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

Shibo Yu
The University of Hong Kong
shibo-yu01@connect.hku.hk

Yingzhou Wang
The University of Hong Kong
wyingz@connect.hku.hk

Yan Chen
The University of Hong Kong
yanc25@hku.hk

Guodong Li
The University of Hong Kong
gdli@hku.hk

Jin-Hong Du
The University of Hong Kong
jinhongd@hku.hk

Abstract

Pairwise comparisons from multiple judges are central to large language model evaluation and preference modeling, yet standard ranking pipelines often pool judgments into a single score vector, treating systematic judge disagreement as noise. We propose *Heterogeneous Judge-Aware (HJA) ranking*, a structured multi-judge ranking framework that separates consensus ranking, judge-specific sensitivity to consensus, and residual preference disagreement. HJA thereby treats ranking, judge sensitivity, and structured disagreement as separate inferential targets. We establish conditions under which this decomposition is identifiable and develop an anchored alternating algorithm that preserves the identifying geometry. For confidence quantification, we study a fixed-panel repeated-comparison regime in which the judge panel may remain fixed or modest while information grows through repeated judgments. This yields uncertainty statements for consensus and judge-specific ranking contrasts, sensitivity parameters, pairwise probabilities, and summaries of residual disagreement. Experiments on synthetic and real multi-judge comparison data show that HJA improves recovery, robustness, uncertainty calibration, and near-tie performance relative to pooled and sensitivity-only baselines. The fitted model also provides diagnostics for judge disagreement and model-affinity patterns, giving a statistically grounded framework for ranking under heterogeneous comparative judgments.

1 Introduction

Comparative judgments have become a standard interface for evaluating and aligning modern language models. Pairwise preferences are used to build leaderboards, benchmark systems, and provide supervision for post-training and preference optimization [1–4]. Increasingly, however, these judgments are not produced by a single homogeneous evaluator, but by panels of human annotators, model-based judges, prompting protocols, or evaluation settings applied to the same collection of candidate outputs [5–8]. In such settings, disagreement among judges is not merely exchangeable noise around one latent ranking. Recent evidence shows substantial variation across judge models, languages, tasks, and evaluation protocols, with especially large consequences for subjective evaluations and near-tie comparisons [9–12]. Thus, moving from one judge to many does not simply increase sample size for a classical ranking problem; it changes the object of inference from a single pooled ordering to a structured multi-judge ranking problem.

To address this problem, we develop a heterogeneous judge-aware ranking framework that treats multi-judge comparisons as evidence for several distinct ranking targets rather than as noisy votes for one pooled ordering. As summarized in Figure 1, the framework separates a consensus ranking

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

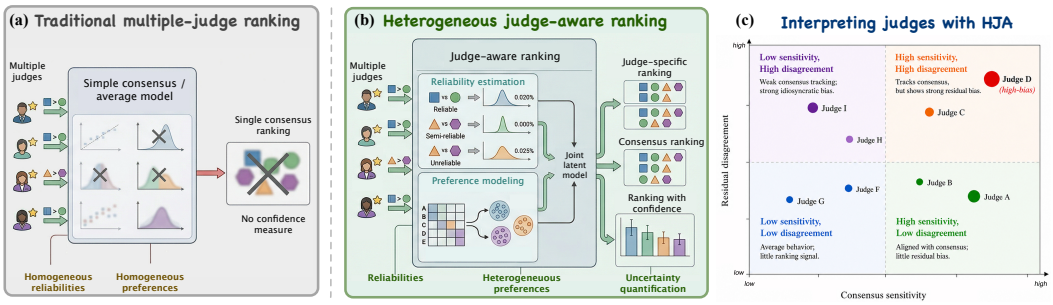


Figure 1: Overview of Heterogeneous Judge-Aware (HJA) ranking. Unlike pooled multi-judge ranking, which collapses heterogeneous judgments into a single consensus score, HJA models judge-specific latent scores through the structured decomposition. This yields consensus and judge-specific rankings, sensitivity assessment, structured disagreement analysis, and uncertainty quantification.

shared across judges, judge-specific sensitivity to that consensus, and structured residual disagreement across judges and items. This distinction is important because two judges may differ not only in how reliably they track the common ranking signal, but also in which types of items or responses they systematically favor. The resulting model supports uncertainty quantification for consensus scores, judge-specific comparisons, consensus sensitivity, and latent disagreement patterns.

A key technical distinction is that the low-rank structure is not used as an arbitrary latent preference space. As formalized in Section 4, HJA decomposes the judge-specific Bradley–Terry–Luce (BTL) score matrix as $S = \gamma\mu^\top + UV^\top$, where μ is the row-average consensus direction, γ_k is judge k's loading on that direction, and $UV^\top$ is constrained to represent residual disagreement orthogonal to consensus. Thus γ_k measures sensitivity to the shared ranking signal, rather than global judge quality, while the low-rank component captures structured deviations after consensus has been removed. The resulting targets are therefore not merely an improved predictor of pairwise preferences, but the identifiable components of this canonical decomposition.

The main technical challenge is to make this separation statistically well-defined, estimable, and inferentially useful. We first show that the consensus component, judge-specific consensus sensitivities, and structured disagreement subspace are identifiable under explicit normalization and rank conditions (Proposition 2). We then estimate the model by constrained maximum likelihood and develop an anchored proximal alternating algorithm that preserves the identifying geometry of the decomposition while enjoying a local linear convergence guarantee (Algorithm 1 and Theorem 3). For inference, we study a repeated-comparison asymptotic regime in which the number of judges may remain fixed or modest while information grows through repeated comparisons, and establish asymptotic normality for the constrained estimator (Theorem 4). This yields confidence intervals for smooth ranking targets, including consensus score contrasts, judge-specific score contrasts, consensus sensitivity, pairwise preference probabilities, and summaries of latent disagreement.

Empirically, we evaluate HJA through synthetic recovery studies and real-data benchmarks (Section 6.1). Synthetic experiments show that HJA recovers consensus rankings, judge-specific behavior, structured heterogeneity, and uncertainty calibration as comparison information increases, while misspecified baselines degrade under stronger heterogeneity. Across four real LLM-evaluation datasets, HJA improves robustness to noisy judges, held-out pairwise prediction, and near-tie performance, especially when judge disagreement is consequential. Beyond predictive benchmarking, we use HJA as a diagnostic tool to report judge-specific rankings with confidence intervals, identify high-leverage judges, and localize structured judge-model affinity patterns, including possible self- or family-affinity effects without treating them as universal findings (Section 6.2).

2 Related Work

Pairwise ranking, annotator heterogeneity, and judge-aware aggregation. Our starting point is the BTL model for ranking from pairwise comparisons [13, 14]. Beyond the homogeneous BTL setting, prior work in crowdsourcing and LLM evaluation shows that aggregation should account for annotator-specific sensitivity, calibration, or sharpness when comparisons are collected from multiple annotators [5, 6]. Motivated by modern multi-judge evaluation settings, we allow disagreement to involve structured preference heterogeneity beyond scalar sensitivity.

LLM-as-a-judge and evaluator disagreement. Pairwise LLM judging became a practical standard through benchmarks and platforms such as MT-Bench and Chatbot Arena [1, 15], and early evidence suggested that strong LLM judges could sometimes serve as useful proxies for human evaluation [16]. More recent work has weakened that optimistic view. Large-scale empirical studies report substantial variation across tasks and benchmarks [9]; multilingual evaluations show poor cross-language stability [10]; and rating-indeterminacy analyses argue that collapsing multiple defensible judgments into a single forced label can itself introduce evaluation bias [11]. Near-tie settings are especially fragile, since small prompt or decoding changes can induce large changes in pairwise outcomes and hence in final leaderboards [12]. These observations suggest that disagreement among judges should be explicitly modeled, rather than averaged away as if all judges were interchangeable.

Judge-aware ranking and uncertainty quantification. JA-Ranking [17] introduces judge-specific discrimination parameters for ranking LLM outputs without ground truth, capturing scalar sensitivity variation across judges. In our notation, this is the sensitivity-only structure $S = \gamma\mu^\top$, the $r = 0$ special case of HJA. Collaborative ranking and preference-completion methods instead learn structured heterogeneous preferences from pairwise data [18, 19], with related work imposing low-rank structure on pairwise preference probabilities or heterogeneous preference populations [20–22]. These approaches show that preference heterogeneity can be low-dimensional, but typically treat the individualized or low-rank preference object as the primary target. HJA uses low-rank structure for a different inferential purpose: it canonically decomposes the judge-specific BTL score matrix into a row-average consensus direction, judge-specific sensitivity to that direction, and an orthogonal residual disagreement component. This yields separate targets for consensus ranking, judge-specific ranking, and latent disagreement, with fixed-panel uncertainty quantification for identifiable components rather than only pairwise prediction. Complementary work on comparative uncertainty argues that evaluation pipelines should retain confidence information rather than reduce judgments to deterministic verdicts [23, 24]; HJA connects this perspective to structured multi-judge ranking with fixed or modest judge panels. See Section A for a detailed comparison.

3 Problem Setup

Pairwise judgments from multiple judges. We consider repeated pairwise comparison settings with multiple judges. Let $[N] = \{1, \dots, N\}$ and $[K] = \{1, \dots, K\}$ index the items and the judges. For each judge $k \in [K]$ and unordered item pair (i, j) with $1 \le i < j \le N$, n_{kij} denotes the number of repeated comparisons between items i and j, and $Y_{kij} \in \{0, \dots, n_{kij}\}$ denotes the number of times judge k prefers item i to item j. In LLM evaluation and preference modeling, pairwise comparisons are routinely used to benchmark systems, rank models without reference answers, and train preference-aligned models [1–4, 8]. Write

$$\Omega := \{(k, i, j) : 1 \le k \le K,\ 1 \le i < j \le N,\ n_{kij} > 0\} \tag{1}$$

for the set of observed judge-item-pair triples. Throughout, statistical information grows through repeated judgments on observed triples, while the number of judges K may remain fixed or modest.

Motivation. A pooled ranking model treats all judges as interchangeable and collapses their responses into a single latent score vector. But this is often too restrictive in multi-judge evaluation. See, e.g., crowdsourced pairwise ranking and MT evaluation, where annotator-specific sensitivity or calibration affects aggregation [5, 6], and is amplified in LLM-as-a-judge settings with judge, language, and protocol sensitivity [9, 10, 12]. Judges differ not only in how strongly they align with an overall ranking signal, but also in systematic ways that cannot be explained by a common score plus independent noise. For example, as will be shown in Section 4.1, the well-documented difference in domain expertise [25, 26] can easily confuse the pooled-based models. These considerations suggest that the target of inference should not be a single pooled ranking alone, but a decomposition that preserves a consensus ranking while allowing structured judge-specific departures from it.

Consensus ranking with sensitivity, preference heterogeneity, and confidence. A judge-aware ranking model should do more than output a single pooled ordering of the items. In many evaluation settings, the substantive questions are not only which items rank highest on average, but also which judges track the shared ranking signal most strongly, where systematic preference disagreement remains, and how confident we should be in the ranking conclusions.

Accordingly, we seek a representation that separates three statistical targets: (i) *consensus ranking*, which summarizes the ordering most strongly supported across judges; (ii) *consensus sensi-*

tivity, which is judge-specific sensitivity to the shared consensus ranking direction and measures how strongly a judge tracks the common ranking signal, rather than the judge's overall quality or calibration across all possible preference dimensions; and (iii) *structured preference heterogeneity*, which allows judges to differ not only in sharpness or noise level, but also in systematic preferences that cannot be explained by a scalar rescaling alone. This perspective is aligned with collaborative ranking and preference-completion models, which treat individualized preferences as structured low-dimensional signals rather than independent noise [18, 19, 27].

In addition to recovering these latent components, the model should support confidence statements for the inferential outputs they induce. In particular, we would like to quantify uncertainty in the consensus ranking, in judge-specific comparisons, and in the estimated sensitivity and disagreement structure. This is especially important in near-tie regimes, where small perturbations in judgments can produce large changes in the final ordering [11, 12]. These considerations motivate a structured latent-score model that separates shared ranking signal from judge-specific alignment and residual disagreement. The next section introduces such a decomposition.

4 Heterogeneous Judge-Aware Ranking Model

4.1 Judge-specific latent variable model

We model each judge as having a latent score vector over the N items, with pairwise judgments determined by differences in those scores. Let $S \in \mathbb{R}^{K\times N}$ denote the score matrix where the kth row $S_{k,:}$ represents the score vector for judge k. Conditional on S, the counts $\{Y_{kij} : (k,i,j) \in \Omega\}$ are assumed to be independent and follow Bradley–Terry–Luce (BTL) model [13, 28]:

$$Y_{kij} \sim \text{Binomial}(n_{kij}, p_{kij}), \qquad \text{logit}(p_{kij}) = S_{ki} - S_{kj}, \qquad (k,i,j) \in \Omega. \tag{2}$$

Equivalently, $p_{kij} = \exp(S_{ki})/\{\exp(S_{ki})+\exp(S_{kj})\}$. For each judge k, let $\mathcal{G}_k = ([N], \mathcal{E}_k)$ denote the comparison graph with edge set $\mathcal{E}_k = \{\{i,j\} : (k,i,j) \in \Omega\}$. For the main development, we assume that $\mathcal{G}_k$ is connected for every k, as is standard in BTL estimation [14]. This condition implies that each judge-specific score vector is identifiable up to an additive constant from that judge's observed comparisons. Together with the row-centering constraint

$$S1_N = 0_K, \tag{3}$$

this removes the usual additive invariance of the BTL model, so the collection $\{p_{kij} : (k,i,j) \in \Omega\}$ uniquely identifies S. We allow for disconnected individual comparison graphs in the partial sampling regime of Section D.4, under additional identification constraints for the latent score matrix.

To separate shared ranking information from structured disagreement across judges, we model the latent score matrix through a consensus-plus-heterogeneity decomposition:

$$S = \gamma\mu^\top + UV^\top, \tag{4}$$

where $\mu \in \mathbb{R}^N$ is the consensus item score vector, $\gamma \in \mathbb{R}^K$ collects judge-specific consensus-sensitivity parameters, $U \in \mathbb{R}^{K\times r}$ contains judge loadings on latent disagreement directions, and $V \in \mathbb{R}^{N\times r}$ contains the corresponding item coordinates. Equivalently, one has

$$\text{logit}(p_{kij}) = \gamma_k(\mu_i - \mu_j) + U_{k,:}(V_{i,:} - V_{j,:})^\top, \qquad (k,i,j) \in \Omega. \tag{5}$$

This decomposition separates two distinct sources of variation across judges. The term $\gamma_k(\mu_i - \mu_j)$ captures how strongly judge k responds to the common consensus signal, and therefore determines how sharply judge k discriminates along the shared ranking direction μ. In the unlikely case where $\gamma_k < 0$, the judge k is anti-aligned with the consensus direction or may be *malicious* [5, 29]. The residual term $U_{k,:}(V_{i,:} - V_{j,:})^\top$ captures systematic departures from that consensus that cannot be explained by a scalar rescaling alone. As a result, the model distinguishes judges who differ mainly in sharpness from judges who exhibit structured preference heterogeneity across items.

This distinction is particularly useful in LLM-as-a-judge settings. Some judges may be weakly aligned with the common ordering and therefore behave as comparatively noisy evaluators, while others may display systematic deviations tied to answer style, language, task domain, or other latent attributes. Under (4), the consensus vector μ represents the population-level ranking target, whereas $UV^\top$ encodes interpretable disagreement structure around that target.

