage. PC2 family clustering covers LLaMA-2-chat, GPT, WizardLM, UltraLM, and Alpaca variants. Whether the same pattern holds for newer families (Claude, Mistral-fine-tunes, DeepSeek) is a natural follow-up.

No new theory. Our contribution is empirical. The classical tools we use (BT, HodgeRank, SVD, effective resistance) are not extended; rather, we map when each applies to LLM preference data.

Single-prompt evaluation. All findings are about per-prompt or per-pair structure within a single conversation turn. Multi-turn agent trajectories, where preferences accumulate across turns, are unaddressed.

Ethics Statement

This work analyzes public preference datasets (Chatbot Arena, MT-Bench, UltraFeedback, HelpSteer2, PKU-SafeRLHF, Arena-Hard, GSM8K, MATH) and generates additional pairwise comparisons via OpenRouter. Open-weight models used: Llama-3.1-8B-Instruct (candidate generator), Llama-3.3-70B-Instruct (strong judge), and the Appendix J frontier-2026 cohort (Llama-4-Maverick, DeepSeek-V3.2, Qwen3-235B). Commercial APIs: GPT-4o-mini (cheap judge), Claude Opus 4.6 (second strong judge, Appendix E; also in frontier-2026 cohort), GPT-5 and Gemini-2.5-Pro (frontier-2026 cohort only). ArmoRM (RLHFlow/ArmoRM-Llama3-8B-v0.1, Appendix I) is open-weight and run locally on consumer GPU. No human subjects were recruited. The PKU-SafeRLHF analysis includes labels concerning harmful content; we report aggregate disagreement statistics only, never reproducing or amplifying specific harmful generations. Our findings on the helpful–safer trade-off may inform alignment training pipelines: we view this as a positive externality (better visibility into preference-data structure for safety-relevant axes) rather than a dual-use risk.

AI-assistant use. During paper preparation, the authors used Claude (Anthropic) as a coding and writing assistant: drafting plotting code, writing experiment scripts, summarizing literature, and revising prose. All experimental results, claims, and final wording are the authors' own; the assistant did not generate empirical numbers or design experiments. We disclose this in compliance with ACL/ARR policy.

References

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, and 1 others. 2022. Constitutional ai: Harmlessness from ai feedback. *arXiv:2212.08073*.

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

Yuxin Chen, Jianqing Fan, Cong Ma, and Kaizheng Wang. 2019. Spectral method and regularized mle are both optimal for top-k ranking. *Annals of Statistics*, 47(4):2204–2235.

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 1 others. 2024. Chatbot arena: An open platform for evaluating llms by human preference. In *ICML*.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, and 1 others. 2021. Training verifiers to solve math word problems. *arXiv:2110.14168*.

Ganqu Cui, Lifan Yuan, Ning Ding, Guanming Yao, Wei Zhu, Yuan Ni, Guotong Xie, Zhiyuan Liu, and Maosong Sun. 2024. Ultrafeedback: Boosting language models with high-quality feedback. In *ICML*.

Abhimanyu Dubey and 1 others. 2024. The llama 3 herd of models. *arXiv:2407.21783*.

Arpad E Elo. 1978. *The Rating of Chessplayers, Past and Present*. Arco Pub.

Moein Falahatgar, Yi Hao, Alon Orlitsky, Venkatadheeraj Pichapati, and Vaishakh Ravindrakumar. 2017. Maxing and ranking with few assumptions. In *NeurIPS*.

Evan Frick, Connor Chen, Joseph Tennyson, Tianle Li, Wei-Lin Chiang, Anastasios N Angelopoulos, and Ion Stoica. 2025. Prompt-to-leaderboard. In *ICML*. ArXiv:2502.14855.

Reinhard Heckel, Max Simchowitz, Kannan Ramchandran, and Martin J Wainwright. 2018. Approximate ranking from pairwise comparisons. In *AISTATS*.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. Measuring mathematical problem solving with the math dataset. In *NeurIPS Datasets and Benchmarks*.

Stefan Ilic and 1 others. 2025. Iq test for llms: An evaluation framework for uncovering core skills in llms. *arXiv:2507.20208*.

Jiaming Ji, Mickel Liu, Juntao Dai, Xuehai Pan, Chi Zhang, Ce Bian, Ruiyang Sun, Yizhou Wang, and Yaodong Yang. 2024. Beavertails: Towards improved safety alignment of llm via a human-preference dataset. In *NeurIPS Datasets and Benchmarks*.

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

Richard M Karp and Robert Kleinberg. 2007. Noisy binary search and its applications. In *SODA*.

Tianle Li, Wei-Lin Chiang, Evan Frick, Lisa Dunlap, Tianhao Wu, Banghua Zhu, Joseph E Gonzalez, and Ion Stoica. 2024. From crowdsourced data to high-quality benchmarks: Arena-hard and benchbuilder pipeline. In *arXiv:2406.11939*.

Hongkang Liu and 1 others. 2025. Lost in benchmarks? rethinking large language model benchmarking with item response theory. In *AAAI*. ArXiv:2505.15055.

Tianqi Liu, Zhen Qin, Junru Wu, Jiaming Shen, Misha Khalman, Rishabh Joshi, Yao Zhao, Mohammad Saleh, Simon Baumgartner, Jialu Liu, and 1 others. 2024. Lipo: Listwise preference optimization through learning-to-rank. In *arXiv:2402.01878*.

Jiawei Lou and 1 others. 2025. Learning compact representations of llm abilities via item response theory. *arXiv:2510.00844*.

Kenneth Massey. 1997. Statistical models applied to the rating of sports teams. *Bluefield College, Bluefield, VA*.

Lucas Maystre and Matthias Grossglauser. 2017. Just sort it! a simple and effective approach to active preference learning. In *ICML*.

