Heterogeneous Judge-Aware Ranking with Sensitivity, Disagreement, and Confidence

Shibo Yu (Presenter), Yingzhou Wang, Yan Chen, Guodong Li, Jin-Hong Du

The University of Hong Kong | Contact: shibo-yu01@connect.hku.hk

NeurIPS 2026 · Sydney

1. Motivation: Why Pooling & Scalar Sensitivity Fail

In multi-judge LLM evaluation, $[N] = \{1, \dots, N\}$ candidate models are compared by $[K] = \{1, \dots, K\}$ heterogeneous evaluators over observed pairs $\Omega := \{(k, i, j) : 1 \le k \le K, 1 \le i < j \le N, n_{kij} > 0\}$.

Example 1: Latent Expertise Reversal

Consider coding responses (A, B, C) where A is correct/terse, B is polished with a fatal bug, and C is incomplete. Judge 1 is a coding expert; Judges 2–3 are general evaluators. Under $S = \gamma\mu^\top + UV^\top$ with $\gamma = 1_3$, $\mu = (0, 1, -1)^\top$:

$$S = \begin{pmatrix} 1 & 1/2 & -3/2 \\ -1/2 & 5/4 & -3/4 \\ -1/2 & 5/4 & -3/4 \end{pmatrix} \Longrightarrow \begin{cases} \text{Expert (Judge 1): } A \succ B \succ C \\ \text{General (Judges 2,3): } B \succ A \succ C \end{cases}$$

Under equal sampling, the pooled win probability is:

$$\bar{p}_{AB} = \frac{1}{3}\{\sigma(1/2) + 2\sigma(-7/4)\} \approx 0.306 < \frac{1}{2}.$$

Consequence: Both Pooled BTL and sensitivity-only JA-Ranking ($\gamma = 1_3$) rank $B \succ A$ with certainty as $n \to \infty$. **HJA** separates consensus μ from the orthogonal expertise interaction $UV^\top$.

2. Heterogeneous Judge-Aware (HJA) Probabilistic Model

Conditional on the row-centered score matrix $S \in \mathbb{R}^{K\times N}$ ($S1_N = 0_K$), pairwise win counts $Y_{kij} \in \{0, \dots, n_{kij}\}$ follow:

$$Y_{kij} \sim \text{Binomial}(n_{kij}, p_{kij}), \qquad \text{logit}(p_{kij}) = S_{ki} - S_{kj}, \quad (k, i, j) \in \Omega.$$

We decompose S into **consensus**, **consensus sensitivity**, and **rank-r residual disagreement**:

$$\boxed{S = \gamma\mu^\top + UV^\top \iff \eta_{kij}(\theta) = \gamma_k(\mu_i - \mu_j) + U_{k,:}(V_{i,:} - V_{j,:})^\top}$$

- $\mu \in \mathbb{R}^N$: Shared cross-judge **consensus ranking direction**.
- $\gamma \in \mathbb{R}^K$: Judge-specific **consensus sensitivity** ($\gamma_k \in \mathbb{R}$ unconstrained; $\gamma_k < 0$ detects anti-aligned/adversarial judges).
- $U \in \mathbb{R}^{K\times r}, V \in \mathbb{R}^{N\times r}$: Low-rank **residual preference disagreement** ($r \le \min\{K - 1, N - 2\}$).

3. Canonical Identifiability Geometry

To make (γ, μ, U, V) statistically identifiable Inferential targets rather than an arbitrary bilinear factorization, we constrain $\theta \in \Theta_r$:

Condition 1 (Canonical Normalization Space Θ_r)

1. **Centering:** $1_N^\top\mu = 0$ and $V^\top 1_N = 0_r$ *(removes additive BTL invariance)*
2. **Scaling:** $1_K^\top U = 0_r^\top$ and $1_K^\top\gamma = K$ *(anchors μ as row-average score)*
3. **Orthogonality:** $\mu^\top V = 0_r^\top$ *(separates disagreement from consensus)*
4. **Subspace Anchoring:** $N^{-1}V^\top V = I_r$ and $K^{-1}U^\top U = D = \text{diag}(d_1 > \cdots > d_r > 0)$.

Under Condition 1, μ, γ, and the residual subspace $\tilde{S}$ are closed-form projections of S:

$$\mu^\top = \frac{1}{K}1_K^\top S, \qquad \gamma = \frac{S\mu}{\|\mu\|_2^2} = \frac{KSS^\top 1_K}{\|1_K^\top S\|_2^2}, \qquad \tilde{S} := SP^\perp_{S^\top 1_K} = UV^\top.$$

Proposition 2 (Global Model Identification)

Suppose each judge comparison graph $\mathcal{G}_k = ([N], \mathcal{E}_k)$ is connected, $S1_N = 0_K$, $S^\top 1_K \neq 0_N$, and $\tilde{S}$ has r distinct positive singular values $\sigma_j = \sqrt{KNd_j}$ (Condition 2). Then:

1. S is uniquely identified from the generative BTL probabilities $\{p_{kij}\}_{(k,i,j)\in\Omega}$.
2. There exists a **unique** $\theta = (\gamma, \mu, U, V) \in \Theta_r$ such that $S = \gamma\mu^\top + UV^\top$, up to column-wise sign flips $(\pm u_j, \pm v_j)$.