We also offer an intuitive example of latent expertise to exemplify our notion of heterogeneity, and to show why simpler models fail in this common phenomenon.

Example 1 (Latent expertise under HJA). *Consider an example of coding skill evaluation with three responses A, B, C, where A is technically correct but terse, B is polished but contains a subtle fatal bug, and C is incomplete. Judge 1 is a coding expert, whereas judges 2 and 3 are general-purpose evaluators. Suppose the HJA model is correctly specified with judgment-specific preference given by*

$$\Pr_k(i \succ j) = \sigma(S_{ki} - S_{kj}), \qquad S = \gamma\mu^\top + UV^\top, \tag{6}$$

where S_{ki} denotes judge k's latent BTL score for response i, and "$\succ$" denotes preference over. With coordinates ordered as (A, B, C), set $\gamma = 1_3$, $\mu = (0, 1, -1)^\top$, $U = (\sqrt{2}/2, -\sqrt{2}/4, -\sqrt{2}/4)^\top$, $V = (\sqrt{2}, -\sqrt{2}/2, -\sqrt{2}/2)^\top$. The resulting score matrix is

$$S = \begin{pmatrix} 1 & 1/2 & -3/2 \\ -1/2 & 5/4 & -3/4 \\ -1/2 & 5/4 & -3/4 \end{pmatrix}.$$

Thus, the coding expert ranks $A \succ B \succ C$, whereas both general-purpose judges rank $B \succ A \succ C$. Under equal sampling across judges, the pooled probability that A defeats B is

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

Consequently, the population pooled-BTL fit ranks $B \succ A$, and increasing the number of comparisons only makes this misleading conclusion more certain for an expertise-conditioned coding target. Since $\gamma = 1_3$, a sensitivity-only model cannot represent the reversal either, because the conflict arises entirely from the latent expertise–response interaction $UV^\top$. HJA instead separates the panel consensus $B \succ A \succ C$ from the judge-specific conclusion $A \succ B$, thereby exposing the disagreement that pooling removes. Our model in effect abstracts this type of phenomenon.

Another example which can serve as an extension is to model the domain expertise d in another dimension, when domain-specific comparisons are available, rather than aggregate all comparisons into a marginal judge-by-candidate score matrix. For example, one may have a mixture model

$$\mathrm{logit}(p_{kijd}) = \gamma_{kd}(\mu_{id} - \mu_{jd}) + \textstyle\sum_{\ell=1}^{r} U_{k\ell} W_{d\ell} (V_{i\ell} - V_{j\ell}), \tag{7}$$

where $W_{d\ell}$ links latent direction ℓ to a domain such as coding or mathematics. The resulting judge-by-candidate-by-domain tensor would distinguish judges reliable in particular domains from candidate LLMs strong in those domains. Section 6.1(B) and details in Section E.6 showcase our model's performance under such misspecified data, when domain labels are hidden from the model.

4.2 Identification of latent components

The decomposition in (4) introduces interpretable latent components for consensus, consensus sensitivity, and structured heterogeneity, but the factorization is not identified by the likelihood alone: many parameter tuples can generate the same row-centered score matrix S. The role of the following constraints is therefore not merely to choose a convenient scale, rotation, or centering convention. They define the canonical representative in which μ is the cross-judge consensus direction, γ measures judge-specific sensitivity to that direction, and $UV^\top$ represents residual disagreement.

Condition 1 (Normalization). *The parameters (γ, μ, U, V) satisfy:*

- *(i) Centering: $1_N^\top \mu = 0$ and $V^\top 1_N = 0_r$.*
- *(ii) Scaling: $1_K^\top U = 0_r^\top$ and $1_K^\top \gamma = K$.*
- *(iii) Orthogonality: $\mu^\top V = 0_r^\top$.*
- *(iv) Subspace anchoring: $N^{-1} V^\top V = I_r$ and $K^{-1} U^\top U = D$ for some diagonal matrix D with strictly decreasing positive diagonal entries.*

Condition 1(i)-(ii) fix a canonical representative of the decomposition in (4). The centering constraints remove the additive indeterminacy inherited from the BTL model and align the decomposition with the row-centering normalization $S1_N = 0_K$. The scaling constraints then calibrate the consensus component so that μ represents the cross-judge consensus direction and γ measures judge-specific alignment with that direction. Indeed, under Condition 1(ii), if $S^\top 1_K \neq 0_N$, then

$$1_K^\top S/K = 1_K^\top \gamma \mu^\top / K + 1_K^\top U V^\top / K = \mu^\top, \qquad \gamma = S\mu / \|\mu\|_2^2 = KSS^\top 1_K / \|1_K^\top S\|_2^2.$$

Thus μ is the row-average BTL score direction induced by the judge-specific score matrix. Correspondingly, γ_k is the consensus sensitivity of judge k's loading on this consensus direction.

Condition 1(iii)-(iv) separate consensus from heterogeneity and anchor the residual factorization. The constraint $\mu^\top V = 0_r^\top$ requires the heterogeneous item directions to be orthogonal to the consensus direction, while the subspace-anchoring condition removes rotational ambiguity up to column-wise sign changes. Define the heterogeneous component of the latent score matrix:

$$\widetilde{S} := SP^{\perp}_{S^\top 1_K}, \tag{8}$$

where $P^{\perp}_A$ denotes the projection matrix onto the orthogonal complement of A's column space. In the degenerate case $S^\top 1_K = 0_N$, the induced model reduces to pure heterogeneity.

Accordingly, for a fixed rank $r \le R_{\max} := \min\{K-1, N-2\}$, we define

$$\Theta_r = \left\{(\gamma, \mu, U, V) : \mu \in \mathbb{R}^N,\ \gamma \in \mathbb{R}^K,\ U \in \mathbb{R}^{K\times r},\ V \in \mathbb{R}^{N\times r},\ \text{Condition 1 holds}\right\}.$$

While Condition 1 specifies a normalized parameterization, it does not guarantee that an arbitrary centered score matrix S admits such a decomposition. To recover admissible (γ, μ, U, V) uniquely from S, we therefore impose regularity conditions on S and its heterogeneous component $\widetilde{S}$.

Condition 2. *The cross-judge aggregate score vector of S is nonzero, i.e., $S^\top 1_K \neq 0_N$, and the heterogeneous score matrix $\widetilde{S}$ has r distinct positive singular values.*

The next proposition formalizes the identifiability of the model. First, the BTL comparison law identifies the row-centered judge-specific score matrix S. Second, Conditions 1 and 2 recover from S the canonical consensus-plus-heterogeneity decomposition, uniquely up to column-wise sign changes in (U, V). Compared to generic low-rank preference modeling [20–22], the latent factors in HJA are therefore not introduced merely as a flexible approximation, but are identifiable components with distinct meanings as consensus, consensus sensitivity, and residual disagreement.

Proposition 2 (Identification)**.** *Consider the probabilistic preference model (2) subject to (3). Then,*

(i) (BTL identification) The latent score matrix S can be identified from the generative probability distribution of pairwise comparisons in (2).
(ii) (Parameter identification) If Condition 2 holds for S, then there exists $(\gamma, \mu, U, V) \in \Theta_r$ such that (4) holds. Such existence is unique up to some sign changes in the columns of U and V.

Proposition 2 shows that the model parameters correspond to uniquely defined features of the data-generating mechanism, rather than to an arbitrary factorization of S. The only residual ambiguity can be eliminated via some sign conventions. This identification result provides the foundation for the constrained estimation and uncertainty quantification procedures developed in the next section.

5 Estimation and Uncertainty Quantification

5.1 Constrained maximum likelihood estimation

We first consider estimation for a fixed heterogeneity rank $r \le R_{\max} := \min\{K-1, N-2\}$, with practical rank-selection criteria discussed in Section C.1. When the heterogeneity components are negligible or unnecessary, $r = 0$ may be selected, and Section C.1.2 shows our rank selection and inferential procedures can indeed spot the case where a simpler model is preferred.

Let $\theta = (\gamma, \mu, U, V) \in \Theta_r$. For each observed triple $(k, i, j) \in \Omega$, define $\eta_{kij}(\theta) = \gamma_k(\mu_i - \mu_j) + U_{k,:}(V_{i,:} - V_{j,:})^\top$. Dropping constants independent of θ, the negative log-likelihood is

$$\mathcal{L}_n(\theta) := \textstyle\sum_{(k,i,j)\in\Omega} n_{kij}\big[-\bar{Y}_{kij}\eta_{kij}(\theta) + \log\{1 + \exp(\eta_{kij}(\theta))\}\big], \tag{9}$$

where $\bar{Y}_{kij} = Y_{kij}/n_{kij}$. When $r > 0$, we estimate the identifiable components of the HJA model using the constrained maximum-likelihood estimator (MLE)

$$\widehat{\theta} = (\widehat{\gamma}, \widehat{\mu}, \widehat{U}, \widehat{V}) \in \arg\min_{\theta\in\Theta_r} \mathcal{L}_n(\theta). \tag{10}$$

The constraint set Θ_r encodes the normalizations used in Section 4.2 to make the consensus direction, judge-specific consensus sensitivities, and low-rank disagreement component statistically identifiable, so that $\widehat{\theta}$ estimates the canonical representative of the decomposition (4). The resulting optimization problem is nonconvex because of the bilinear factorization in (U, V) and the nonlinear

Algorithm 1 Heterogeneous judge-aware (HJA) ranking by proximal anchored alternating MLE

Require: Data $\{(n_{kij}, Y_{kij})\}_{(k,i,j)\in\Omega}$, rank r, tolerance ε, iterations T, and parameters $\{\tau^{(t)}\}_{t\geq 0}$.
Ensure: Estimated parameter $\widehat{\theta}$, Fisher information matrix $\mathcal{I}_n(\widehat{\theta})$, and a smooth target estimand q.
1: $\theta^{(0)} \leftarrow$ INITIALIZE($\{(n_{kij}, Y_{kij})\}_{(k,i,j)\in\Omega}, r$) based on Algorithm C.1.
2: **for** $t = 0, 1, \ldots, T-1$ **do**
3: Set $\tau \leftarrow \tau^{(t)}$.
4: Compute $(\widetilde{\gamma}, \widetilde{U})$ as the unique minimizer of $\mathcal{L}_n(\gamma, \mu^{(t)}, U, V^{(t)}) + \frac{\tau}{2}\|\gamma - \gamma^{(t)}\|_2^2 + \frac{\tau}{2}\|U - U^{(t)}\|_F^2$ subject to $1_K^\top \gamma = K$ and $1_K^\top U = 0_r^\top$.
5: Compute $(\widetilde{\mu}, \widetilde{V})$ as the unique minimizer of $\mathcal{L}_n(\widetilde{\gamma}, \mu, \widetilde{U}, V) + \frac{\tau}{2}\|\mu - \mu^{(t)}\|_2^2 + \frac{\tau}{2}\|V - V^{(t)}\|_F^2$ subject to $1_N^\top \mu = 0$ and $1_N^\top V = 0_r^\top$.
6: $\theta^{(t+1)} \leftarrow$ REANCHOR($\widetilde{\gamma}, \widetilde{\mu}, \widetilde{U}, \widetilde{V}$) based on Algorithm C.2.
7: **If** $|\mathcal{L}_n(\theta^{(t+1)}) - \mathcal{L}_n(\theta^{(t)})|/[1 + |\mathcal{L}_n(\theta^{(t)})|] < \varepsilon$ **then break**.
8: **end for**
9: $\widehat{\theta} \leftarrow \theta^{(t+1)}$ and $\mathcal{I}_n(\widehat{\theta}) \leftarrow$ ESTIMATEFISHERINFO($\widehat{\theta}$) based on Algorithm C.3.
10: Compute $\widehat{q} = q(\widehat{\theta})$, $\widehat{\mathrm{se}}(\widehat{q})$, and the Wald interval $\widehat{q} \pm z_{1-\alpha/2}\widehat{\mathrm{se}}(\widehat{q})$ using Algorithm C.4.

anchoring constraints in Θ_r. In addition, direct alternating updates over the judge and item blocks do not preserve the nonlinear identification constraints, including $\mu^\top V = 0$, $N^{-1}V^\top V = I_r$, and $K^{-1}U^\top U$. To address this, we develop a proximal alternating algorithm for constrained MLE that maintains the identifying structure of the decomposition while exploiting the blockwise smoothness of the loss function, shown in Algorithm 1. Each of its likelihood-and-gradient evaluation costs up to $\mathcal{O}(KNr + N^2)$ in time complexity. We combine proximal block updates with a re-anchoring step to obtain the canonical representative in Θ_r. Section C.2 offers further implementation details.

To analyze Algorithm 1 in the local nonconvex regime, we work in an identified chart around the target constrained MLE. Assumption A1 ensures that the re-anchoring map is locally well defined and nondegenerate; Assumption A2 imposes an isolated constrained local minimizer with positive curvature in this chart; and Assumption A3 places the initialization in its basin of attraction. Let $\widehat{\theta}$ be defined by (10), and let $\{\theta^{(t)}\}_{t\geq 0}$ denote the anchored sequence generated by Algorithm 1. Throughout the remainder of Section 5, we fix the local column signs of (U, V) by the anchoring convention (e.g., by Remark 1) and work in the corresponding local identified parameterization.

Theorem 3 (Locally linear convergence)**.** *Under Assumptions A1–A3, the iterates of Algorithm 1 remain in the identified local chart, the likelihood $\{\mathcal{L}_n(\theta^{(t)})\}_{t\geq 0}$ are nonincreasing, and $\theta^{(t)} \to \widehat{\theta}$. Moreover, there exists $\rho \in (0,1)$ such that for t large enough, $\|\theta^{(t)} - \widehat{\theta}\|_2 + \mathcal{L}_n(\theta^{(t)}) - \mathcal{L}_n(\widehat{\theta}) \lesssim \rho^t$.*

Theorem 3 suggests that the re-anchored alternating updates of Algorithm 1 are stable within the identified local chart, i.e., the objective decreases monotonically, the iterates converge to the intended constrained MLE, and the final local convergence is geometric. See Section C.3 for its proof.

5.2 Uncertainty quantification

Point estimates alone are insufficient for judge-aware ranking: in near-tie regimes, small perturbations in the observed comparisons can change the consensus ordering, judge-specific rankings, or estimated disagreement structure. We therefore quantify uncertainty for the score contrasts that determine ranking conclusions, rather than only for the latent components themselves. We study this problem in a fixed-rank repeated-comparison regime, where K and N may remain fixed while information accumulates through repeated comparisons on observed triples. This regime matches LLM evaluation settings in which the same judge panel can be queried repeatedly across prompts, random seeds, or judging instances, even when adding trusted judges is costly.

We impose the following regularity conditions for statistical inference in this setting. Assumption A4 assumes stable asymptotic sampling fractions $n_{kij}/n \to w_{kij}$, and Assumption A5 requires the active comparison graph for each judge to be connected. They identify the row-centered score matrix from active pairwise differences and pass this information to model components.

With the local sign convention fixed as in Section 5.1 and Ω_w defined in (D.1), we define the Fisher information for the anchored true parameter $\theta^* \in \Theta_r$ by

$$\mathcal{I}(\theta^*) = \textstyle\sum_{(k,i,j)\in\Omega_w} w_{kij} p^*_{kij}(1 - p^*_{kij}) \nabla_\theta \eta_{kij}(\theta^*) \nabla_\theta \eta_{kij}(\theta^*)^\top. \tag{11}$$

The next theorem shows the inferential properties of the estimate and its smooth target functionals.