Rémi Munos, Michal Valko, Daniele Calandriello, Mohammad Gheshlaghi Azar, Mark Rowland, Zhaohan Daniel Guo, Yunhao Tang, Matthieu Geist, Thomas Mesnard, Andrea Michi, Marco Selvi, Sertan Girgin, Nikola Momchev, Olivier Bachem, Daniel J. Mankowitz, Doina Precup, and Bilal Piot. 2024. Nash learning from human feedback. In *ICML*. ArXiv:2312.00886.

Sahand Negahban, Sewoong Oh, and Devavrat Shah. 2017. Rank centrality: Ranking from pairwise comparisons. *Operations Research*, 65(1):266–287.

OpenAI. 2024. Gpt-4o mini: advancing cost-efficient intelligence. https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. In *NeurIPS*.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. In *NeurIPS*.

Alexandre Ramé, Guillaume Couairon, Mustafa Shukor, Corentin Dancette, Jean-Baptiste Gaya, Laure Soulier, and Matthieu Cord. 2023. Rewarded soups: Towards pareto-optimal alignment by interpolating weights fine-tuned on diverse rewards. In *NeurIPS*. ArXiv:2306.04488.

Nihar B Shah and Martin J Wainwright. 2018. Simple, robust and optimal ranking from pairwise comparisons. *Journal of Machine Learning Research*, 18(199):1–38.

Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. 2024. Scaling llm test-time compute optimally can be more effective than scaling model parameters. *arXiv:2408.03314*.

Feifan Song, Bowen Yu, Minghao Li, Haiyang Yu, Fei Huang, Yongbin Li, and Houfeng Wang. 2024. Preference ranking optimization for human alignment. In *AAAI*.

Haoxiang Wang, Wei Xiong, Tengyang Xie, Han Zhao, and Tong Zhang. 2024a. Interpretable preferences via multi-objective reward modeling and mixture-of-experts. In *EMNLP Findings*. ArXiv:2406.12845.

Yuxuan Wang and 1 others. 2025. Irt-router: Effective and interpretable multi-llm routing via item response theory. In *ACL*. ArXiv:2506.01048.

Zhilin Wang, Yi Dong, Olivier Delalleau, Jiaqi Zeng, Gerald Shen, Daniel Egert, Jimmy J Zhang, Makesh Narsimhan Sreedhar, and Oleksii Kuchaiev. 2024b. Helpsteer2: Open-source dataset for training top-performing reward models. In *arXiv:2406.08673*.

Zheng Yuan, Hongyi Yuan, Chuanqi Tan, Wei Wang, Songfang Huang, and Fei Huang. 2023. Rrhf: Rank responses to align language models with human feedback without tears. In *arXiv:2304.05302*.

Yifan Zhang and 1 others. 2024. Beyond bradley-terry models: A general preference model for language model alignment. *arXiv:2410.02197*.

Siyan Zhao, John Dang, and Aditya Grover. 2024. Group preference optimization: Few-shot alignment of large language models. In *ICLR*. ArXiv:2310.11523.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, and 1 others. 2024. Judging llm-as-a-judge with mt-bench and chatbot arena. In *NeurIPS Datasets and Benchmarks*.

Zhanhui Zhou, Jie Liu, Chao Yang, Jing Shao, Yu Liu, Xiangyu Yue, Wanli Ouyang, and Yu Qiao. 2024. Beyond one-preference-fits-all alignment: Multi-objective direct preference optimization. In *ACL Findings*. ArXiv:2310.03708.

A Active Comparison Selection: full details

Setup. We use the aggregated UltraFeedback per-axis graph at $K = 17$ models. Each unordered model pair (i, j) has a pool of ≈ 9000 axis-prompt verdicts. The oracle on a query is a uniform draw from the pool. We compare five strategies that sequentially choose which pair to query:

1. **Random**: uniform over $\binom{17}{2} = 136$ pairs.
2. **Round-robin**: canonical-order sweep.
3. $\Delta\lambda_2$ **greedy**: query the pair whose addition maximizes the algebraic connectivity $\lambda_2(L)$ of the unweighted comparison graph.
4. $\Delta\mathrm{Tr}(L^+)$ **greedy** (effective resistance / L+-uncertainty).
5. **Top-k unstable**: heuristic; query the lowest-mass edge incident to the current BT top-k.

For each strategy and each comparison budget m, we report Kendall τ between the BT fit on the m queried verdicts and the BT fit on all ~ 1.34M verdicts (target ranking). 25 repeats per (strategy, m) cell.

Result. Figure 5 shows the budget–accuracy curves. $\Delta\mathrm{Tr}(L^+)$ greedy is consistently best or near-best across the data-poor regime ($m \leq 100$), with $\Delta\lambda_2$ greedy second. Random and tournament round-robin lag behind in the small-m regime; all strategies converge by $m \geq 200$.

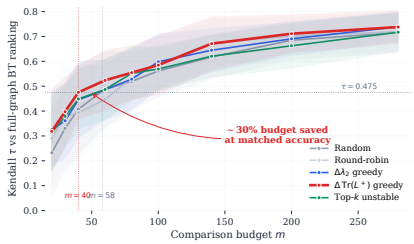

Figure 5: Active-comparison-selection budget–accuracy curves on UltraFeedback aggregated $K = 17$ graph. Effective-resistance greedy ($\Delta\mathrm{Tr}(L^+)$, red) consistently leads in the budget-tight regime ($m \in [20, 100]$). All strategies converge by $m \geq 200$. Random reaches Kendall $\tau = 0.475$ at $m \approx 58$; $\Delta\mathrm{Tr}(L^+)$ reaches the same accuracy at $m = 40$ ($\approx$30% budget savings).

Quantitative summary. At $m = 40$, $\Delta\mathrm{Tr}(L^+)$ reaches Kendall $\tau = 0.475$ vs. random at 0.408. Linear interpolation: random reaches $\tau = 0.475$ at $m \approx 58$. The effective-resistance strategy saves approximately 30% of the comparison budget for matched accuracy in the data-poor regime. Effects are 1–2σ at any single budget cell but the direction is consistent across the full sweep.