4. Maximum Likelihood & Fixed-Panel Asymptotics

Constrained MLE. For $\bar{Y}_{kij} = Y_{kij}/n_{kij}$, minimize the negative log-likelihood over the canonical parameter space Θ_r:

$$\hat{\theta} = (\hat{\gamma}, \hat{\mu}, \hat{U}, \hat{V}) \in \arg\min_{\theta\in\Theta_r} \mathcal{L}_n(\theta), \qquad \mathcal{L}_n(\theta) = \sum_{(k,i,j)\in\Omega} n_{kij}\Big[\log\{1 + e^{\eta_{kij}(\theta)}\} - \bar{Y}_{kij}\eta_{kij}(\theta)\Big].$$

Fixed-panel regime. Let K, N be fixed, $n = \Sigma\, n_{kij} \to \infty$, and $n_{kij}/n \to w_{kij}$ (Assumption A4). Each active comparison graph $\mathcal{G}_k^w = ([N], \mathcal{E}_k^w)$ is connected, with $\Omega_w = \{(k, i, j) : w_{kij} > 0\}$ (Assumption A5).

Proposition (Graph-Induced Fisher Information Positivity)

Under A4–A5, canonical normalization, and $\mu^* \neq 0$, fix the column signs of (U, V) locally. The score map $\Phi(\theta) = \gamma\mu^\top + UV^\top$ has an injective differential on $T_{\theta^*}\Theta_r$. For $H = [h_1, \dots, h_K]^\top \neq 0$ with $H1_N = 0$ and $\omega_{kij} = w_{kij}p^*_{kij}(1-p^*_{kij})$,

$$\langle H, \mathcal{I}_{\mathcal{S}}(S^*)H\rangle = \sum_{k=1}^K \sum_{\{i,j\}\in\mathcal{E}_k^w} \omega_{kij}(H_{ki} - H_{kj})^2 = \sum_{k=1}^K h_k^\top L_k h_k > 0.$$

Connectedness gives $\ker L_k = \text{span}\{1_N\}$. Hence, with J the Jacobian of Φ in local identified coordinates,

$$\boxed{\mathcal{I}(\theta^*) = J^\top\mathcal{I}_{\mathcal{S}}(S^*)J \succ 0.}$$

Thus graph connectivity and the identifying geometry jointly guarantee nonsingular information for the structured parameters.

Theorem 4 (Fixed-Rank Asymptotic Normality & Wald Inference)

Under Assumptions A4–A5 with $\mu^* \neq 0$, $\hat{\theta} \xrightarrow{p} \theta^*$ and, in local identified coordinates,

$$\sqrt{n}(\hat{\theta} - \theta^*) \xrightarrow{d} \mathcal{N}(0, \mathcal{I}(\theta^*)^{-1}).$$

For any smooth functional $q : \Theta_r \to \mathbb{R}$ (e.g., $\mu_i - \mu_j$, $S_{ki} - S_{kj}$, γ_k, or p_{kij}),

$$\sqrt{n}\{q(\hat{\theta}) - q(\theta^*)\} \xrightarrow{d} \mathcal{N}(0, \nabla_\theta q(\theta^*)^\top\mathcal{I}(\theta^*)^{-1}\nabla_\theta q(\theta^*)).$$

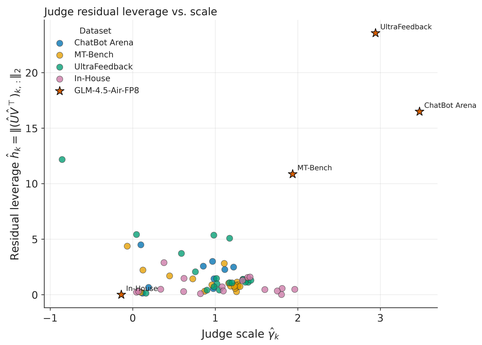
Consistent Rank Selection (Theorem C.1): Minimizing $\text{BIC}(r) = 2\mathcal{L}_n(\hat{\theta}_r) + r(K + N - r - 3)\log n$ satisfies $\mathbb{P}(\arg\min_r \text{BIC}(r) = r^*) \to 1$.

5. Diagnostic Analysis: Sensitivity vs. Residual Leverage

To distinguish sharp consensus alignment from idiosyncratic bias, we pair $\hat{\gamma}_k$ with **residual leverage**:

$$\hat{h}_k = \|(\hat{U}\hat{V}^\top)_{k,:}\|_2, \qquad \hat{\rho}_k = \frac{\|(\hat{U}\hat{V}^\top)_{k,:}\|_2}{|\hat{\gamma}_k|\|\hat{\mu}\|_2 + \epsilon}.$$

- **High-Leverage Outlier Detection:** On Chatbot Arena, `zai-org/GLM-4.5-Air-FP8` has both large $\hat{\gamma}_k$ and extreme residual leverage $\hat{h}_k$, acting as a high-sensitivity, high-bias judge. Removing it stabilizes $\hat{\mu}$ and unmasks moderate-sensitivity judges (e.g., `Llama-4-Scout-17B-16E-In`).
- **Judge-Model Affinity Clustering:** Biclustering $\hat{S} = \hat{\gamma}\hat{\mu}^\top + \hat{U}\hat{V}^\top$ localizes structured family-level preference blocks across evaluators and candidate LLMs.