Theorem 4 (Fixed-rank asymptotics). *For the observed-data model (2) and (4) with $\mu^* \neq 0$, under Assumptions A4 and A5, the constrained MLE (10) satisfies $\widehat{\theta} \xrightarrow{p} \theta^*$, $\mathcal{I}(\theta^*)$ is positive definite, and*

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

Consequently, for any smooth scalar target $q : \Theta_r \to \mathbb{R}$,

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

Theorem 4 shows that repeated judgments can support inference on the structured components of the HJA model even when the judge panel itself does not grow. The Fisher information in (11) can be consistently estimated by the plug-in observed design estimator

$$\mathcal{I}_n(\widehat{\theta}) = \sum_{(k,i,j)\in\Omega} \frac{n_{kij}}{n} \widehat{p}_{kij}(1 - \widehat{p}_{kij}) \nabla_\theta \eta_{kij}(\widehat{\theta}) \nabla_\theta \eta_{kij}(\widehat{\theta})^\top , \tag{12}$$

where $\widehat{p}_{kij} = \{1 + \exp[-\eta_{kij}(\widehat{\theta})]\}^{-1}$. Under the assumptions of Theorem 4, $\mathcal{I}_n(\widehat{\theta}) \xrightarrow{p} \mathcal{I}(\theta^*)$. Thus Wald standard errors for smooth targets $q(\theta)$ are given by $\widehat{\mathrm{se}}\{q(\widehat{\theta})\} = [\frac{1}{n} \nabla q(\widehat{\theta})^\top \mathcal{I}_n(\widehat{\theta})^{-1} \nabla q(\widehat{\theta})]^{1/2}$. Typical choices of q include consensus contrasts $\mu_i - \mu_j$, judge-specific contrasts $S_{ki} - S_{kj}$, consensus-sensitivity parameters γ_k, pairwise probabilities p_{kij}, and smooth summaries of $UV^\top$. These are the natural inferential objects for ranking, as full rankings are discontinuous at ties, whereas pairwise score differences remain smooth and directly quantify ranking uncertainty. Note that when $r^* = 0$, the inferential results above still apply to μ and γ.

6 Empirical Evaluation

6.1 Benchmarking across synthetic and real data

(A) Synthetic simulations. We first use synthetic data to isolate the effect of structured judge heterogeneity under a known latent score matrix and known consensus-plus-heterogeneity decomposition. Data are generated from (2) and (4) with $N = 8$ items, $K = 4$ judges, rank $r = 1$, and 50 random seeds. We compare 4 methods: HJA, JA-Ranking [17], pooled BTL ($\gamma = 1_K, r = 0$), and an unstructured judge-wise BTL estimator followed by a rank-$(r+1)$ truncated SVD of the estimated score matrix; and consider 2 regimes: a sample-size experiment varying the total number of comparisons n_{cmp} at fixed heterogeneity scale $h = 1$, and a heterogeneity experiment varying h at fixed $n_{\mathrm{cmp}} = 800$, where h scales the residual component $UV^\top$. Full data-generating details, metric definitions, convergence diagnostics, and sensitivity-only simulations are deferred to Section E.

Figure 2 shows that HJA improves score recovery, consensus ranking, uncertainty calibration, and judge-specific sign prediction as comparison information increases. The advantage is most pronounced as heterogeneity grows: pooled BTL and JA-Ranking are structurally misspecified because they cannot represent residual judge-item preference structure, and hence concentrate around biased score matrices in larger samples. The unstructured BTL+SVD baseline is less accurate and has higher score-level variance because it estimates the full score matrix before imposing low-rank structure. The ablation in Section E.5 further shows that the structured optimization and reanchor step are necessary to stabilize the identified decomposition and improve score-level recovery.

(B) Misspecified synthetic data. We use a hidden-expertise setting with $N = 12$ items and $K = 9$ judges to illustrate how our model can identify the potential latent domains of judges' expertise, when the synthetic data do not agree with our model. With the two preference (expertise) blocs (6 maths-reliable, 3 coding-reliable, each judge nearly random outside its own domain) expressed by a mixture model based on (7) with its $U = 0$, the likelihood cannot be expressed via (9), thus lying outside our model's data generating mechanism. Every judge contributes 4,800 comparisons, and no domain label d is supplied to any model. The details in Section E.6 will show that a pooled-ranking can reverse the pattern, while HJA's $UV^\top$ term approximates the pattern accurately with correct interpretation; indeed, across 50 repetitions, HJA correctly recovers the ranking in 48 runs, versus 35/0 for the sensitivity-only/pooled-BTL, with high held-out accuracy. We show that HJA both mitigates the reversal and identifies the (misspecified) latent structure responsible for it.

(C) Real-data robustness evaluation. We next use four representative real-world evaluation datasets to test whether modeling structured judge heterogeneity improves ranking under noisy,

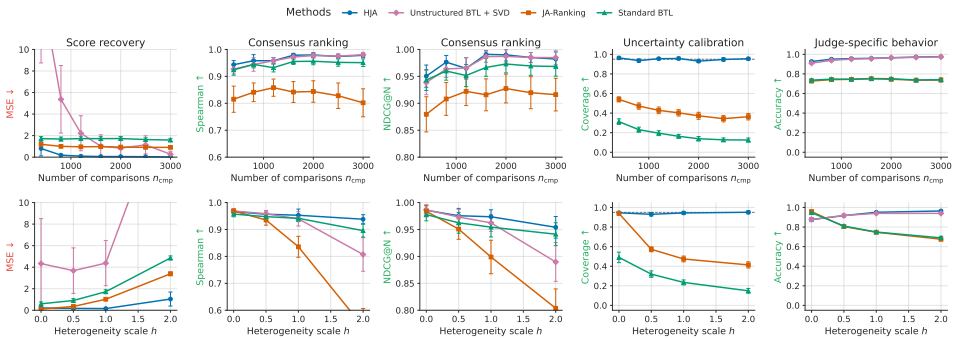


Figure 2: Benchmarking on synthetic simulations, with (a) varying numbers of pairwise comparisons $T \in \{400, 800, 1200, 1600, 2000, 2500, 3000\}$ at heterogeneity scale 1.0, and (b) varying heterogeneity scales over $\{0, 0.5, 1, 2\}$ and fixed $T = 800$ comparisons. Each panel reports the mean over 50 independent repetitions with 95% normal-approximation error bars. Columns show performance on score recovery (MSE), consensus rank quality (Spearman and NDCG@N), uncertainty calibration (nominal 95% score-entry coverage), and judge-specific behavior (sign accuracy).

uneven, and near-tie judgments: Chatbot Arena [15], MT-Bench [1], UltraFeedback [30], and an in-house judgment dataset [17]. We compare the induced consensus rankings by HJA, JA-Ranking, and BTL along three axes: robustness to injected noisy judges, held-out pairwise prediction, and accuracy on empirically ambiguous near-tie pairs, where automatic rankings are especially fragile [11, 12]. Noisy-judge robustness measures how stable the estimated ranking remains as additional noisy judges are injected; held-out prediction evaluates record-level generalization; and near-tie accuracy focuses on item pairs with empirical win rates close to 0.5, where small modeling biases can flip the induced ordering. See Section F.1 for dataset and experimental details.

Table 1 shows that HJA performs strongly on the three public benchmarks, Chatbot Arena, MT-Bench, and UltraFeedback, with clear gains in held-out prediction and consistent improvements in robustness and near-tie accuracy. On the in-house dataset [17], though HJA does not dominate JA-Ranking under the strongest perturbations or all near-tie slices, it remains competitive. Overall, these results support the benefit of modeling structural judge heterogeneity beyond scalar sensitivity.

(D) Stress test. We inject a source with a systematic preference for a fixed candidate cluster and with approximately the comparison volume of the entire clean panel to each of the four datasets, mimicking, say, strong self-preference bias [31]. Table 2 shows that HJA better preserves clean rankings on the three public datasets. In in-house data, however, all methods reverse the ranking due to the volume of the biased injection. See implementation details in Section F.1.

6.2 Consensus, sensitivity, and disagreement in real data

Beyond benchmark comparison, we use HJA as a full-data diagnostic tool for real multi-judge evaluation. We focus on Chatbot Arena, a large crowdsourced LLM comparison dataset, and report

Table 1: Real-data comparison. Robustness is evaluated under increasing noisy-judge perturbation; near-tie accuracy is evaluated across increasingly less ambiguous near-tie groups. Entries are mean $\pm$ standard deviation in 20 random seeds; bold marks the best mean per dataset-metric group.

Dataset	Method	Noisy-judge robustness			Hold-out accuracy	Near-tie accuracy		
		Low	Mid	High		Closest	Mid	Farthest
Chatbot Arena	HJA	**0.95±0.07**	**0.85±0.09**	**0.82±0.07**	**0.65±0.01**	**0.68±0.05**	**0.69±0.04**	**0.67±0.06**
	JA-Ranking	0.77±0.09	0.51±0.12	0.46±0.11	0.58±0.01	0.58±0.03	0.54±0.05	0.50±0.06
	BTL	0.47±0.16	0.33±0.13	0.21±0.11	0.58±0.01	0.57±0.04	0.54±0.05	0.51±0.06
MT-Bench	HJA	**1.00±0.00**	**1.00±0.00**	**1.00±0.00**	**0.76±0.01**	**0.74±0.02**	**0.84±0.01**	**0.89±0.01**
	JA-Ranking	0.67±0.00	0.67±0.00	0.67±0.00	0.70±0.01	0.62±0.02	0.78±0.01	0.86±0.01
	BTL	0.92±0.15	0.87±0.17	0.85±0.20	0.70±0.01	0.62±0.02	0.78±0.01	0.86±0.01
Ultra-Feedback	HJA	0.81±0.06	**0.65±0.11**	**0.67±0.13**	**0.70±0.01**	**0.69±0.05**	**0.68±0.04**	**0.71±0.06**
	JA-Ranking	**0.83±0.11**	0.63±0.11	0.62±0.16	0.62±0.01	0.55±0.06	0.58±0.06	0.60±0.04
	BTL	0.57±0.09	0.32±0.12	0.30±0.10	0.61±0.01	0.55±0.06	0.58±0.06	0.60±0.04
In-house	HJA	**0.94±0.05**	**0.83±0.05**	0.33±0.05	**0.62±0.01**	**0.60±0.07**	**0.56±0.10**	0.58±0.10
	JA-Ranking	0.65±0.06	0.43±0.09	**0.36±0.08**	0.61±0.01	0.58±0.07	**0.56±0.08**	**0.59±0.09**
	BTL	0.35±0.08	0.18±0.06	0.12±0.04	0.61±0.01	0.57±0.07	**0.56±0.08**	**0.59±0.09**

Table 2: Stress-test ranking preservation: Spearman correlation with each method's clean ranking (mean ± sd over 20 replications)

Dataset	HJA	Sensitivity-only	Pooled BTL
Chatbot Arena	$\mathbf{0.1748 \pm 0.0164}$	0.0553 ± 0.0102	0.0339 ± 0.0073
MT-Bench	$\mathbf{0.2343 \pm 0.0469}$	-0.1229 ± 0.0280	-0.1086 ± 0.0287
UltraFeedback	$\mathbf{0.6245 \pm 0.3283}$	-0.2705 ± 0.0130	-0.2636 ± 0.0145
In-house	$\mathbf{-0.3263 \pm 0.0046}$	-0.3324 ± 0.0024	-0.3293 ± 0.0020

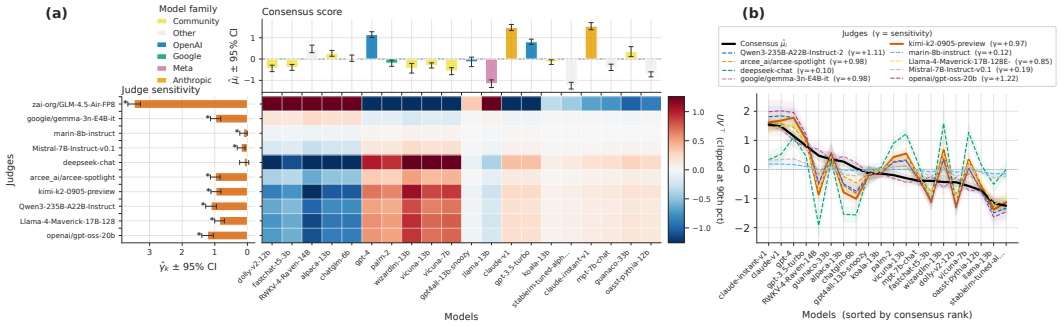


Figure 3: HJA diagnostic analysis on Chatbot Arena. (a) Heatmap shows the heterogeneous preference matrix $\widehat{U}\widehat{V}^\top$; barplots show judge sensitivity $\widehat{\gamma}$ and consensus score $\widehat{\mu}$, with 95% confidence intervals. The models are colored by LLM model families. (b) Judge-specific score profiles $\widehat{S}_{k,:}$, ordered by consensus score $\widehat{\mu}$ with the high-leverage outlier excluded.

analogous analyses for the other real datasets in Section F.2. The fitted decomposition separates three quantities: the consensus ranking, judge sensitivity, and residual judge-model interactions. As shown in Figure 3(a), this separation is useful for judge assessment: a large $\widehat{\gamma}_k$ indicates a strong response to the consensus direction, but reliability also requires small residual leverage. The judge `zai-org/GLM-4.5-Air-FP8` has both unusually large consensus sensitivity and unusually large residual disagreement, making it a high-leverage judge rather than simply a strong judge; this agrees with the broader diagnostic in Figure F6. HJA also provides judge-specific rankings through the fitted score profiles $\widehat{S}_{k,:} = \widehat{\gamma}_k \widehat{\mu}^\top + \widehat{U}_{k,:}\widehat{V}^\top$. Figure 3(b) shows these profiles ordered by the consensus ranking. Unlike fitting each judge independently, HJA estimates these rankings jointly through a shared consensus direction and a low-rank disagreement structure. This pooling is important in sparse or noisy pairwise data: judge-specific deviations are allowed, but they are regularized through common latent disagreement directions rather than estimated as unrelated score vectors.

Finally, the residual component $\widehat{U}\widehat{V}^\top$ localizes structured judge-model affinities. After identifying high-leverage outliers, we can refit or visualize the model with those judges filtered to obtain a more stable consensus and a less saturated interaction heatmap; see Figure F8. The same residual structure can be inspected by clustering judges and models, as in Figures F9 and F10, to reveal groups with similar evaluation behavior. These analyses suggest that HJA can flag unreliable or high-leverage judges, recover judge-specific ranking preferences, and expose systematic affinity patterns, while treating self- or family-preference effects as diagnostic hypotheses rather than universal conclusions.

7 Conclusion and discussion

We introduced HJA ranking, an identifiable multi-judge BTL framework that separates consensus ranking, judge sensitivity, and structured residual disagreement. With constrained estimation and fixed-panel uncertainty quantification, HJA yields rankings that are more robust, interpretable, and diagnostic under heterogeneous judgments. However, the framework is not intended to certify rankings under adversarial judge panels: severe coordinated bias may affect the consensus estimate, and the nonconvex optimization remains locally initialized. Hence, deliberate manipulations may cause HJA to output biased rankings, which poses a potential negative impact. Thus, HJA should be viewed as a statistically grounded diagnostic for heterogeneous evaluation before benchmark design.