Caveats. The gain is modest (not order-of-magnitude). It applies in the regime where every pair has roughly equal-quality verdicts; if some pairs are inherently more informative (high $\sigma'(\Delta\theta)$), the gap may widen. The result establishes that the comparison-graph Laplacian is not purely descriptive but drives a real, if contained, methodological improvement.

B Mechanism diagnostics for Variant E

Computed from the 156 discriminative Arena-Hard prompts (those with non-flat strong-judge sum-of-axes ground truth).

Stratified pass@1 by axis-divergence.

Distribution of distinct per-axis cheap-judge winners. 1 distinct winner: 15 prompts (9.6%); 2 distinct: 141 (90.4%); ≥ 3: 0. Cheap-judge per-axis BT distinguishes mostly two candidates per prompt: a primary dominator and an axis-specialist runner-up.

distinct axis winners	n	scalar BT	Variant E
1 (single dominator)	15	0.467	0.467
≥ 2 (axes diverge)	141	0.170	**0.355**

Table 5: Variant E gains pass@1 exactly when the cheap judge produces axis-divergent rankings and exactly ties scalar BT when the cheap judge agrees across axes.

Scalar BT picks the axis-discriminative-range maximizer. The number of cheap-judge axes on which scalar BT's pick is the top-1 candidate: 89.1% of picks are top on exactly one axis (the largest-discriminative-range axis dominates the pooled BT, here typically conciseness), 9.6% are top on all four axes (the single-dominator regime), and 0.6% are top on zero axes (rare BT-induced pairwise quirks). By construction, Variant E's pick is top on ≥ 1 axis on 100% of prompts.

Method agreement. Scalar BT and Variant E pick the same candidate on only 10.3% of discriminative prompts; on the 89.7% disagreement prompts the two methods literally choose different outputs. The diversity is real, not noise.

C Arena-Hard reranker pass@1: full per-axis table

Full 5-oracle version of table 4.

D Per-category Arena-Hard breakdown

We roll up Arena-Hard's fine-grained subject tags into five coarse categories via a keyword classifier (released with the code). Per-category Variant E vs. scalar BT pass@1 on the overall sum-of-axes oracle:

E Multi-judge robustness on Arena-Hard

To address the LLM-on-LLM evaluation concern in the body's Arena-Hard reranking result, we re-rated all 200 prompts $\times$ 8 candidates with a second strong judge from a different model family: Claude Opus 4.6 via OpenRouter (model id `anthropic/claude-opus-4.6`, accessed 2026-05-21), using the *identical* 1–5 per-axis rubric we used for the Llama-3.3-70B run. Because the two judges come from different model families and providers, agreement between them reduces the likelihood that the reranking result is specific to one judge's rubric.

Coverage. Both judges fully rated $174/200$ prompts (87.0%). The remaining 26 prompts split into: Llama-3.3-70B parse / formatting failures on 23 prompts; Opus-4.6 safety refusals on 3 prompts (offensive-security topics such as TFTP fuzzers and exploit tooling that Llama scored without refusal). The differential refusal pattern is itself a form of judge-disagreement.

Judge agreement on top-1. On the 159 both-rated prompts, the two judges agree on which candidate is overall best (sum across axes) only 33.3% of the time ($53/159$). Per-axis agreement is higher but still modest: helpfulness 45.3%, factuality 47.8%, conciseness 45.9%, fluency 48.4%. Two strong LLM judges from different model families disagree more than they agree on overall top-1 on multi-axis open-ended preferences, itself a striking finding about multi-axis LLM-as-judge evaluation.

Variant E wins under both judges independently, and *more* on their intersection. Reading: the previously-reported $\approx 3.3\sigma$ advantage of Variant E over scalar BT under Llama-3.3-70B *survives* under a stricter, different-family judge (in the same direction, with a smaller magnitude) and *strengthens* on the subset where both judges agree on the top-1 candidate. On the agreement subset, Variant E and maximin are $\sim 5\times$ better than scalar BT (0.480 / 0.500 vs. 0.220).

Why the gap differs by judge: axis-binding. The smaller Opus-4.6 gap has an interpretable cause that is itself a finding about LLM-as-judge dispositions. Cross-axis Pearson correlations in the per-candidate ratings:
Llama treats the four axes as relatively independent dimensions ($r \approx 0.38$–0.49); Opus collapses helpfulness, factuality, and fluency into something close to a single quality judgment ($r \approx 0.72$–0.80). Variant E captures gain by promoting candidates that are the cheap judge's per-axis top-1 on at least one axis; that gain converts to ground-truth pass@1 only when the ground-truth judge also treats per-axis distinctions as meaningfully separable. Llama does; Opus largely does not, so the per-axis specialist often coincides with Opus's overall winner: exactly the regime in which scalar BT and Variant E converge. The both-judges-agree subset is the cleanest test because it isolates prompts on which the two judges' very different rating styles converge on the same winner. We treat the sur-

Method	Overall (n=156)	Helpfulness (n=108)	Factuality (n=92)	Conciseness (n=144)	Fluency (n=124)
(i) Single-utility baselines					
scalar BT	0.199	0.185	0.087	0.222	0.161
per-axis PC1 (A)	0.205	0.194	0.076	0.229	0.153
HodgeRank grad.	0.167	0.157	0.065	0.188	0.129
Rank Centrality	0.186	0.176	0.076	0.208	0.145
(ii) Linear multi-axis aggregations					
Borda	0.205	0.194	0.087	0.229	0.153
average rank	0.205	0.194	0.087	0.229	0.153
weighted plurality	0.205	0.194	0.087	0.229	0.153
(iii) Non-linear multi-axis aggregations					
Variant E (plurality)	**0.365**	**0.481**	**0.630**	**0.417**	**0.548**
maximin	**0.385**	0.380	0.446	0.403	0.395

Table 6: Full pass@1 on Arena-Hard vs. Llama-3.3-70B per-axis ground truth, all 5 oracles. Plurality dominates every per-axis oracle (factuality 7.2×, helpfulness 2.6×, fluency 3.4×, conciseness 1.9×); maximin edges plurality on the sum-of-axes overall by +2pp.