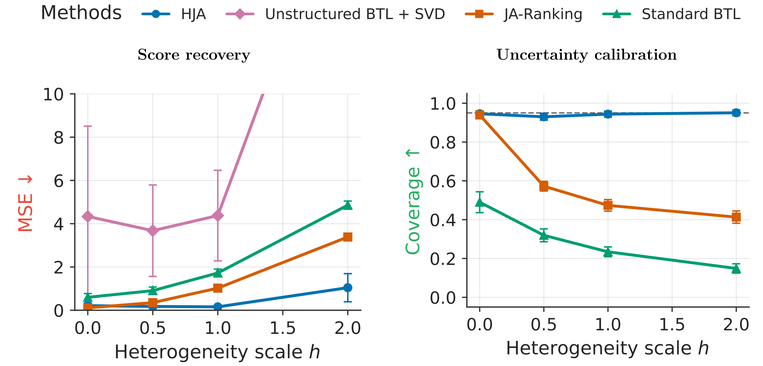


Residual leverage versus consensus sensitivity across four datasets; stars mark GLM-4.5-Air-FP8.

6. Synthetic & Misspecified Hidden-Expertise Recovery

- **Well-Specified Regime** ($N = 8, K = 4, r = 1$)**:** Across varying comparison budgets $n_{\text{cmp}} \in [400, 3000]$ and heterogeneity scales $h \in \{0, 0.5, 1, 2\}$, HJA achieves the lowest Score MSE, highest Consensus Spearman/NDCG@N, and **95% empirical coverage**, while Pooled BTL and JA-Ranking suffer severe undercoverage as h grows.
- **Misspecified Latent-Domain Mixture** ($N = 12, K = 9$)**:** With 6 math-expert and 3 coding-expert judges ($Y_{kij} \mid D = d \sim \text{Bernoulli}[\sigma(\gamma_{kd}(a_{id} - a_{jd}))]$, domain d hidden), pooled win probabilities reverse the true order ($p_{kAB} = 0.49 < 0.5$). Across 50 runs:

Metric (50 Seeds)	HJA (Ours)	Sensitivity-Only	Pooled BTL
Correct Pairwise Order ($A \succ B$)	**48 / 50**	35 / 50	0 / 50
Held-Out Prediction Accuracy	**0.714**	0.542	0.551

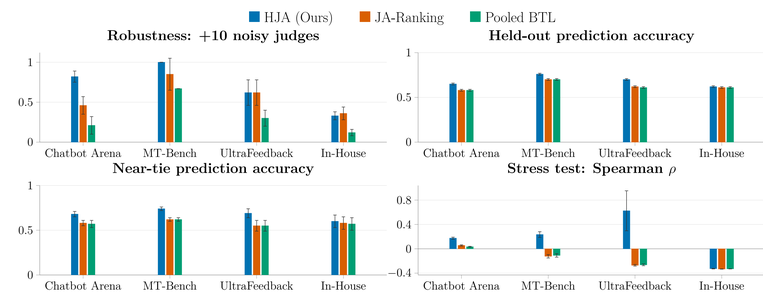


Selected panels from the well-specified synthetic experiment. As heterogeneity grows, HJA maintains low score MSE and near-nominal 95% interval coverage; sensitivity-only and pooled BTL under-cover.

7. Real-Data Benchmarks & Stress Tests

Chatbot Arena ($N = 20, K = 10$), **MT-Bench** ($N = 6, K = 20$), **UltraFeedback** ($N = 17, K = 20$), and **In-House** ($N = 45, K = 18$).

Test interpretation. Robustness measures ranking accuracy against the original ranking after adding 10 noisy judges. Held-out and near-tie accuracy measure prediction on unseen comparisons and nearly tied pairs. Stress-test ρ measures agreement with the clean ranking after injecting a biased source matching the clean panel's comparison volume; negative ρ indicates reversal. Higher is better for all metrics.



Bars show the reported means; error bars show ±1 standard deviation. Each cluster is one dataset.

Selected references

Bradley & Terry. Rank Analysis of Incomplete Block Designs: I. The Method of Paired Comparisons. *Biometrika*, 1952.

Park et al. Preference Completion: Large-scale Collaborative Ranking from Pairwise Comparisons. *ICML*, 2015.

Xu et al. A Judge-Aware Ranking Framework for Evaluating Large Language Models without Ground Truth. *ICML*, 2026.

Han, Tang & Xu. Statistical Inference for Pairwise Comparison Models. *arXiv*, 2025.

Zheng et al. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. *NeurIPS*, 2023.

Cui et al. UltraFeedback: Boosting Language Models with Scaled AI Feedback. *ICML*, 2024.