Acknowledgements and disclosures

We appreciate the Area Chair and reviewers' thoughtful suggestions, which greatly improved the content and clarity of this paper. The last author acknowledges support from the Institute of Data Science (IDS) at the University of Hong Kong through the HKU100 award.

References

[1] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems*, volume 36, pages 46595–46623, 2023.

[2] Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

[3] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744, 2022.

[4] 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, pages 53728–53741, 2023.

[5] Xi Chen, Paul N Bennett, Kevyn Collins-Thompson, and Eric Horvitz. Pairwise ranking aggregation in a crowdsourced setting. In *Proceedings of the sixth ACM international conference on Web search and data mining*, pages 193–202, 2013.

[6] Naoki Otani, Toshiaki Nakazawa, Daisuke Kawahara, and Sadao Kurohashi. IRT-based aggregation model of crowdsourced pairwise comparison for evaluating machine translations. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 511–520, 2016.

[7] Justin Zhao, Flor Miriam Plaza-del Arco, Benjamin Genchel, and Amanda Cercas Curry. Language model council: Democratically benchmarking foundation models on highly subjective tasks. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 12395–12450, 2025.

[8] Amit Dhurandhar, Rahul Nair, Moninder Singh, Elizabeth Daly, and Karthikeyan Natesan Ramamurthy. Ranking large language models without ground truth. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2431–2452, 2024.

[9] Anna Bavaresco, Raffaella Bernardi, Leonardo Bertolazzi, Desmond Elliott, Raquel Fernández, Albert Gatt, Esam Ghaleb, Mario Giulianelli, Michael Hanna, Alexander Koller, André F. T. Martins, Philipp Mondorf, Vera Neplenbroek, Sandro Pezzelle, Barbara Plank, David Schlangen, Alessandro Suglia, Aditya K Surikuchi, Ece Takmaz, and Alberto Testoni. LLMs instead of human judges? a large scale empirical study across 20 NLP evaluation tasks. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 238–255, 2025.

[10] Xiyan Fu and Wei Liu. How reliable is multilingual LLM-as-a-judge? In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 11040–11053. Association for Computational Linguistics, 2025.

[11] Luke Guerdan, Solon Barocas, Kenneth Holstein, Hanna M. Wallach, Steven Z. Wu, and Alexandra Chouldechova. Validating LLM-as-a-judge systems under rating indeterminacy. In *Advances in Neural Information Processing Systems*, volume 38, pages 112282–112350, 2025.

[12] Mingqi Gao, Yixin Liu, Xinyu Hu, Xiaojun Wan, Jonathan Bragg, and Arman Cohan. Re-evaluating automatic LLM system ranking for alignment with human preference. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 4605–4629, 2025.

[13] 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.

[14] David R Hunter. MM algorithms for generalized Bradley–Terry models. *The Annals of Statistics*, 32(1):384–406, 2004.

[15] Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Banghua Zhu, Hao Zhang, Michael Jordan, Joseph E. Gonzalez, and Ion Stoica. Chatbot Arena: An open platform for evaluating LLMs by human preference. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 8359–8388. PMLR, 2024.

[16] Cheng-Han Chiang and Hung-yi Lee. Can large language models be an alternative to human evaluations? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15607–15631, 2023.

[17] Mingyuan Xu, Xinzi Tan, Jiawei Wu, and Doudou Zhou. A judge-aware ranking framework for evaluating large language models without ground truth. In *Forty-third International Conference on Machine Learning*, 2026.

[18] Yu Lu and Sahand N Negahban. Individualized rank aggregation using nuclear norm regularization. In *2015 53rd Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 1473–1479. IEEE, 2015.

[19] Dohyung Park, Joe Neeman, Jin Zhang, Sujay Sanghavi, and Inderjit Dhillon. Preference completion: Large-scale collaborative ranking from pairwise comparisons. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 1907–1916. PMLR, 2015.

[20] Arun Rajkumar and Shivani Agarwal. When can we rank well from comparisons of $O(n \log(n))$ non-actively chosen pairs? In *29th Annual Conference on Learning Theory*, volume 49 of *Proceedings of Machine Learning Research*, pages 1376–1401. PMLR, 2016.

[21] Tao Jin, Pan Xu, Quanquan Gu, and Farzad Farnoud. Rank aggregation via heterogeneous Thurstone preference models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(04):4353–4360, 2020.

[22] Rui Wu, Jiaming Xu, Rayadurgam Srikant, Laurent Massoulié, Marc Lelarge, and Bruce Hajek. Clustering and inference from pairwise comparisons. In *Proceedings of the 2015 ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Systems*, pages 449–450, 2015.

[23] Yassir Fathullah and Mark Gales. Generalised probabilistic modelling and improved uncertainty estimation in comparative LLM-as-a-judge. In *Proceedings of the Forty-first Conference on Uncertainty in Artificial Intelligence*, volume 286 of *Proceedings of Machine Learning Research*, pages 1266–1288. PMLR, 2025.

[24] Victor Wang, Michael JQ Zhang, and Eunsol Choi. Improving LLM-as-a-judge inference with the judgment distribution. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 23173–23199, Suzhou, China, nov 2025. Association for Computational Linguistics.

[25] Annalisa Szymanski, Noah Ziems, Heather A Eicher-Miller, Toby Jia-Jun Li, Meng Jiang, and Ronald A Metoyer. Limitations of the LLM-as-a-judge approach for evaluating LLM outputs in expert knowledge tasks. In *Proceedings of the 30th international conference on intelligent user interfaces*, pages 952–966, 2025.

[26] Jiachun Li, David Simchi-Levi, and Will Wei Sun. LLM evaluation as tensor completion: Low rank structure and semiparametric efficiency. *arXiv preprint arXiv:2604.05460*, 2026.

[27] Jianqing Fan, Hyukjun Kwon, and Xiaonan Zhu. Uncertainty quantification for ranking with heterogeneous preferences. *arXiv preprint arXiv:2509.01847*, 2025.

[28] R. Duncan Luce. *Individual Choice Behavior: A Theoretical Analysis*. Wiley, New York, 1959.

[29] Jing Li, Rafal Mantiuk, Junle Wang, Suiyi Ling, and Patrick Le Callet. Hybrid-MST: A hybrid active sampling strategy for pairwise preference aggregation. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

[30] Ganqu Cui, Lifan Yuan, Ning Ding, Guanming Yao, Bingxiang He, Wei Zhu, Yuan Ni, Guotong Xie, Ruobing Xie, Yankai Lin, Zhiyuan Liu, and Maosong Sun. ULTRAFEEDBACK: Boosting language models with scaled AI feedback. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 9722–9744. PMLR, 21–27 Jul 2024.

[31] Zhi-Yuan Chen, Hao Wang, Xinyu Zhang, Enrui Hu, and Yankai Lin. Beyond the surface: Measuring self-preference in LLM judgments. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 1653–1672. Association for Computational Linguistics, 2025.

Appendix

This serves as an appendix to the main paper. Below, we provide an outline for the appendix along with a summary of the notation used in the main paper and the appendix.

Organization. The content of the appendix is organized as follows.

Appendix		Content
Section A		Relation to alternative formulations.
Section B	B.1	Proof of Proposition 2 on model identification.
Section C	C.1	Rank selection and reduction to simpler models.
	C.2	Algorithms and implementation details.
	C.3	Proof of Theorem 3 on convergence analysis.
Section D	D.1	Proof of consistency part of Theorem 4.
	D.2	Proof of asymptotic normality part in Theorem 4 via graph theory.
	D.3	Auxiliary lemmas used for the proofs.
	D.4	Partial-sampling (pooled-graph treatment) regime.
Section E	E.1	Main heterogeneous DGP specification.
	E.2	Comparison design, fitting protocol, and metrics.
	E.3	Convergence diagnostics.
	E.4	JA-Ranking DGP specification and results.
	E.5	Ablation study.
	E.6	Latent-expertise (misspecified) data experiment.
Section F	F.1	Real-data evaluation details.
	F.2	Interpretation on real data analyses.

Notation. An overview of some general notations used in the main paper and the appendix is as follows.

In $\mathbb{R}^d$, the j-th standard basis vector is denoted by e_j and the zero (or all-one) vector is denoted by 0_d (or 1_d) or simply 0 (or 1) if the dimension is clear from the context. The cardinality of a set $\mathcal{S}$ is denoted by $|\mathcal{S}|$. The indicator function is denoted by $\mathbb{1}_{\{\cdot\}}$.

We use "o" and "$\mathcal{O}$" to denote the little-o and big-O notations; "$o_{\mathbb{P}}$" and "$\mathcal{O}_{\mathbb{P}}$" are their probabilistic counterparts. For sequences $\{a_n\}$ and $\{b_n\}$, we write $a_n \lesssim b_n$ if $a_n = \mathcal{O}(b_n)$; and $a_n \asymp b_n$ if $a_n = \mathcal{O}(b_n)$ and $b_n = \mathcal{O}(a_n)$. Convergence almost surely, in probability, and in distribution is denoted by "$\xrightarrow{\text{a.s.}}$", "$\xrightarrow{\text{p}}$", and "$\xrightarrow{\text{d}}$", respectively. When "positivity" property is imposed or stated on a matrix, it means positive definiteness. We use $\mathrm{KL}(p \,\|\, q)$ to denote the Kullback–Leibler (KL) divergence between two distributions p and q.

Reproducibility. All experiment results in this paper can be reproduced with datasets openly accessible. The code files are available at https://github.com/JinHongDu-Lab/HJA-Ranking.

A Relation to Alternative Formulations

Relation to JA-Ranking. JA-Ranking [1] and related judge-aware BTL models represent judge heterogeneity primarily through a judge-specific discrimination or sensitivity parameter, leading to a sensitivity-only score matrix of the form $S = \gamma\mu^\top$. HJA contains this model as the special case $r = 0$, up to the normalization $1_K^\top\gamma = K$. The main distinction is that, for $r > 0$, HJA allows judges to deviate from the consensus direction through the residual component $UV^\top$. Thus γ_k should be interpreted as judge k's sensitivity to the consensus ranking direction, rather than as a global sensitivity parameter scaling all aspects of the judge-specific score vector. A judge can be weakly aligned with the consensus while still having strong, structured preferences in the residual disagreement subspace. This differs from formulations in which a single positive discrimination parameter scales the entire latent score vector. If one instead wishes γ to scale the full judge-specific score, one may set $U = \mathrm{diag}(\gamma)\widetilde{U}$ and write

$$S = \gamma\mu^\top + \mathrm{diag}(\gamma)\widetilde{U}V^\top = \mathrm{diag}(\gamma)\big(1_K\mu^\top + \widetilde{U}V^\top\big).$$

Under this alternative parameterization, γ plays the role of a global sensitivity parameter, since it scales both the consensus and heterogeneous components simultaneously.

Additionally, our formulation does not impose strict positivity on the coefficients γ_k. This is useful in settings with noisy or partially adversarial judges: $\gamma_k = 0$ means that judge k does not respond to the consensus direction and contributes only through structured disagreement around consensus, whereas $\gamma_k < 0$ indicates systematic opposition to the consensus signal. By contrast, JA-ranking imposes a strict positivity restriction on the corresponding discrimination parameters.

Thirdly, JA-Ranking assumes a balanced comparison design in which each unordered model pair is sampled uniformly for each judge. HJA instead works directly with the observed triple set Ω in (1), allowing arbitrary and unbalanced cell counts n_{kij}, including designs in which some judge–pair cells are observed much more often than others or are missing entirely. This flexibility is important in realistic LLM evaluation pipelines, where certain comparisons may be difficult to obtain because of API rate limits, model availability, annotation cost, safety filters, or adaptive sampling policies that focus effort on near-tie or high-disagreement pairs. In the main theory, identification and inference are guaranteed when each judge-specific active comparison graph is connected. Beyond this regime, Section D.4 gives a pooled-graph treatment: disconnected judge-specific graphs can still be handled when the pooled comparison graph is connected and the structural low-rank constraints rule out unidentified local perturbations.

Relation to low-rank BTL and collaborative ranking models. Low-rank preference and collaborative ranking models [2–4] also exploit low-dimensional structure in heterogeneous preferences. However, HJA uses low rank for a different inferential purpose. Rather than treating the factorization of the full score matrix as the primary object, HJA first defines the consensus direction as the row-average score vector and then constrains the residual factor $UV^\top$ to be orthogonal to that consensus direction. Consequently, the low-rank component represents structured disagreement around consensus, not an arbitrary rotated latent preference space. The normalization and rank conditions in Proposition 2 make this separation identifiable from the BTL comparison law, up to the usual column-wise sign ambiguity of the factors.

Relation to two-step BTL plus spectral decomposition. A natural alternative is to first estimate an unrestricted judge-specific BTL score matrix $\widehat{S}$, and then recover $(\widehat{\gamma}, \widehat{\mu}, \widehat{U}, \widehat{V})$ by applying the decomposition in Proposition 2 followed by a truncated singular value decomposition (SVD). This two-step procedure is useful as a diagnostic and as a computational baseline, especially when each judge has many repeated comparisons over a dense comparison graph. However, it separates score estimation from the structural constraints and therefore does not exploit the low-rank consensus-plus-heterogeneity model during likelihood fitting. In finite samples, this can lead to high variance in the estimated score matrix before the spectral step is applied, as shown in Figure 2, when the heterogeneity level increases. By contrast, the constrained MLE estimates the identifiable components jointly, borrows strength across judges and items through the structural model, and directly supports the Fisher-information-based uncertainty quantification.

Relation to random graph theory. The inference properties of HJA rely on the connectivity structure of the active comparison graphs $\mathcal{G}_k^w$ (where edge (i, j) exists if $(k, i, j) \in \Omega$) for each judge k. The weighted graph-Laplacian framework in Proposition D.2 shows that Fisher information non-

degeneracy is controlled by properties of the graph, similar to spectral properties in random graph theory [5]. With these techniques, we can establish the properties of MLE without requiring a fully connected comparison graph, which is important for real-world evaluation settings, as judges may only compare subsets of items.

In the asymptotic regime where the comparison graph $\mathcal{G}$ (where edge (i, j) exists if $(k, i, j) \in \Omega$ for some k) is fixed, connectivity is imposed as a deterministic structural assumption (Assumption A5). However, when comparisons are allocated randomly, or the graph grows dynamically (e.g., active learning over time), one could appeal to results from random graph theory, such as properties of Erdős-Rényi [6], to establish that connectivity holds with high probability. Such connections suggest that HJA's asymptotic theory could be extended to settings with random or growing comparison graphs, though the current paper focuses on fixed-panel designs common in LLM evaluation.

B Identification

B.1 Proof of Proposition 2

Proof of Proposition 2. The proof consists of three parts. First for (i), given any set of pairwise probabilities, S is uniquely identified if each of its row is restricted to have zero sum (i.e., Eq. (3)). The second part proves that the mapping induced by (4) from Θ_r to the space of S is injective. Finally, Part (iii) imposes (3) on the space of S, which makes such mapping bijective (one-to-one). Part (iii), combined with Part (ii), establishes the unique existence, which is identification.