Category	n	scalar BT	Variant E	Δ
advice / business	19	0.053	**0.421**	+37
creative / roleplay	13	0.231	**0.462**	+23
coding	87	0.172	**0.345**	+17
math / reasoning	19	0.316	0.316	±0
factual / explain	5	0.400	0.400	±0

Table 7: Variant E wins big on advice, creative, and coding (where helpfulness vs. conciseness vs. factuality genuinely diverge across candidates) and exactly ties scalar BT on math/reasoning and factual explanation (where an underlying correctness signal aligns all axes). The axis-divergence-vs.-convergence mechanism that distinguishes Arena-Hard from MATH at the dataset level also distinguishes advice-style from math-style prompts within Arena-Hard.

Method	Llama oracle (n=156)	Opus oracle (n=195)	Both agree (n=50)
scalar BT	0.199	0.231	0.220
per-axis PC1 (A)	0.205	0.215	0.220
HodgeRank grad.	0.167	0.205	0.180
Rank Centrality	0.186	0.226	0.200
Borda	0.205	0.226	0.220
Variant E (plurality)	**0.365**	**0.277**	**0.480**
maximin	**0.385**	**0.292**	**0.500**

Table 8: Pass@1 on Arena-Hard overall sum-of-axes oracle under each strong judge separately, plus the intersection subset (both judges agree on top-1). The reranking gap survives independent evaluation: Variant E beats scalar BT by +5pp under Opus-4.6 and by +17pp under Llama-3.3-70B. The gap is largest (+26pp) on the subset where the two judges agree on the correct top-1, the cleanest possible ground truth available without human annotation. The linear-vs.-non-linear dichotomy reproduces under both judges.

viving direction and the agreement-subset effect together as strong evidence against the "Variant E only matches one LLM-judge's idiosyncratic rubric" objection; differential axis-binding across judge families is a useful methodological footnote in its own right. The small remaining gap to a human-annotated benchmark is our most important next-step validation. Full pairwise verdicts, per-candidate ratings from both judges, and analysis code are released with the paper.

F Per-prompt diagnostic plot

Companion visualization to table 2: binned-mean relationships between per-prompt diagnostics and leave-one-axis-out disagreement on UltraFeedback (n = 55,538 prompts). The two strong predictors (min Kendall τ and cross-axis disagreement rate) are monotone and tight; the per-prompt ρ_{grad} panel is nearly flat, confirming that HodgeRank curl is too noisy at $K = 4$.

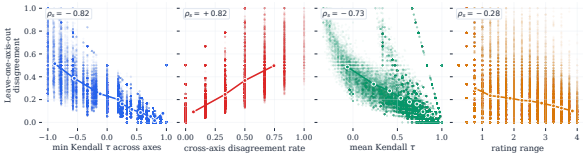

Figure 6: Per-prompt diagnostics vs. leave-one-axis-out disagreement on UltraFeedback (binned means with stderr).

G HelpSteer2 PC2 axis-loading bootstrap

Per-axis PC2 loadings on HelpSteer2, bootstrap mean and 95% CI (n_{boot} = 200 prompt resamples).

Axis pair	Llama-3.3-70B	Opus-4.6
helpfulness × factuality	$+0.38$	$\mathbf{+0.80}$
helpfulness × fluency	$+0.49$	$\mathbf{+0.73}$
factuality × fluency	$+0.44$	$\mathbf{+0.72}$
conciseness × fluency	$+0.64$	$+0.61$

Table 9: Per-candidate axis-rating correlations under each strong judge.

Axis	PC2 mean	95% CI
verbosity	-0.802	$[-0.814, -0.790]$
complexity	-0.583	$[-0.600, -0.567]$
coherence	$+0.115$	$[+0.096, +0.135]$
helpfulness	$+0.045$	$[+0.037, +0.052]$
correctness	$+0.033$	$[+0.025, +0.041]$

Table 10: HelpSteer2 PC2 axis loadings. Every CI is tight (width ≤ 0.04); all five axes have loadings statistically distinguishable from zero. The direction is verbosity/complexity vs. the quality cluster, determined to four decimal places.

H Open Questions: Detailed Results

We summarize results for three open questions tested with data already in hand. None of these is a headline claim; each provides partial evidence that informs the program's scope.

Q-stability: PC2 across prompt-domain subsets (UltraFeedback). We split the 55,684 UltraFeedback prompts by their `source` label and refit per-axis BT plus SVD on each of the four largest source subsets (sharegpt, $n = 17.8$k; flan_v2_niv2, 11.8k; evol_instruct, 9.2k; ultrachat, 9.1k). The 16 models present in all four subsets give cross-source Pearson correlations of per-model PC coordinates:

Statistic	PC1	PC2
mean off-diagonal r	**+0.962**	**+0.655**
range of off-diagonal r	$[+0.941, +0.985]$	$[+0.236, +0.915]$

Table 11: Per-model PC coordinate stability across UltraFeedback source subsets.

PC1 is essentially universal across prompt domains. PC2 transfers strongly between similar prompt distributions (sharegpt $\leftrightarrow$ ultrachat: $r = 0.915$; $\leftrightarrow$ evol_instruct: 0.867) but weakly between distant ones (flan_v2_niv2 $\leftrightarrow$ ultrachat: 0.236). Task-instruction-style data (flan_v2_niv2) stresses the multi-axis trade-off the most (PC2 explains 6.4% of variance there, vs. 0.8% on sharegpt). The "model personality" framing survives but with the caveat that the trade-off axis direction is somewhat domain-conditioned. The per-source V_2 direction cosine to the full-data V_2 is ≥ 0.85 for all four sources, so the axis is broadly consistent in direction even when explained-variance differs.

Q-benchmarks: does UF-PC2 predict TruthfulQA? We matched 12 of the 17 UltraFeedback models to entries in the (deprecated) Open LLM Leaderboard v1 (proprietary models gpt-3.5-turbo, gpt-4, bard, plus two not on HF leaderboard, are excluded). Raw Pearson r between PC2 and each of six benchmarks is essentially zero (-0.05 to $+0.04$), because PC1 dominates everything (overall capability correlates with every benchmark). Controlling for PC1 via partial correlation:

Benchmark	partial r(PC2 | PC1)	p
ARC	-0.217	0.50
HellaSwag	-0.050	0.88
MMLU	-0.149	0.64
TruthfulQA	$\mathbf{-0.450}$	**0.14**
Winogrande	-0.154	0.63
GSM8K	-0.127	0.69

Table 12: Partial Pearson r of UF-PC2 with Open LLM Leaderboard v1 benchmarks, controlling for PC1, $n = 12$.

The PC2–TruthfulQA partial correlation is the strongest among the six benchmarks and in the predicted direction (higher PC2 = more helpfulness-leaning $\Rightarrow$ lower TruthfulQA). It does not reach significance at $n = 12$, and ARC/MMLU show similar weak negative partials, so we cannot rule out "PC2 measures RLHF-helpfulness-tuning intensity, which has a small generic academic-benchmark cost" rather than "PC2 specifically predicts TruthfulQA." Adding proprietary models (gpt-3.5-turbo, gpt-4, at low PC2 with high TruthfulQA) would clarify, but those are not on the leaderboard.

Q-scale: PC2 vs. model scale within a fixed training procedure. We compare PC2 magnitudes across model sizes within two families with bootstrap CIs from the E6 stability test:

Within LLaMA-2-chat (same training procedure, scale varies) PC2 magnitude is smaller at 70B than at 7B/13B, consistent with "more capable models satisfy multi-axis requirements more equally." WizardLM however flips sign between 7B and 13B; different WizardLM sizes use different training mixtures, so scale and training pro-

Model	PC2	95% CI
llama-2-7b-chat	$+0.271$	$[+0.257, +0.284]$
llama-2-13b-chat	$+0.316$	$[+0.304, +0.329]$
llama-2-70b-chat	$+0.163$	$[+0.151, +0.176]$
wizardlm-7b	-0.135	n/a
wizardlm-13b	$+0.110$	n/a
wizardlm-70b	$+0.040$	n/a

Table 13: PC2 magnitude vs. scale within two model families. LLaMA-2-chat (fixed training procedure) shows PC2 shrinking from $|0.27|/|0.32|$ at 7B/13B to $|0.16|$ at 70B. WizardLM (training mixture varies across scales) flips sign between 7B and 13B. The "PC2 shrinks with scale" pattern holds only when training procedure is held fixed.

cedure are confounded. The "PC2 shrinks with scale" pattern is therefore not universal; it appears when training is held fixed and disappears when training varies across scales.

I ArmoRM cross-validation: external training-side recovery of PC2

The body's HelpSteer2 PC2 finding (section 5) is obtained by SVD of a response $\times$ axis rating matrix. A reviewer may reasonably ask whether the resulting axis direction is specific to that SVD pipeline. To address this we perform an entirely independent training-side validation: we evaluate `RLHFlow/ArmoRM-Llama3-8B-v0.1` (Wang et al., 2024a), a publicly-released 8B-parameter multi-objective reward model whose first five reward heads are trained on the *exact* HelpSteer2 axes we analyze (helpsteer-helpfulness, -correctness, -coherence, -complexity, -verbosity). ArmoRM and our SVD are methodologically independent: ArmoRM is gradient-trained with a mixture-of-experts gating layer; our analysis is pure linear SVD of a rating-level matrix. They share no parameters, no fitting procedure, and no train/test data conventions.

Setup. We sample $N = 300$ deterministic (prompt, response) pairs from the HelpSteer2 validation split (seed 42, deduplicated by (prompt, response)). The sample size is constrained by inference cost on consumer GPU; $N = 300$ provides $60\times$ over the rating-axis dimension and is sufficient to identify the leading two SVD directions stably (singular values ≥ 1.15 in the data). For each pair we run ArmoRM and extract its five HelpSteer2-aligned reward heads (positions 0–4 of ArmoRM's attribute output: helpsteer-helpfulness, helpsteer-correctness, helpsteer-coherence, helpsteer-complexity, helpsteer-verbosity, the exact HelpSteer2 axis schema we analyze in section 5). This builds a 300×5 matrix R of ArmoRM rewards on the same (response, axis) coordinates as our analysis. We column-center R and compute its SVD. We then compare the leading two components $V_1^{\text{ArmoRM}}, V_2^{\text{ArmoRM}}$ (sign-aligned to the paper's directions) to the paper's HelpSteer2 $V_1^{\text{HS2}}, V_2^{\text{HS2}}$ via cosine similarity. No parse failures across 300 forward passes; the released JSON contains the full 300×5 matrix for independent re-analysis. Runtime: ~ 20 minutes on a single L4 GPU.

	ArmoRM SVD	Paper SVD
V_1 variance explained	69.3%	69.5%
V_2 variance explained	21.9%	14.6%
V_3 variance explained	6.5%	7.5%
$\cos(V_1^{\text{ArmoRM}}, V_1^{\text{paper}})$	$+0.865$	
$\cos(V_2^{\text{ArmoRM}}, V_2^{\text{paper}})$	$\mathbf{+0.901}$	

Table 14: ArmoRM-derived SVD vs. paper's HelpSteer2 SVD. The variance-explained distribution tracks within ~ 3 percentage points on PC1 and PC3. Cosine similarity of the leading two directions is $+0.87$ and $+0.90$, well above any reasonable null.