Part (i): Identification of BTL score matrix. For each fixed judge k, the pairwise probabilities determine all score differences $S_{ki} - S_{kj}$ through the inverse logit (2). Hence, if $\widetilde{S}$ is another matrix generating the same pairwise probabilities, then

$$\widetilde{S}_{ki} - \widetilde{S}_{kj} = S_{ki} - S_{kj} \qquad \text{for all } i, j.$$

Therefore, for each k, the vector $\widetilde{S}_{k,:} - S_{k,:}$ is constant across items, i.e., there exists $c_k \in \mathbb{R}$ such that

$$\widetilde{S}_{k,:} = S_{k,:} + c_k 1_N^\top.$$

Right-multiplying both sides by 1_N yields

$$\widetilde{S}_{k,:}1_N = S_{k,:}1_N + c_k 1_N^\top 1_N = S_{k,:}1_N + N c_k.$$

But it is required that $S1_N = 0$, with the same applying to $\widetilde{S}$, and therefore $c_k = 0$ for all $k \in \{1, \dots, K\}$, i.e., $S = \widetilde{S}$.

Part (ii): Identification of latent components. Below, we prove that if $S = \gamma\mu^\top + UV^\top$ with $(\gamma, \mu, U, V) \in \Theta_r$, the parameters are uniquely identified subject to some sign changes to be specified later.

Multiplying both sides of (4) with $\frac{1}{K}1_K^\top$ yields

$$\frac{1}{K}1_K^\top S = \frac{1}{K}1_K^\top \gamma\mu^\top + \frac{1}{K}1_K^\top UV^\top = \mu^\top, \tag{B.1}$$

so μ is immediately identified by the row average of S.

To further separate the consensus item direction from the heterogeneity item subspace, by Condition 1(iii),

$$S\mu = \gamma\|\mu\|^2 + U(V^\top\mu) = \gamma\|\mu\|^2 \Longrightarrow \gamma = \frac{S\mu}{\|\mu\|^2}, \tag{B.2}$$

and thus γ is identified.

We proceed to identify U and V from $\widetilde{S} := S - \gamma\mu^\top = UV^\top$. Write $U = [u_1, \dots, u_r]$ and $V = [v_1, \dots, v_r]$. By Condition 1(iv), the vectors $v_1, \dots, v_r$ are orthonormal after normalization by $\sqrt{N}$, and the columns of U are mutually orthogonal with $\|u_j\|_2^2 = Kd_j$. Together with the orthogonality condition (Condition 1(iii)), this yields the orthogonal decomposition

$$\widetilde{S} = \sum_{j=1}^{r} u_j v_j^\top,$$

where the left vectors $\gamma, u_1, \dots, u_r$ are mutually orthogonal and the right vectors $\mu, v_1, \dots, v_r$ are mutually orthogonal. Therefore, after normalization, this is a SVD of $\widetilde{S}$. Its nonzero singular values are

$$\sigma_j = \|u_j\|_2\|v_j\|_2 = \sqrt{KN\, d_j}, \qquad j = 1, \dots, r.$$

The strict inequalities $d_1 > \cdots > d_r > 0$ imply that $\sigma_1, \dots, \sigma_r$ are pairwise distinct and therefore simple.

Now let $\widetilde{S} = \widetilde{U}\widetilde{V}^\top$ be another representation satisfying the same conditions. The same argument shows that this representation also produces an SVD of $\widetilde{S}$ with the same simple nonzero singular

values. By the uniqueness of singular vectors associated with simple singular values, there exist signs $\varepsilon_1, \ldots, \varepsilon_r \in \{\pm 1\}$ such that

$$\widetilde{u}_j = \varepsilon_j u_j, \qquad \widetilde{v}_j = \varepsilon_j v_j, \quad j = 1, \ldots, r.$$

For each j, the pair $(\widetilde{u}_j, \widetilde{v}_j)$ may be replaced by $(-\widetilde{u}_j, -\widetilde{v}_j)$ without changing the product $\widetilde{u}_j \widetilde{v}_j^\top$. And therefore U and V are identified up to the sign changes in their corresponding columns.

Remark 1 (Sign of factors and loadings)**.** In subsequent analysis where such sign ambiguities also need ruling out, we may require the first non-zero elements of each column of U to be positive, so that the signs of U and V are anchored and fixed.

Part (iii): Existence of parameters recovered from S. We are left to show that given any matrix $S \in \mathbb{R}^{K \times N}$ satisfying $S1_N = 0$ and Condition 2, these parameters can indeed be constructed. Let $\mu = \frac{1}{K} S^\top 1_K$, and we have $1_N^\top \mu = 0$. Then,

$$\gamma := \frac{1}{\|\mu\|^2} S\mu = \frac{K}{\|1_K^\top S\|^2} SS^\top 1_K$$

satisfies $1_K^\top \gamma = K$, which is the second part of Condition 1(ii).

After recovering the consensus part, the remaining $\widetilde{S} = S - \alpha SS^\top 1_K 1_K^\top S$, where $\alpha = 1/\|1_K^T S\|^2$. Hence, $\widetilde{S} = SP^\perp_{S^\top 1_K}$ where $P^\perp_A$ is the projection matrix onto the orthogonal complement of $\mathrm{col}(A)$. Therefore, $1_K^\top \widetilde{S} = 1_K^\top S P^\perp_{S^\top 1_K} = 0^\top$ automatically holds. Take the SVD of $\widetilde{S}$ to get $\widetilde{S} = P\Sigma R^\top$. Let $V = \sqrt{N} R$ and $U = \frac{1}{\sqrt{N}} P\Sigma$, and Condition 1(iv) is satisfied. The rank r is also determined to be $\mathrm{rank}(\widetilde{S})$.

To verify the first part of Condition 1(ii), we left-multiply $1_K^\top$ to the equation $\widetilde{S} = UV^\top$ and get $(1_K^\top U)V^\top = 1_K^\top \widetilde{S} = 0^\top$, i.e., $V(1_K^\top U)^\top = 0$. But $V \in \mathbb{R}^{N \times r}$ is constructed to have full column rank, and consequently Condition 1(ii) must hold.

Finally, we study $V^\top \mu$. Note that by (8), $\widetilde{S}\mu = \frac{1}{K} SP^\perp_{S^\top 1_K} 1_K^\top S = 0$, and hence $UV^\top \mu = 0$. But U is also constructed to have full column rank, and $V^\top \mu$ must be zero. Taking the transpose yields Condition 1(iii). A similar argument verifies $V^\top 1_N = 0$.

Parts (ii) and (iii) jointly corroborate Proposition 2(ii). □

C Estimation

C.1 Rank selection

C.1.1 Rank selection methods

The rank r of the factorization $UV^\top$ serves as a structural hyperparameter that controls the model's capacity to capture structured disagreement around consensus, thereby affecting model complexity. We offer three ways to select the rank r of HJA. Note that all methods below allow the case of the $r = 0$, i.e., when our model is reduced to the sensitivity-only (JA) model.

Cross Validation. In experiments with dense comparison graphs, we can mask a validation subset of the pairwise comparisons $\Omega_{\text{val}} \subset \Omega$. For each candidate rank r, we estimate the parameters $\widehat{\theta}_r$ on the training set and select the r that minimizes the negative log-likelihood on Ω_{val}. This guards against overfitting by measuring generalizability and is hence suitable for empirical practices with real-world datasets. We adopt this method when dealing with real-world datasets in this paper; see Section F.1.

Spectral Residual Analysis. According to (8), estimation of the preference residuals $UV^\top$ as a whole does not depend on r. Therefore, as a straightforward diagnostic, r can be selected by analyzing the spectral decay of $UV^\top$, which can be estimated by fitting the consensus-only model ($S_0 = \gamma\mu^\top$) and computing $S - S_0$. A scree plot of the SVD of this residual matrix can help locate the optimal r, which typically corresponds to the elbow of the plot, reflecting the dominant aspects of disagreement. However, since the optimization is structural, this method may serve as a heuristic for future work. Another potential problem with this approach lies in the instability of $\widehat{S}$ directly estimated from the data, which is shown in Figure 2.

Information-theoretic approach: BIC. We also propose the BIC method for hyperparameter selection, which is statistically guaranteed to consistently estimate the true rank r^* of the underlying judge preference matrix $(UV^\top)^*$. Note that the BIC for a candidate rank r is defined as

$$\text{BIC}(r) = 2\mathcal{L}_n(\widehat{\theta}_r) + d_r \log n, \tag{C.1}$$

where $\mathcal{L}_n(\widehat{\theta}_r)$ is the minimized negative log-likelihood under rank r, and $d_r := r(K+N-r-3)$ is the effective degrees of freedom (having accounted for the identifiability constraints). The following theorem shows the consistency of BIC under the setting of Theorem 4.

Theorem C.1 (Consistency of rank selection)**.** *Under Assumptions A4 and A5 with fixed comparison graph and $\mu^* \neq 0$, the rank selected by minimizing BIC defined in (C.1) is statistically consistent, i.e.,*

$$\lim_{n\to\infty} \mathbb{P}\left\{\arg\min_{r\in\{0,\dots,R_{\max}\}} \text{BIC}(r) = r^*\right\} = 1.$$

We now provide the formal proof for Theorem C.1. The goal is to show that minimizing the Bayesian Information Criterion (BIC) consistently selects the true heterogeneity rank r^* as the total number of comparisons $n \to \infty$ while the comparison graph Ω remains fixed. Note that the sample log-likelihood is $-\mathcal{L}_n(\theta)$ throughout the paper.

Proof of Theorem C.1. Let $\mathbb{P}^*$ denote the true data-generating distribution parameterized by $\theta^* \in \Theta_{r^*}$, where Θ_r is the constrained parameter space corresponding to rank r (adhering to the identifiability constraints in Proposition 2). To prove $\lim_{n\to\infty} \mathbb{P}\left(r^* = \arg\min_r \text{BIC}(r)\right) = 1$, we must show that for any $r \neq r^*$, $\mathbb{P}\left(\text{BIC}(r) - \text{BIC}(r^*) > 0\right) \to 1$. We partition the analysis into two cases: the under-parameterized regime ($r < r^*$) and the over-parameterized regime ($r > r^*$).

Case 1: Under-parameterized regime ($r < r^*$). When $r < r^*$, the candidate model space Θ_r does not contain the true parameter θ^*. Because the BTL model is a strictly concave Generalized Linear Model (GLM) mapping bounded logits to probabilities, and our identifiability constraints make the parameterization injective, θ^* cannot be perfectly approximated by any $\theta \in \Theta_r$.

Following the argument of Section D.1 (especially (D.4)), by the properties of the KL divergence, the maximum expected log-likelihood under the restricted space Θ_r is strictly less than under the true parameter:

$$\delta_r := M(\theta^*) - \sup_{\theta \in \Theta_r} M(\theta) > 0$$

By the Uniform Law of Large Numbers (since Θ is compact and the BTL log-likelihood is continuous and bounded), we have $\frac{1}{n}\mathcal{L}_n(\widehat{\theta}_r) \xrightarrow{\text{p}} \sup_{\theta \in \Theta_r} M(\theta)$ and $\frac{1}{n}\mathcal{L}_n(\widehat{\theta}_{r^*}) \xrightarrow{\text{p}} M(\theta^*)$. Hence $\mathcal{L}_n(\widehat{\theta}_r) - \mathcal{L}_n(\widehat{\theta}_{r^*}) = n\delta_r + o_{\mathbb{P}}(n)$.

The difference in BIC is

$$\begin{aligned} \text{BIC}(r) - \text{BIC}(r^*) &= -2\left(\mathcal{L}_n(\widehat{\theta}_{r^*}) - \mathcal{L}_n(\widehat{\theta}_r)\right) + (d_r - d_{r^*})\log(n) \\ &= -2n\left(\frac{1}{n}\mathcal{L}_n(\widehat{\theta}_{r^*}) - \frac{1}{n}\mathcal{L}_n(\widehat{\theta}_r)\right) + (d_r - d_{r^*})\log(n). \end{aligned} \tag{C.2}$$

Since the term inside the first parenthesis converges in probability to $\delta_r > 0$, the likelihood difference scales as $\mathcal{O}_p(n)$. But the penalty difference $(d_r - d_{r^*})\log(n)$ only scales as $\mathcal{O}(\log n)$, making the linear growth of the likelihood deficit dominate. Thus, for any $r < r^*$,

$$\lim_{n\to\infty} \mathbb{P}\left(\text{BIC}(r) - \text{BIC}(r^*) > 0\right) = 1.$$

Case 2: Over-parameterized regime ($r > r^*$) When $r > r^*$, the candidate model space Θ_r strictly nests the true model space Θ_{r^*} in terms of closure (i.e., $\overline{\Theta_{r^*}} \subset \overline{\Theta_r}$). Therefore, $\theta^* \in \overline{\Theta_r}$, and the model can be viewed as correctly specified but over-parameterized.

Because the true model is nested within the larger model, we invoke theory on the Likelihood Ratio (LR) test. Under the identifiability constraints (Section 4.2) and assumptions required by asymptotic normality, the Fisher Information Matrix evaluated at θ^* is non-singular. Note that under the null hypothesis that the true parameter lies in the restricted space Θ_{r^*}, the log-likelihood ratio statistic may not converge in law to a chi-squared distribution, as the true parameter θ^* essentially lies on the boundary of Θ_r. However, $-2\left(\mathcal{L}_n(\widehat{\theta}_r) - \mathcal{L}_n(\widehat{\theta}_{r^*})\right)$ is still $\mathcal{O}_p(1)$ by theory on the boundary behaviour of LR [7, Theorem 3] and tests on reduced-rank regression [8, Theorem 3.2]. Therefore, the maximum improvement in log-likelihood gained by fitting $r - r^*$ spurious components is bounded in probability. But $r > r^*$ leads to $d_r > d_{r^*}$ and consequently the penalty term $(d_r - d_{r^*})\log(n) \to \infty$ as $n \to \infty$. Thus, for any $r > r^*$:

$$\lim_{n\to\infty} \mathbb{P}\left(\text{BIC}(r) - \text{BIC}(r^*) > 0\right) = 1.$$

Combining Case 1 and Case 2, the true rank r^* strictly minimizes the BIC with probability tending to 1 as the sample size $n \to \infty$, i.e., $\lim_{n\to\infty} \mathbb{P}\left(\arg\min_r \text{BIC}(r) = r^*\right) = 1$. □

Note that the theory above also applies to the case of $r^* = 0$. We also add a proof-of-concept simulation on the rank selection consistency of BIC under the oracle model; see Figure C1.

C.1.2 Practical suggestions and reduction to simpler models

Based on the theory and simulations, we prefer BIC when the true data fit our method well or when computational resources are too limited to run cross-validation.

It is possible that the selected rank is zero, indicating that a simpler model might suffice and be preferred for sake of parsimony. We provide a targeted model-reduction experiment with 3000 balanced comparisons and five-fold validation over $r \in \{0, 1, 2, 3\}$; see Table C1.

Thus, cross-validation does not introduce a disagreement component merely to absorb random variation: it selects $r = 0$ when the data contain no meaningful residual structure, and $r > 0$ in every structured run. A selected $r = 0$ implies the data do not support a reproducible $UV^\top$ component, though not that population heterogeneity is exactly zero. After selecting $r = 0$, we evaluate the further reduction to pooled BTL by checking whether a joint 95% confidence region for the $K - 1$ contrasts of $\gamma - 1_K$ contains zero; see Table C2.

This suggests a practical hierarchy: select r by validation likelihood or BIC; if $r = 0$, assess whether $\gamma = 1_K$ is compatible; and report held-out negative log-likelihood alongside either simplification.

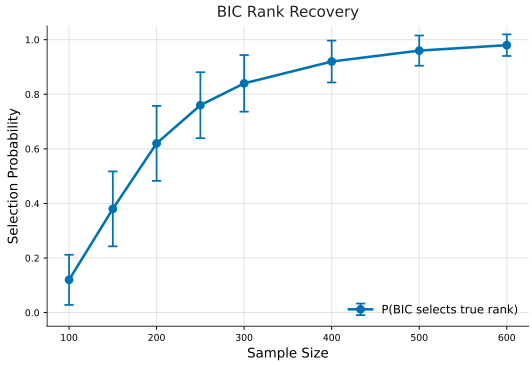


Figure C1: BIC rank selection consistency. The selection probability (y-axis) is estimated by the frequency of correct rank estimation over 50 repetitions. Error bars denote $\pm$ twice the standard error.

Table C1: Testing reduction to sensitivity-only model ($r^* = 0$)

Data-generating regime	$r = 0$ **selected**	**Interpretation**
Exact sensitivity-only model	50/50	No residual heterogeneity
Very small residual component	48/50	Residual signal is negligible
10% independent label flips	48/50	Unstructured noise only
Structured rank-one alternative	0/50	Reproducible residual heterogeneity

C.2 Algorithm

C.2.1 Initialization

We initialize our parameters by fitting a pooled BTL model for μ, all ones for γ, and a judgewise BTL model for each row of S. See Algorithm C.1.

Algorithm C.1 INITIALIZE: construct $(\gamma^{(0)}, \mu^{(0)}, U^{(0)}, V^{(0)})$

Require: Aggregated counts $\{(n_{kij}, Y_{kij})\}_{(k,i,j)\in\Omega}$, target rank r

Ensure: Initial value $\theta^{(0)} = (\gamma^{(0)}, \mu^{(0)}, U^{(0)}, V^{(0)}) \in \Theta_r$

1: Fit a pooled centered BTL model to obtain $\mu^{(0)}$.
2: Re-center $\mu^{(0)} \leftarrow \mu^{(0)} - \frac{1}{N}(1_N^\top \mu^{(0)})1_N$.
3: Set $\gamma^{(0)} \leftarrow 1_K$.
4: **for** $k = 1, \ldots, K$ **do**
5: Fit a centered judgewise BTL model using judge k's comparisons to obtain $\widetilde{S}^{(0)}_{k,:}$.
6: **end for**
7: Form the initial residual matrix $R^{(0)} \leftarrow \widetilde{S}^{(0)} - \gamma^{(0)}(\mu^{(0)})^\top$.
8: Judge-center the residual: $R_c^{(0)} \leftarrow R^{(0)} - \frac{1}{K}1_K 1_K^\top R^{(0)}$.
9: Compute the rank-r truncated SVD $R_c^{(0)} \approx P_r \Sigma_r Q_r^\top$.
10: Set $U^{(0)} \leftarrow P_r \Sigma_r^{1/2}$ and $V^{(0)} \leftarrow Q_r \Sigma_r^{1/2}$.
11: Compute $(\gamma^{(0)}, \mu^{(0)}, U^{(0)}, V^{(0)}) \leftarrow$ REANCHOR$(\gamma^{(0)}, \mu^{(0)}, U^{(0)}, V^{(0)})$ based on Algorithm C.2.
12: **return** $\theta^{(0)} = (\gamma^{(0)}, \mu^{(0)}, U^{(0)}, V^{(0)})$.

C.2.2 Re-anchor and enforce identifiability

The re-anchoring step restores the nonlinear identification constraints without changing the implied score matrix $\widetilde{S} = \widetilde{\gamma}\widetilde{\mu}^\top + \widetilde{U}\widetilde{V}^\top$. Concretely, REANCHOR removes the component of $\widetilde{V}$ along $\widetilde{\mu}$, compensates for this change through the update $\gamma^+ = \widetilde{\gamma} + \widetilde{U}a$, and then refactorizes the heterogeneity term by a normalized SVD. If the candidate loading violates the local nondegeneracy conditions,

Table C2: Further reduction to pooled-BTL

Data-generating regime	Zero contained	Suggested reduction
Exact pooled model	49/50	Pooled BTL
Slight sensitivity variation	46/50	Often compatible with pooled BTL
Independent symmetric label flips	45/50	Often compatible with pooled BTL
Material sensitivity heterogeneity	0/50	Retain judge-specific γ

Algorithm C.2 REANCHOR: canonical re-anchoring with feasibility check

Require: Tentative $(\widetilde{\gamma}, \widetilde{\mu}, \widetilde{U}, \widetilde{V})$ from the two block updates, tolerance $\delta_\mu > 0$, spectral tolerance $\delta_\sigma > 0$.

Ensure: Either a feasible anchored point $(\gamma^+, \mu^+, U^+, V^+) \in \Theta_r$ or FAIL.

Set $\mu^+ \leftarrow \widetilde{\mu}$.
if $\|\mu^+\|_2 < \delta_\mu$ **then**
  return FAIL ▷ Avoid instability incurred on updating V
end if
Compute $a \leftarrow \widetilde{V}^\top \mu^+ / \|\mu^+\|_2^2$.
Set $\bar{V} \leftarrow \widetilde{V} - \mu^+ a^\top$ and $\gamma^+ \leftarrow \widetilde{\gamma} + \widetilde{U} a$.
Form $\bar{H} \leftarrow \widetilde{U} \bar{V}^\top$.
Compute the thin SVD $\bar{H} = P \Sigma Q^\top$, with diagonal entries of Σ arranged in decreasing order.
if $\sigma_r(\Sigma) < \delta_\sigma$ **then**
  return FAIL ▷ Avoid ill-conditioned matrices caused by a large leap
end if
Set $U^+ \leftarrow P\Sigma/\sqrt{N}$, and $V^+ \leftarrow \sqrt{N}\, Q$. Flip the corresponding columns of (U^+, V^+) so that the first nonzero entry of each column of U^+ is positive.
return $(\gamma^+, \mu^+, U^+, V^+)$

the routine returns FAIL, and the main algorithm increases the proximal parameter before recomputing the block updates.

Unlike JA-Ranking [1], we do not log-reparameterize the judge-specific coefficient γ. In our model, γ is an affine-anchored consensus-sensitivity vector recovered from the structured score matrix, not a strictly positive global discrimination parameter. A log transform would impose unnecessary positivity and would not be compatible with the S-preserving re-anchoring map. The proximal terms make the block updates well posed and provide descent, while re-anchoring restores the canonical representative without changing the likelihood.

In exact arithmetic, the post-anchor candidate automatically satisfies $1_K^\top \gamma^+ = K$ because the judge block enforces $1_K^\top \widetilde{\gamma} = K$ and $1_K^\top \widetilde{U} = 0_r^\top$.

C.2.3 Confidence intervals for target functionals

Algorithm C.3 ESTIMATEFISHERINFO: plug-in Fisher information

Require: Final estimator $\widehat{\theta} = (\widehat{\gamma}, \widehat{\mu}, \widehat{U}, \widehat{V})$, data $\{(n_{kij}, Y_{kij})\}_{(k,i,j)\in\Omega}$.

Ensure: Plug-in Fisher information matrix $\mathcal{I}_n(\widehat{\theta})$.

Use the local identified parameterization fixed in Section 5.1; all gradients below are taken in this parameterization.
for each $(k, i, j) \in \Omega$ **do**
  Compute $\widehat{\eta}_{kij} = \widehat{\gamma}_k(\widehat{\mu}_i - \widehat{\mu}_j) + \widehat{U}_{k,:}(\widehat{V}_{i,:} - \widehat{V}_{j,:})^\top$.
  Set $\widehat{p}_{kij} = \{1 + \exp(-\widehat{\eta}_{kij})\}^{-1}$.
  Compute the local gradient $\widehat{g}_{kij} = \nabla_\theta \eta_{kij}(\widehat{\theta})$.
end for
Set $\mathcal{I}_n(\widehat{\theta}) \leftarrow \sum_{(k,i,j)\in\Omega} \frac{n_{kij}}{n} \widehat{p}_{kij}(1 - \widehat{p}_{kij}) \widehat{g}_{kij} \widehat{g}_{kij}^\top$.
return $\mathcal{I}_n(\widehat{\theta})$.

Algorithm C.4 UNCERTAINTYQUANTIFICATION: delta-method confidence intervals

Require: Final estimator $\widehat{\theta}$, plug-in information matrix $\mathcal{I}_n(\widehat{\theta})$, target functionals $\mathcal{Q}$, confidence level $1-\alpha$.
Ensure: Wald confidence intervals for $\{q(\theta) : q \in \mathcal{Q}\}$.
1: Use the local identified parameterization fixed in Section 5.1.
2: Set $\widehat{\Sigma}_\theta \leftarrow \frac{1}{n}\mathcal{I}_n(\widehat{\theta})^{-1}$.
3: **for** each smooth scalar target $q \in \mathcal{Q}$ **do**
4: Compute $\widehat{q} = q(\widehat{\theta})$.
5: Compute the local gradient $\widehat{a}_q = \nabla_\theta q(\widehat{\theta})$.
6: Compute $\widehat{\mathrm{se}}(\widehat{q}) = \left(\widehat{a}_q^\top \widehat{\Sigma}_\theta \widehat{a}_q\right)^{1/2}$.
7: Report $\widehat{q} \pm z_{1-\alpha/2}\widehat{\mathrm{se}}(\widehat{q})$.
8: **end for**

C.3 Convergence analysis

Assumption A1 (Local algorithmic chart)**.** There exists a compact set $\Theta_{r,0} \subset \Theta_r$ such that: (i) θ^* lies in the relative interior of $\Theta_{r,0}$ after fixing the local anchor signs of the columns of U and V; (ii) for every $\theta = (\gamma, \mu, U, V) \in \Theta_{r,0}$, one has $\|\mu\|_2 \geq \delta_\mu$, and the nonzero singular values of $UV^\top$ are separated from each other and from zero by at least δ_σ; (iii) the re-anchoring map is well-defined and maps $\Theta_{r,0}$ into itself.

Assumption A2 (Local isolated minimizer)**.** For the empirical criterion $\mathcal{L}_n$, there exists a unique local minimizer $\widehat{\theta}_n \in \Theta_{r,0}$, after fixing the anchor signs, and $\nabla^2_\xi \mathcal{L}_n(\widehat{\theta}_n)$ is positive definite in the local free-coordinate chart ξ.

Assumption A3 (Initialization in basin)**.** The initialized iterate $\theta^{(0)}$ lies in the basin of attraction of $\widehat{\theta}_n$ within $\Theta_{r,0}$.

Proof of Theorem 3. The proof of Theorem 3 follows from Proposition C.2. □

Proposition C.2 (Convergence of the proximal anchored alternating algorithm)**.** *Consider the sequence $\{\theta^{(t)}\}_{t\geq 0}$ generated by Algorithm 1, where each block subproblem includes the proximal terms*

$$\frac{\tau_t}{2}\|\gamma - \gamma^{(t)}\|_2^2 + \frac{\tau_t}{2}\|U - U^{(t)}\|_F^2 \quad \text{and} \quad \frac{\tau_t}{2}\|\mu - \mu^{(t)}\|_2^2 + \frac{\tau_t}{2}\|V - V^{(t)}\|_F^2,$$

with proximal parameters uniformly bounded away from zero ($\tau_t \geq \underline{\tau} > 0$) if τ is allowed to change during different iterations.

Part I: Global descent properties. *Suppose the structural assumptions (Assumption A1) hold. Then:*

1. *the sequence $\{\mathcal{L}_n(\theta^{(t)})\}_{t\geq 0}$ is nonincreasing and convergent;*
2. *the iterates satisfy $\sum_{t=0}^{\infty} \|\theta^{(t+1)} - \theta^{(t)}\|^2 < \infty$.*

Part II: Local sequence convergence. *Suppose further that Assumptions A2 and A3 concerning the basin hold. Then:*

3. *the iterates remain in $\Theta_{r,0}$ and the full anchored sequence $\theta^{(t)}$ converges to $\widehat{\theta}_n$;*
4. *equivalently, before anchoring, the factor pair $(U^{(t)}, V^{(t)})$ converges to the equivalence class of $(\widehat{U}_n, \widehat{V}_n)$ up to the column-sign convention, and after re-anchoring the convergence is ordinary parameter convergence.*

Part III: Local linear rate. *Following Part II, the convergence is locally linear: there exist constants $C > 0$ and $\rho \in (0, 1)$ such that*

$$\|\theta^{(t)} - \widehat{\theta}_n\| \leq C\rho^t \quad \text{and} \quad \mathcal{L}_n(\theta^{(t)}) - \mathcal{L}_n(\widehat{\theta}_n) \leq C\rho^t$$

for all sufficiently large t.

In particular, whenever the sample criterion admits a unique local minimizer $\widehat{\theta}_n$ in the identified chart $\Theta_{r,0}$, the output of Algorithm 1 coincides with that local constrained MLE for all sufficiently accurate runs.

Proof of Proposition C.2. Let the parameter space be partitioned into two blocks: the judge block $\phi_1 = (\gamma, U)$ and the item block $\phi_2 = (\mu, V)$, such that the full parameter vector is $\theta = (\phi_1, \phi_2)$.

Part I: Global Descent Properties. This part relies only on the structural conditions in Assumption A1 as the properties of proximal alternating minimization.

Claim 1 (Nonincreasing and convergent objective): For the judge block $\phi_1^{(t+1)}$, the update satisfies:

$$\phi_1^{(t+1)} = \underset{\phi_1}{\operatorname{argmin}} \left\{ \mathcal{L}_n(\phi_1, \phi_2^{(t)}) + \frac{\tau_t}{2} \|\phi_1 - \phi_1^{(t)}\|^2 \right\}.$$

Comparing the value at the minimum $\phi_1^{(t+1)}$ against the previous iterate $\phi_1^{(t)}$ yields the sufficient decrease inequality for the first block:

$$\mathcal{L}_n(\phi_1^{(t+1)}, \phi_2^{(t)}) \leq \mathcal{L}_n(\phi_1^{(t)}, \phi_2^{(t)}) - \frac{\tau_t}{2} \|\phi_1^{(t+1)} - \phi_1^{(t)}\|^2.$$

Applying identical logic to the item block update $\phi_2^{(t+1)}$ yields

$$\mathcal{L}_n(\phi_1^{(t+1)}, \phi_2^{(t+1)}) \leq \mathcal{L}_n(\phi_1^{(t+1)}, \phi_2^{(t)}) - \frac{\tau_t}{2} \|\phi_2^{(t+1)} - \phi_2^{(t)}\|^2.$$

Adding these two inequalities provides the global sufficient decrease condition for a full iteration:

$$\mathcal{L}_n(\theta^{(t+1)}) \leq \mathcal{L}_n(\theta^{(t)}) - \frac{\tau_t}{2} \|\theta^{(t+1)} - \theta^{(t)}\|^2 \leq \mathcal{L}_n(\theta^{(t)}). \tag{C.3}$$

The re-anchoring step (Algorithm C.2 and by Assumption A1) acts purely on the rotational and scaling equivalence classes of the bilinear factorization. Because it is an invariant transformation with respect to S, the objective value $\mathcal{L}_n$ remains unchanged before and after re-anchoring. Therefore, the sequence of objective values $\{\mathcal{L}_n(\theta^{(t)})\}$ is monotonically nonincreasing. Since the negative log-likelihood is bounded from below (by zero), the sequence converges to some limit $L^* \geq 0$.