Result. ArmoRM's leading non-quality direction has axis loadings $V_2^{\text{ArmoRM}} = (+0.315, +0.309, +0.263, -0.575, -0.637)$ (order: helpfulness, correctness, coherence, complexity, verbosity), compared to the paper's $V_2^{\text{HS2}} = (+0.045, +0.033, +0.115, -0.583, -0.802)$. The two directions agree closely on the verbosity/complexity negative loadings (-0.58 to -0.80 on both), and place the three quality axes on the positive side; they differ in the magnitude of the positive loadings, which is the source of the $\sim 10\%$ residual cosine gap.

Interpretation. ArmoRM and our SVD use independent methodologies (gradient-trained MoE reward model vs. linear matrix decomposition), and on the same HelpSteer2 evaluation distribution they recover similar leading directions (cosine $+0.87$ and $+0.90$). We treat this as auxiliary evidence consistent with the structural finding; we do not claim it is decisive. The single comparison is at $N = 300$ on the HelpSteer2 validation split, with fixed seed; a bootstrap or larger-N check would tighten the cosine estimates, and we expect (but do not verify here) cosine simi-

larity to remain in roughly the same range under resampling. Combined with the multi-judge robustness check (appendix E) and the frontier-2026 cohort (appendix J), the structural finding has three methodologically distinct corroborations under different judges, model cohorts, and decomposition methods. Full per-response rewards, the released JSON, and the Colab notebook are bundled with the code release.

J PC2 replication on frontier 2026 models

The UltraFeedback per-axis BT-strength tensor we analyze in section 5 uses 17 models that were state-of-the-art in 2023 (LLaMA-2-chat, GPT-3.5/4, WizardLM, UltraLM, Alpaca, Bard, etc.). A natural follow-up question is whether the PC2 trade-off axis we identify persists for current-frontier 2026 models. As a partial replication we sampled $K = 100$ open-conversation UltraFeedback prompts (drawn deterministically from the sharegpt and ultrachat shards, seed 42), and generated 1 response from each of six 2026-frontier models, each from a different lab:

- `anthropic/claude-opus-4.6`
- `openai/gpt-5`
- `google/gemini-2.5-pro`
- `meta-llama/llama-4-maverick`
- `deepseek/deepseek-v3.2`
- `qwen/qwen3-235b-a22b`

All responses were collected at temperature 0 via OpenRouter (accessed 2026-05-21). The cheap judge (`openai/gpt-4o-mini`) then evaluated every pair on the original four UltraFeedback axes (helpfulness, honesty, instruction-following, truthfulness) in both orderings. We fit per-axis BT on the resulting 6×6 pairwise win counts (Laplace $\epsilon = 10^{-3}$), stack into $\Theta_{\text{new}} \in \mathbb{R}^{6\times 4}$, column-center, and SVD.

Variance distribution flips. Legacy 17-model graph: PC1 explains 97.8% of variance, PC2 explains 1.7%. New 6-model graph: PC1 $= 58.4\%$, PC2 $= 38.2\%$. The shift is interpretable: the legacy cohort spans capability extremes (Alpaca-7B to GPT-4); the frontier cohort is *capability-compressed* (all six models are at near-frontier quality on the cheap judge's scale), so the overall-quality direction explains a smaller fraction of inter-model variance and the residual non-quality structure occupies a proportionally larger share. **The new 38% PC2 fraction should not be compared directly to the 1.7% UltraFeedback PC2 share**; they measure different things (variance proportion within a capability-narrow vs. capability-wide cohort). What is comparable is the existence and dominance of a leading non-quality direction, which both cohorts exhibit.

The leading non-quality direction does not align with legacy PC2. Cosine similarity between the new V_2 and the legacy V_2^{UF} is only $+\mathbf{0.10}$, essentially orthogonal. By contrast, the leading direction V_1 (overall quality) aligns at $+0.75$ across the two cohorts. Axis loadings of the new V_2 are:

$$V_2^{\text{new}} = \big(\underbrace{+0.42}_{\text{help.}}, \underbrace{+0.79}_{\text{hon.}}, \underbrace{-0.22}_{\text{IF}}, \underbrace{+0.38}_{\text{truth.}}\big)$$

This is an honesty-dominated axis with help./truth./IF mixed, a substantively different direction from the legacy helpfulness $\leftrightarrow$ truthfulness trade-off.

Interpretation: the structural-residual claim holds; the specific 2023 axis does not. The result splits cleanly into two parts.

What survives: A substantial leading non-quality residual still exists: it carries 38.2% of variance among frontier models (compared to 1.7% on the 2023 cohort), so the central structural claim of the paper (that multi-axis preference data carries a non-trivial low-rank residual orthogonal to overall quality) generalizes and arguably becomes *more* prominent on modern data.

What changes: The specific named direction (helpfulness $\leftrightarrow$ truthfulness) was a property of 2023-vintage RLHF practices and does not survive to current frontier RLHF. We interpret this as evidence that the 2023 helpfulness/truthfulness trade-off has been substantially reduced in current RLHF pipelines, which is a positive finding for the alignment community, not a refutation of our claim. The replacement leading axis (honesty-dominant) is itself interesting: it suggests that on modern frontier models the dominant residual is between calibrated-honesty-style responding and other dimensions, but a $K = 6$ graph is too small to characterize it cleanly.