Claim 2 (Square-summability): Summing the sufficient decrease inequality (C.3) from iteration $t = 0$ to T yields a telescoping sum:

$$\sum_{t=0}^{T} \frac{\tau_t}{2} \|\theta^{(t+1)} - \theta^{(t)}\|^2 \leq \mathcal{L}_n(\theta^{(0)}) - \mathcal{L}_n(\theta^{(T+1)}) \leq \mathcal{L}_n(\theta^{(0)}).$$

Using the strict lower bound $\tau_t \geq \underline{\tau} > 0$ and taking the limit as $T \to \infty$, we obtain:

$$\sum_{t=0}^{\infty} \|\theta^{(t+1)} - \theta^{(t)}\|^2 \leq \frac{2}{\underline{\tau}} \mathcal{L}_n(\theta^{(0)}) < \infty.$$

This finite sum guarantees square-summability of the iterate differences, which also trivially implies that the step sizes converge to zero: $\|\theta^{(t+1)} - \theta^{(t)}\| \to 0$ as $t \to \infty$.

Part II: Local Sequence Convergence. This part invokes the statistical and initialization conditions in Assumptions A2 and A3.

Claim 3 (Convergence to the local minimizer $\widehat{\theta}_n$): By Assumption A2, $\widehat{\theta}_n$ is an isolated local minimizer. Therefore, there exists a compact neighborhood $\mathcal{B} \subset \Theta_{r,0}$ around $\widehat{\theta}_n$ such that $\widehat{\theta}_n$ is the strictly unique stationary point within $\mathcal{B}$, and $\mathcal{L}_n(\theta) > \mathcal{L}_n(\widehat{\theta}_n)$ for all $\theta \in \mathcal{B} \setminus \{\widehat{\theta}_n\}$.

By Assumption A3, the initialization $\theta^{(0)}$ lies close enough to $\widehat{\theta}_n$ such that the connected component of the sublevel set $\mathcal{S} := \{\theta \in \mathcal{B} \mid \mathcal{L}_n(\theta) \leq \mathcal{L}_n(\theta^{(0)})\}$ is strictly contained in the interior of $\mathcal{B}$, given the positive definiteness of the Hessian matrix. Because $\mathcal{B}$ is compact and $\mathcal{L}_n$ is continuous, the sublevel set $\mathcal{S}$ is also compact.

By the global descent property established in Part I, $\mathcal{L}_n(\theta^{(t+1)}) \leq \mathcal{L}_n(\theta^{(t)})$ for all t. Thus, the sequence cannot enter the region where $\mathcal{L}_n(\theta) > \mathcal{L}_n(\theta^{(0)})$, meaning the entire sequence $\{\theta^{(t)}\}_{t=0}^{\infty}$ is strictly confined within the compact set $\mathcal{S}$. The re-anchoring step under Assumption A1 preserves this containment.

Because $\mathcal{S}$ is compact, the sequence $\{\theta^{(t)}\}$ must possess at least one limit point $\theta^* \in \mathcal{S}$ by the Bolzano-Weierstrass theorem. To characterize any such limit point, we invoke the first-order optimality conditions of the exact proximal block updates. For the judge block,

$$\nabla_{\phi_1} \mathcal{L}_n(\phi_1^{(t+1)}, \phi_2^{(t)}) + \tau_t(\phi_1^{(t+1)} - \phi_1^{(t)}) = 0.$$

Taking the norm yields $\|\nabla_{\phi_1} \mathcal{L}_n(\phi_1^{(t+1)}, \phi_2^{(t)})\| = \tau_t \|\phi_1^{(t+1)} - \phi_1^{(t)}\|$. By Part I and the boundedness of τ_t, the right-hand side converges to zero, and thus the gradient mapping uniformly converges to zero. Because $\mathcal{L}_n$ is continuously differentiable, any subsequence converging to θ^* satisfies $\nabla \mathcal{L}_n(\theta^*) = 0$.

Therefore, every limit point of the sequence is a stationary point. But $\mathcal{S} \subset \mathcal{B}$, and by construction, $\widehat{\theta}_n$ is the strictly unique stationary point within $\mathcal{B}$. Therefore, $\theta^* = \widehat{\theta}_n$ is the unique limit point of the sequence. Since $\{\theta^{(t)}\}$ is a sequence in a compact set with exactly one limit point, the entire sequence globally converges to $\widehat{\theta}_n$.

Claim 4 (Pre- and post-anchoring equivalence): Because the joint parameters converge, the sequence of latent evaluation scores $S^{(t)} = \gamma^{(t)}(\mu^{(t)})^\top + U^{(t)}(V^{(t)})^\top$ converges to the optimal empirical score matrix $\widehat{S}_n$. Before the re-anchoring step is applied, the unanchored factor matrices $(U^{(t)}, V^{(t)})$ converge to the equivalence class $[\widehat{U}_n, \widehat{V}_n]$ defined by orthogonal rotations and scale-sign permutations. The re-anchoring map is a continuous function acting on $\Theta_{r,0}$ that uniquely identifies a representative from this equivalence class by fixing local anchor signs, and hence guarantees the ordinary parameter-wise convergence in Euclidean space.

Part III: Local linear rate. This part draws the final conclusion on local convergence rate from the assumed and established properties of the objective function, the algorithm, and the topological structures of the global optimum.

We restrict the domain of the objective function in $\mathcal{S}$, where $\mathcal{L}$ demonstrates strong convexity. By [9, Theorem 2], $\mathcal{L}$ satisfies the Polyak-Łojasiewicz (PL) inequality [10], and by Appendix F of [9], satisfies the proximal-PL property, and by Appendix G of [9], has a local Kurdyka-Łojasiewicz (KL) exponent $\theta = 1/2$. The local linear convergence rate hence directly follows from [11, Theorem 3.4]. □

D Statistical Inference

D.1 Consistency of constrained MLE

Recall that the structural parameter is estimated by the constrained maximum likelihood estimator

$$\widehat{\theta} := (\widehat{\gamma}, \widehat{\mu}, \widehat{U}, \widehat{V}) \in \arg\min_{\theta \in \Theta_r} \mathcal{L}(S(\theta)), \qquad S(\theta) := \gamma\mu^\top + UV^\top.$$

Since the likelihood depends on θ only through the induced score matrix $S(\theta)$, the proof is most naturally carried out on the centered score space

$$\mathcal{S} := \{S \in \mathbb{R}^{K\times N} : S1_N = 0, \text{ and Condition 2 holds}\}.$$

The signs of the columns of (U, V) can be anchored and fixed via, say, Remark 1, and then Proposition 2 ensures $\mathcal{S} = \{S(\theta) : \theta \in \Theta_r\}$.

We formalize the regularity conditions mentioned in Section 5.2 as the assumptions below.

Assumption A4 (Asymptotic sampling fractions)**.** There exist fixed constants $w_{kij} \geq 0$ such that $n_{kij}/n \xrightarrow{\text{p}} w_{kij}$ for all $(k, i, j) \in \Omega$ as $n \to \infty$.

If n_{kij} is non-random, i.e., the design of the experiment is deterministic, the convergence in probability is reduced to convergence in the ordinary sense. Note that a triple can be observed ($n_{kij} > 0$) but still have asymptotic fraction $w_{kij} = 0$. To account for this in subsequent analysis, we define:

$$\Omega_w := \{(k, i, j) \in \Omega : w_{kij} > 0\} \tag{D.1}$$

as the set of asymptotically active triples.

Assumption A5 (Graph connectivity)**.** For each judge $k \in [K]$, the comparison graph $\mathcal{G}_k^w = ([N], \mathcal{E}_k^w)$ with edge set $\mathcal{E}_k^w = \{\{i, j\} : (k, i, j) \in \Omega_w\}$ is connected, and $0 < p^*_{kij} < 1$ for each active triple $(k, i, j) \in \Omega_w$.

By connectedness, we mean an undirected graph where a path (a sequence of connecting edges) exists between every pair of nodes (items). Therefore, under Assumption A5, all the items assessed by a judge should be linked via one or more comparisons. However, it is not required that all possible pairs be directly compared, and this scenario is accounted for via Ω_w.

These assumptions can guarantee the consistency of $\widehat{\theta}$, which is shown below.

Proof of Theorem 4: consistency part. For $S \in \mathcal{S}$, write

$$\eta_{kij}(S) := S_{ki} - S_{kj}, \qquad p_{kij}(S) := \frac{1}{1 + \exp\{-\eta_{kij}(S)\}}.$$

Also write $\bar{Y}_{kij} := Y_{kij}/n_{kij}$. Then the normalized negative log-likelihood, viewed as a function of S, is

$$M_n(S) := \frac{1}{n}\mathcal{L}(S) = \sum_{(k,i,j)\in\Omega} \frac{n_{kij}}{n}\Big[- \bar{Y}_{kij}\eta_{kij}(S) + \log\big(1 + \exp\{\eta_{kij}(S)\}\big)\Big]. \tag{D.2}$$

By the strong law, for each fixed $(k, i, j) \in \Omega$,

$$\bar{Y}_{kij} \xrightarrow{\text{a.s.}} p^*_{kij},$$

and by Assumption A4,

$$\frac{n_{kij}}{n} \to w_{kij}.$$

Since Ω is fixed and each summand in (D.2) is continuous in S, it follows that $M_n(S)$ converges uniformly on every compact subset of $\mathcal{S}$ to

$$M(S) := \sum_{(k,i,j)\in\Omega} w_{kij}\Big[- p^*_{kij}\eta_{kij}(S) + \log\big(1 + \exp\{\eta_{kij}(S)\}\big)\Big]. \tag{D.3}$$

Let $S^* = S(\theta^*) \in \mathcal{S}$. Then, by the usual Bernoulli Kullback–Leibler identity,

$$M(S) - M(S^*) = \sum_{(k,i,j)\in\Omega} w_{kij}\,\mathrm{KL}\big(\mathrm{Bern}(p^*_{kij}) \,\|\, \mathrm{Bern}(p_{kij}(S))\big) \geq 0. \tag{D.4}$$

Equality holds if and only if $p_{kij}(S) = p^*_{kij}$ for all $(k,i,j) \in \Omega_w$, equivalently,

$$S_{ki} - S_{kj} = S^*_{ki} - S^*_{kj} \qquad \text{for all } (k,i,j) \in \Omega_w.$$

By Lemma D.3, equality in (D.4) implies $S = S^*$. Hence S^* is the unique minimizer of M on $\mathcal{S}$.

We first prove the consistency of the score-space MLE $\widehat{S} \in \arg\min_{S\in\mathcal{S}} \mathcal{L}(S)$.

Fix any open neighborhood $\mathcal{O}$ of S^* in $\mathcal{S}$. Since M is continuous and uniquely minimized at S^*, there exists $\delta > 0$ such that

$$\inf_{S\in\mathcal{S}\setminus\mathcal{O}} M(S) \geq M(S^*) + 3\delta. \tag{D.5}$$

By Lemma D.4, there exists $R < \infty$ such that

$$\inf_{S\in\mathcal{S}:\,\|S\|_F > R} M(S) \geq M(S^*) + 4\delta. \tag{D.6}$$

Define the truncated score set

$$K_R := \{S \in \mathcal{S} : \|S\|_F \leq R\}.$$

Since K_R is closed and bounded in the finite-dimensional Euclidean space $\mathbb{R}^{K\times N}$, it is compact. By uniform convergence on compact subsets,

$$\sup_{S\in K_R} |M_n(S) - M(S)| \xrightarrow{\mathrm{p}} 0.$$

Hence, with probability tending to one,

$$\sup_{S\in K_R} |M_n(S) - M(S)| < \delta.$$

On this event, for every $S \in K_R$,

$$M_n(S^*) \leq M(S^*) + \delta,$$

whereas for every $S \in K_R \setminus \mathcal{O}$, by (D.5),

$$M_n(S) \geq M(S) - \delta \geq M(S^*) + 2\delta.$$

Thus no minimizer of M_n over K_R can lie in $K_R \setminus \mathcal{O}$.

Further, by the sample coercivity part of Lemma D.4, with probability tending to one,

$$\inf_{S\in\mathcal{S}:\,\|S\|_F > R} M_n(S) > M(S^*) + 2\delta.$$

Hence no global minimizer of M_n over $\mathcal{S}$ can lie outside K_R. Combining the two conclusions, any global minimizer $\widehat{S}$ of M_n over $\mathcal{S}$ must belong to $\mathcal{O}$ with probability tending to one, i.e.,

$$\mathbb{P}(\widehat{S} \in \mathcal{O}) \to 1.$$

Since $\mathcal{O}$ was arbitrary, we obtain

$$\widehat{S} \xrightarrow{\mathrm{p}} S^*.$$

We now pass the consistency in the score space into the parameter space after fixing the sign convention of Remark 1. By Lemma D.4, the criterion M_n remains coercive on $\mathcal{S}$ with high probability, ensuring that any global minimizer of M_n belongs to the identified score space $\mathcal{S}$.

Define the structured score map

$$\Phi : \Theta_r \to \mathcal{S}, \quad \Phi(\theta) = \gamma\mu^\top + UV^\top.$$

By Proposition 2, after fixing column signs, Φ is invertible with smooth inverse $\Phi^{-1} : \mathcal{S} \to \Theta_r$. Since both Φ and its inverse are continuous, the continuous mapping theorem yields

$$\widehat{\theta} = \Phi^{-1}(\widehat{S}) \xrightarrow{\mathrm{p}} \Phi^{-1}(S^*) = \theta^*.$$

□

D.2 Asymptotic normality and nondegeneracy of the Fisher information

The asymptotic normality of the constrained maximum likelihood estimator requires the nonsingularity of the Fisher information matrix

$$I_\Theta(\theta^*) = \sum_{(k,i,j)\in\Omega_w} w_{kij} p^*_{kij}(1-p^*_{kij}) \nabla_\theta \eta_{kij}(\theta^*) \nabla_\theta \eta_{kij}(\theta^*)^\top.$$

In this section, we first argue that such nondegeneracy can be derived from the active comparison graph structure together with the local regularity of the structured score map (Proposition D.1), and then we establish the asymptotic normality. The argument is inspired by the graph-Laplacian viewpoint for pairwise-comparison Fisher information developed in [12], where positivity of the Fisher information is obtained from graph connectivity (after quotienting out the additive invariance of pairwise-comparison models). In our setting, the argument proceeds in two layers:

1. establish positive definiteness of the Fisher information in the ambient row-centered score space $\mathcal{S}$ via judge-wise connected active graphs; and
2. pull this positivity back to the structured parameter space Θ_r through the differential of the map $\Phi(\theta) = S(\theta) = \gamma\mu^\top + UV^\top$.

The second step is needed because our inferential target is not the unrestricted score matrix S, but the constrained latent parameter $\theta = (\gamma, \mu, U, V)$.

We first show the local regularity which passes score-space nondegeneracy to parameter-space nondegeneracy.

Proposition D.1 (Local map injectivity)**.** *Let $\theta^* = (\gamma^*, \mu^*, U^*, V^*) \in \Theta_r$ denote the true parameter with $\mu^* \neq 0$. After fixing the column signs of (U, V) according to a local sign convention, there exists a neighborhood of θ^* in Θ_r that is parameterized by a smooth local chart on the manifold. Moreover, the differential of the structured score map*

$$\Phi : \Theta_r \to \mathcal{S}, \qquad \Phi(\theta) = S(\theta) = \gamma\mu^\top + UV^\top,$$

is injective at θ^* on the tangent space $T_{\theta^*}\Theta_r$. That is,

$$D\Phi(\theta^*)[\delta\theta] = 0,\ \delta\theta \in T_{\theta^*}\Theta_r \quad \Longrightarrow \quad \delta\theta = 0.$$

Proof of Proposition D.1. The proof is divided into two steps.