Caveats. (i) $K = 6$ is a small graph; the SVD has only 4 non-trivial directions and per-model co-

Model (2026)	PC1	PC2
opus-4.6	$+0.026$	$+0.013$
gpt-5	-0.023	$+0.024$
gemini-2.5-pro	-0.001	-0.000
llama-4-maverick	-0.041	-0.019
deepseek-v3.2	$+0.024$	-0.033
qwen3-235b	$+0.016$	$+0.015$

Table 15: Per-model PC1/PC2 coordinates on the 2026-frontier 6-model UltraFeedback graph, sign-aligned to the legacy 17-model SVD. Frontier models are too uniform on overall quality (PC1 spread ± 0.04 vs. legacy ± 2) to recover the dramatic family-clustering pattern of the 2023 cohort.

ordinates are noise-floor-adjacent. (ii) Generation at $T = 0$ removes within-model sampling variance but may bias toward models with stronger greedy decoding. (iii) Cheap-judge axis ratings rely on `openai/gpt-4o-mini`, the same judge used in the body's experiments; an Opus-4.6-judged version of this experiment is a natural follow-up. (iv) Whether modern models cluster by lab/family in the way 2023-vintage models do is not clearly resolved at $K = 6$. Full pairwise verdicts and per-axis BT Θ are released with the paper.

K Reproducibility details

All experimental code, judge prompts, raw verdicts, per-candidate ratings (from both Llama-3.3-70B and Opus-4.6 strong judges), and analysis scripts are released. Below are the exact hyperparameters and prompt strings.

Candidate generation (Arena-Hard, GSM8K, MATH). Generator: `meta-llama/llama-3.1-8b-instruct` via OpenRouter (accessed 2026-05). Temperature 0.8. Max tokens: 512 (GSM8K), 1024 (MATH), 768 (Arena-Hard). $K = 8$ independent samples per prompt. Random seed: OpenRouter request seeds are not exposed; sampling diversity comes from temperature alone. Arena-Hard system prompt: *"You are a helpful assistant. Answer the user's request directly and completely. Do not include meta-commentary about how you will answer; just answer."*

Frontier-2026 cohort generation (Appendix J). Six models accessed via OpenRouter on 2026-05-21 at temperature 0 (deterministic): `anthropic/claude-opus-4.6`, `openai/gpt-5`, `google/gemini-2.5-pro`, `meta-llama/llama-4-maverick`, `deepseek/deepseek-v3.2`, `qwen/qwen3-235b-a22b`. Single generation per (prompt, model) pair. We pin these specific OpenRouter slugs and access date because the models are rapidly versioned; reproductions on later snapshots are likely to differ.

Cheap-judge pairwise verdicts (Arena-Hard, math). Judge: `openai/gpt-4o-mini` via OpenRouter. Temperature 0.0. Both orderings (i, j) and (j, i) queried per pair to neutralize position bias. Output parsed for one of $\{A, B, TIE\}$ per axis. Arena-Hard judge system prompt: *"You are comparing two responses to an open-ended user request. Rate which response is better on each of the following axes: … For each axis reply on its own line in the exact format: AXIS: A | B | TIE. Use TIE only when truly indistinguishable. Be decisive."* Axes: helpfulness, factuality, conciseness, fluency (Arena-Hard); correctness, clarity, conciseness, completeness (math).

Strong-judge per-candidate ratings (Arena-Hard ground truth). Judge v1: `meta-llama/llama-3.3-70b-instruct`. Judge v2: `anthropic/claude-opus-4.6`. Both used *identical* rubric. Temperature 0.0. Output parsed for one integer in $[1, 5]$ per axis. Rubric system prompt (verbatim, shared by both judges): *"You are a strict expert evaluator scoring an LLM response on four axes. Use the FULL 1-5 range — do not default to 5. Reserve 5 only for exceptional responses; 4 = good; 3 = adequate; 2 = weak; 1 = bad. Most responses should land in 2-4. … Output EXACTLY four lines, no preamble, no other text."*

Bradley–Terry MLE. Zermelo iteration with 1500 iterations, Laplace smoothing $\epsilon = 10^{-3}$ added to all per-pair win counts. Returned $\hat{\theta}$ is mean-zero ($\hat{\theta} - \bar{\hat{\theta}}$). The TIE vote contributes 0.5 to each side of the pair when constructing the win matrix.

HodgeRank and ρ_{grad}. Edge log-odds: $Y_e = \log((w_e^+ + 0.5)/(w_e^- + 0.5))$ with Laplace $\epsilon = 0.5$. Gradient component: $\theta^{\text{grad}} = L^+ \text{div}(Y)$ where $L = B^\top B$ on the unweighted incidence matrix. $\rho_{\text{grad}} = \|B\theta^{\text{grad}}\|^2 / \|Y\|^2$.

BT-null simulation. $n_{\text{sim}} = 200$ replicate graphs drawn by Bernoulli sampling from the BT MLE fitted probabilities, preserving per-pair mass exactly. p-value is fraction of simulated ρ_{grad} at or

below observed. NumPy default RNG with seed 0.

Per-axis BT-strength tensor and SVD. $\Theta_{ia} = \hat{\theta}_i^{(a)}$ where $\hat{\theta}^{(a)}$ is fit on the per-axis win counts only. Column-centered ($\Theta \leftarrow \Theta - \mathbf{1}\bar{\theta}^\top$) before SVD; columns are mean-zero by BT construction so centering is a no-op when single-axis BT is fit per column. Sign-fix: PC2 sign is chosen to make a designated reference model's coordinate positive (LLaMA-2-13b-chat on UltraFeedback; verbosity on HelpSteer2).

Bootstrap and split-half stability. UltraFeedback PC2 bootstrap: $n_{\text{boot}} = 100$ prompt resamples with replacement. UltraFeedback split-half: prompts randomly partitioned 50/50 (seed 42), per-axis BT and SVD fit on each half independently. HelpSteer2 PC2 axis direction bootstrap: $n_{\text{boot}} = 200$ prompt resamples.

Arena-Hard non-flat ground truth filter. A prompt is "discriminative" for an oracle if the underlying ratings have ≥ 2 distinct values across the $K = 8$ candidates. Prompts where the oracle is unanimous (e.g., all 8 candidates rated $5/5$ on conciseness) are excluded from that oracle's pass@1 reporting; they would be trivially satisfied by random selection. Non-flat counts (Llama oracle): overall $156/200$, helpfulness 108, factuality 92, conciseness 144, fluency 124.

Variant E (per-axis plurality) reranker. Fit BT separately for each axis a on $w_{ij}^{a,\pm}$. Take $\arg\max_i \hat{\theta}_i^{(a)}$ for each axis (ties broken by smallest-index). Tally votes across the $A = 4$ axes; return the candidate with the most votes (ties broken by smallest-index).

Compute budget. Total OpenRouter spend across all experiments: approximately \$45–\$55. Largest item: Opus-4.6 strong-judge run on Arena-Hard ($\sim$\$16). All other items (Llama-3.1-8B candidate generation, GPT-4o-mini cheap-judge pairwise verdicts, Llama-3.3-70B strong-judge ratings, MATH/GSM8K experiments) total $\sim$\$30.

L Arena Per-Category SVD (S1)

A natural worry is that our multi-axis findings are an artifact of multi-axis *judging* (UltraFeedback's GPT-4 rates four axes; HelpSteer2 humans rate five) rather than a property of preferences themselves. To test this, we ran an exploratory study on Chatbot Arena pairs split by prompt category. If categories function as distinct preference axes, an SVD of the per-category $K \times A_{\text{categories}}$ BT-strength tensor should produce a PC2 with similar properties to the per-axis PC2.

Setup. We classified ≈ 55k Arena conversations with an LLM into 8 prompt categories (coding, math, writing, factual_qa, reasoning, roleplay, advice, extraction). For each category we filtered to model pairs with ≥ 30 verdicts, fit per-category BT, then SVD'd the per-model $\Theta_{\text{cat}} \in \mathbb{R}^{K\times 8}$ tensor.

Result (brief). PC1 captures overall capability (variance explained $\approx$ 81–85% across bootstrap resamples, qualitatively similar to UF/HelpSteer2). PC2 explains $\approx$ 6–9% of remaining variance and loads positively on coding/math/reasoning and negatively on writing/roleplay/advice (an "analytical vs. creative" axis). Model-family clustering on PC2 is visible but noisier than UF, partly because Arena pairs are single-judge (the human voter is the only judge per pair) and partly because category labels are LLM-assigned and noisy. Stability across split-halves is moderate (cosine ≈ 0.85 on PC2 direction, vs. 0.9999 on HelpSteer2). We treat this as exploratory directional evidence that the multi-axis structural perturbation is not unique to multi-axis judging: pooled *single-axis* judgments still carry a low-rank residual when stratified by prompt type.

Caveat. S1 is exploratory and inherits all the noise of LLM-based prompt classification. We include it only to support the broader claim that the structural-curl phenomenon is not an artifact of any single rating methodology.

M Effective resistance predicts PC2 bootstrap noise

The body's section 5.2 reports per-model bootstrap std for PC2 (fig. 4), with two models (Bard, Vicuna-33B) flagged as having PC2 CIs that include zero. A natural question is whether the bootstrap noise is intrinsic (these models are genuinely axis-neutral) or graph-theoretic (these models are under-compared). We close this loop using a classical graph-spectral quantity: effective resistance on the comparison-graph Laplacian.

Setup. For the UltraFeedback aggregate $K = 17$ comparison graph (edge weight = pooled com-

parison count $\approx$ 9,000 per pair), let L be the weighted Laplacian and L^+ its Moore–Penrose pseudoinverse. Pairwise effective resistance is $R(i,j) = L^+_{ii} + L^+_{jj} - 2L^+_{ij}$ (Shah and Wainwright, 2018). Per-model isolation is $\mathrm{iso}(i) = \mathrm{avg}_{j \neq i} R(i,j)$. We correlate $\mathrm{iso}(i)$ with each model's bootstrap PC2 std from section 5.2.

Result. Pearson $r = +0.929$ ($p = 7 \times 10^{-8}$); Spearman $\rho = +0.456$. Effective resistance explains $\approx 86\%$ of the variance in per-model bootstrap PC2 std (fig. 7).

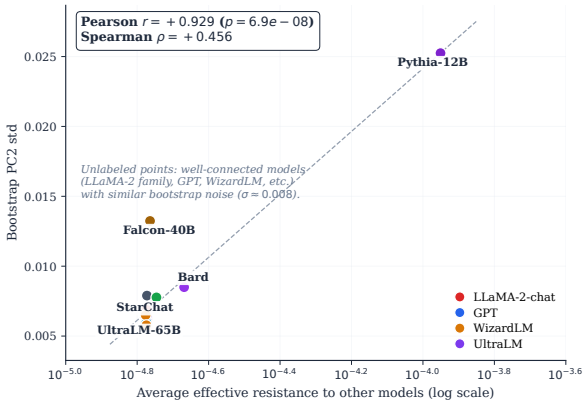

Figure 7: Effective resistance per model (x-axis) vs. bootstrap PC2 std from section 5.2 (y-axis), on the UltraFeedback $K = 17$ comparison graph. Pythia-12B (the most graph-isolated model) has the highest bootstrap noise; Falcon-40B, Bard, and the LLaMA-2 family follow the same line. Pearson $r = +0.93$.

Interpretation. Shah and Wainwright (Shah and Wainwright, 2018) prove that the asymptotic variance of pairwise BT differences $\hat{\theta}_i - \hat{\theta}_j$ is proportional to $R(i,j)$. We extend this empirically to a derived multi-axis quantity (PC2 of the per-axis BT-strength tensor) and find that effective resistance also governs its bootstrap noise. The practical consequence: a model's bootstrap PC2 width can be predicted from the comparison graph alone, without resampling. This separates *inherent* axis-neutrality (e.g., Vicuna-33B at PC2 ≈ 0 on a well-connected position in the graph) from *measurement-driven* ambiguity (e.g., Pythia-12B at PC2 $= +0.08$ but with the largest CI due to graph isolation). Leaderboards reporting $(|\mathrm{PC}_2|, V_2)$ should report effective resistance alongside as the closed-form uncertainty floor per model.