Step 1: Local chart of Θ_r. By Condition 1, near θ^* the set Θ_r is the zero set of

$$\psi(\theta) := \Big(1_K^\top\gamma - K,\ 1_N^\top\mu,\ 1_K^\top U,\ V^\top 1_N,\ \mu^\top V,\ N^{-1}V^\top V - I_r,\ \mathrm{offdiag}(U^\top U)\Big),$$

where $\mathrm{offdiag}(\cdot)$ keeps the off-diagonal entries, since the remaining requirement on the diagonal of $K^{-1}U^\top U$ is an open condition. Its differential at θ^* is

$$D\psi(\theta^*)[\delta\theta] = \Big(1_K^\top\delta\gamma,\ 1_N^\top\delta\mu,\ 1_K^\top\delta U,\ \delta V^\top 1_N,\ \delta\mu^\top V^* + \mu^{*\top}\delta V,\ N^{-1}(\delta V^\top V^* + V^{*\top}\delta V),\ \mathrm{offdiag}(\delta U^\top U^* + U^{*\top}\delta U)\Big).$$

We show that $D\psi(\theta^*)$ is surjective. Let $D = K^{-1}U^{*\top}U^* = \mathrm{diag}(d_1, \ldots, d_r)$ as in Condition 1. For any target $(a_1, a_2, b_1^\top, b_2, b_3^\top, \Xi_1, \Xi_2)$, where Ξ_1 is symmetric and Ξ_2 is symmetric with zero diagonal, take

$$\delta\gamma = \frac{a_1}{K}1_K, \quad \delta\mu = \frac{a_2}{N}1_N, \quad \delta U = \frac{1}{K}1_K b_1^\top + U^* X, \quad \delta V = \frac{1}{N}1_N b_2^\top + \frac{\mu^* b_3^\top}{\|\mu^*\|_2^2} + \frac{1}{2}V^*\Xi_1,$$

where $X_{ii} = 0$ and $X_{ij} = (\Xi_2)_{ij}/(2Kd_i)$ for $i \neq j$; this is well defined since $d_i > 0$ and $\mu^* \neq 0$. By Condition 1, the first six components of $D\psi(\theta^*)[\delta\theta]$ match the target directly, and for the last one,

$$\delta U^\top U^* + U^{*\top}\delta U = K(X^\top D + DX), \qquad K(d_j X_{ji} + d_i X_{ij}) = (\Xi_2)_{ij} \quad (i \neq j).$$

Hence $D\psi(\theta^*)$ is surjective. By the regular value theorem, after fixing the column signs of (U, V), Θ_r is a smooth embedded submanifold near θ^*, with

$$T_{\theta^*}\Theta_r = \ker D\psi(\theta^*).$$

Step 2: Injectivity of $D\Phi(\theta^*)$. Let $\delta\theta \in \ker D\psi(\theta^*)$ satisfy

$$D\Phi(\theta^*)[\delta\theta] = \delta\gamma\, {\mu^*}^\top + \gamma^*\, \delta\mu^\top + \delta U\, {V^*}^\top + U^*\, \delta V^\top = 0.$$

Left-multiplying by $K^{-1}1_K^\top$ and using $\delta\theta \in \ker D\psi(\theta^*)$ and Condition 1, we get

$$0 = \frac{1}{K}(1_K^\top \delta\gamma){\mu^*}^\top + \frac{1}{K}(1_K^\top \gamma^*)\delta\mu^\top + \frac{1}{K}(1_K^\top \delta U){V^*}^\top + \frac{1}{K}(1_K^\top U^*)\delta V^\top = \delta\mu^\top.$$

Then right-multiplying by μ^*, and noting $\delta V^\top \mu^* = -{V^*}^\top \delta\mu = 0$, we get

$$0 = \|\mu^*\|_2^2\, \delta\gamma + \delta U\, ({V^*}^\top \mu^*) + U^*\, (\delta V^\top \mu^*) = \|\mu^*\|_2^2\, \delta\gamma,$$

so $\delta\gamma = 0$ since $\mu^* \neq 0$.

It remains that $\delta U\, {V^*}^\top + U^*\, \delta V^\top = 0$. Let $A := N^{-1}{V^*}^\top \delta V$, which is skew-symmetric since $\delta\theta \in \ker D\psi(\theta^*)$. Right-multiplying by $N^{-1}V^*$ gives $\delta U = -U^* A^\top = U^* A$, and therefore

$$\operatorname{offdiag}(\delta U^\top U^* + {U^*}^\top \delta U) = K \operatorname{offdiag}(DA - AD) = 0 \quad \Longrightarrow \quad (d_i - d_j)A_{ij} = 0 \quad (i \neq j).$$

Since $d_1 > \cdots > d_r$ and A is skew-symmetric, $A = 0$ and thus $\delta U = 0$. Finally, $U^*\, \delta V^\top = 0$ and ${U^*}^\top U^* = KD$ is invertible, so $\delta V = 0$. Hence $\delta\theta = 0$. □

Proposition D.1 is the infinitesimal analog of parameter identifiability at the local chart of θ^*. Now we can state the next proposition, which is crucial to the proof of asymptotic normality.

Proposition D.2 (Graph-induced nondegeneracy of the Fisher information)**.** *Suppose Assumptions A4 and A5 hold. Let*

$$S^* := S(\theta^*) = \gamma^* {\mu^*}^\top + U^* {V^*}^\top.$$

Then the Fisher information matrix

$$I_\Theta(\theta^*) = \sum_{(k,i,j)\in\Omega_w} \omega_{kij}\, \nabla_\theta \eta_{kij}(\theta^*) \nabla_\theta \eta_{kij}(\theta^*)^\top$$

is positive definite on the identified tangent space $T_{\theta^*}\Theta_r$; that is,

$$\delta\theta^\top I_\Theta(\theta^*)\delta\theta > 0 \qquad \textit{for every } 0 \neq \delta\theta \in T_{\theta^*}\Theta_r.$$

Consequently, $I_\Theta(\theta^*)$ is nonsingular in the local identified coordinates.

We begin to prove the second part of Theorem 4. Proof of Proposition D.2 follows afterwards.

Proof of Theorem 4: asymptotic normality. In Section D.1, we have established $\widehat{S} \xrightarrow{\text{p}} S^*$. Asymptotic normality of $\widehat{S}$ can follow this consistency. Since $\mathcal{S}$ is a finite-dimensional Euclidean space and M_n is a smooth criterion (the binomial log-likelihood is twice continuously differentiable in S), standard asymptotic MLE theory applies, i.e., consistency combined with the local positive definiteness of the Hessian (equivalently, the Fisher information in the S-parameterization) implies asymptotic normality. But Proposition D.2 guarantees the nonsingularity of the Fisher information in the S-space, so the local positive definiteness condition is satisfied. This yields $\sqrt{n}(\widehat{S} - S^*) \xrightarrow{\text{d}} \mathcal{N}(0, \mathcal{I}_\mathcal{S}(S^*)^{-1})$.

We now pass this normality from the S-space to the parameter space Θ_r. By the delta method applied to Φ^{-1},

$$\sqrt{n}(\widehat{\theta} - \theta^*) = \sqrt{n} D\Phi^{-1}(S^*)(\widehat{S} - S^*) + o_p(1) \xrightarrow{\text{d}} \mathcal{N}(0, D\Phi^{-1}(S^*)\Sigma_S[D\Phi^{-1}(S^*)]^\top),$$

The limiting covariance in the S-space is $\Sigma_S = \mathcal{I}_\mathcal{S}(S^*)^{-1}$, where

$$\mathcal{I}_\mathcal{S}(S^*) = \sum_{(k,i,j)\in\Omega_w} w_{kij} p^*_{kij}(1 - p^*_{kij}) \nabla_S \eta_{kij}(S^*) \nabla_S \eta_{kij}(S^*)^\top$$

is positive definite by Proposition D.2 (Step 1–2 of its Proof).

The likelihood depends on θ only through the linear predictor $\eta_{kij}(\theta)$. By the chain rule, for each active triple $(k, i, j) \in \Omega_w$,

$$\frac{\partial \eta_{kij}}{\partial \theta} = \frac{\partial \eta_{kij}}{\partial S} \cdot \frac{\partial S}{\partial \theta} = \nabla_S \eta_{kij}(S^*) \cdot D\Phi(\theta^*), \tag{D.7}$$

where $D\Phi(\theta^*)$ is the Jacobian matrix of the score map $\Phi(\theta) = \gamma\mu^\top + UV^\top$ evaluated at θ^*.

The Fisher information in the structured parameter space is therefore

$$\begin{aligned}
\mathcal{I}_\Theta(\theta^*) &= \sum_{(k,i,j)\in\Omega_w} w_{kij} p^*_{kij}(1 - p^*_{kij}) \left(\frac{\partial \eta_{kij}}{\partial \theta}\right) \left(\frac{\partial \eta_{kij}}{\partial \theta}\right)^\top \\
&= \sum_{(k,i,j)\in\Omega_w} w_{kij} p^*_{kij}(1 - p^*_{kij}) D\Phi(\theta^*)^\top \nabla_S \eta_{kij}(S^*) \nabla_S \eta_{kij}(S^*)^\top D\Phi(\theta^*) \\
&= D\Phi(\theta^*)^\top \left[\sum_{(k,i,j)\in\Omega_w} w_{kij} p^*_{kij}(1 - p^*_{kij}) \nabla_S \eta_{kij}(S^*) \nabla_S \eta_{kij}(S^*)^\top \right] D\Phi(\theta^*) \\
&= D\Phi(\theta^*)^\top \, \mathcal{I}_\mathcal{S}(S^*) \, D\Phi(\theta^*).
\end{aligned} \tag{D.8}$$

Proposition D.1 ensures that $D\Phi(\theta^*)$ is injective. Combined with the positive definiteness of $\mathcal{I}_\mathcal{S}(S^*)$, the conjugation by $D\Phi(\theta^*)$ preserves positive definiteness.

Therefore,

$$\sqrt{n}(\widehat{\theta} - \theta^*) \xrightarrow{\mathrm{d}} \mathcal{N}(0, \mathcal{I}_\Theta(\theta^*)^{-1}),$$

and for any smooth scalar target functional $q(\theta)$, the delta method yields

$$\sqrt{n}\{q(\widehat{\theta}) - q(\theta^*)\} \xrightarrow{\mathrm{d}} \mathcal{N}\left(0, \nabla_\theta q(\theta^*)^\top \mathcal{I}(\theta^*)^{-1} \nabla_\theta q(\theta^*)\right),$$

completing the proof. □

Proof of Proposition D.2. The proof is divided into three steps.

Step 1: Score-space Fisher information as a weighted Laplacian form. Temporarily parameterize the model by the row-centered score matrix

$$\mathcal{S} := \{S \in \mathbb{R}^{K\times N} : S1_N = 0\}.$$

For a fixed active triple $(k, i, j) \in \Omega_w$, recall that

$$\eta_{kij}(S) := S_{ki} - S_{kj}.$$

Under the binomial Bradley–Terry–Luce model, the population Fisher information in the ambient score parameterization is

$$I_\mathcal{S}(S^*) = \sum_{(k,i,j)\in\Omega_w} \omega_{kij} \, \nabla_S \eta_{kij}(S^*) \nabla_S \eta_{kij}(S^*)^\top,$$

where $\omega_{kij} = w_{kij} p^*_{kij}(1 - p^*_{kij}) > 0$.

Let $H = (H_{kn}) \in \mathcal{S}$ be an arbitrary score perturbation. Then the corresponding quadratic form is

$$\langle H, I_\mathcal{S}(S^*) H \rangle = \sum_{(k,i,j)\in\Omega_w} \omega_{kij} (H_{ki} - H_{kj})^2. \tag{D.9}$$

Indeed, for each triple (k, i, j), the directional derivative of $\eta_{kij}(S)$ along H is

$$D\eta_{kij}(S^*)[H] = H_{ki} - H_{kj},$$

which gives (D.9).

Now fix a judge $k \in [K]$, and write $h_k = (H_{k1}, \ldots, H_{kN})^\top \in \mathbb{R}^N$. Then the k-th summand in (D.9) becomes

$$\sum_{\{i,j\}\in E^w_k} \omega_{kij} (h_{k,i} - h_{k,j})^2 = h_k^\top L_k h_k,$$

where L_k is the weighted graph Laplacian of the active graph $G_k^w = ([N], E_k^w)$ with edge weights ω_{kij}. Therefore

$$\langle H, I_{\mathcal{S}}(S^*)H\rangle = \sum_{k=1}^{K} h_k^\top L_k h_k. \tag{D.10}$$

Step 2: Judge-wise graph connectivity implies positivity on $\mathcal{S}$. By Assumption A5, each active graph G_k^w is connected. Because L_k is the weighted Laplacian of the connected graph G_k^w with strictly positive edge weights, for every $x \in \mathbb{R}^N$,

$$x^\top L_k x = \sum_{\{i,j\}\in E_k^w} \omega_{kij}(x_i - x_j)^2 \geq 0.$$

Hence $x \in \ker(L_k)$ implies $x^\top L_k x = 0$, so $x_i = x_j$ for every edge $\{i,j\} \in E_k^w$. Since G_k^w is connected, x must be constant on all vertices, that is, $x = c1_N$. Conversely, $L_k 1_N = 0$. Therefore,

$$\ker(L_k) = \operatorname{span}\{1_N\}.$$

Hence, for each k,

$$h_k^\top L_k h_k = 0 \quad \Longleftrightarrow \quad h_k = c_k 1_N \text{ for some } c_k \in \mathbb{R}.$$

But $H \in \mathcal{S}$ means every row of H is centered, i.e.

$$H1_N = 0.$$

Equivalently,

$$1_N^\top h_k = 0 \qquad \text{for all } k \in [K].$$

If $h_k = c_k 1_N$, then $0 = 1_N^\top h_k = N c_k$, so $c_k = 0$, and therefore $h_k = 0$. Since this holds for every k, we conclude that

$$\langle H, I_{\mathcal{S}}(S^*)H\rangle = 0 \quad \Longrightarrow \quad H = 0.$$

Thus $I_{\mathcal{S}}(S^*)$ is positive definite on $\mathcal{S}$:

$$\langle H, I_{\mathcal{S}}(S^*)H\rangle > 0 \qquad \forall\, 0 \neq H \in \mathcal{S}. \tag{D.11}$$

Step 3: Pulling back positivity to the structured parameter space Θ_r. Let $\delta\theta \in T_{\theta^*}\Theta_r$ be an arbitrary tangent perturbation, and define the induced score perturbation

$$H := D\Phi(\theta^*)[\delta\theta] \in \mathcal{S}.$$

Since

$$\Phi(\theta) = \gamma\mu^\top + UV^\top,$$

the first-order variation of $S(\theta)$ at θ^* is

$$H = \delta\gamma\, \mu^{*\top} + \gamma^*\, \delta\mu^\top + \delta U\, V^{*\top} + U^*\, \delta V^\top. \tag{D.12}$$

For each active triple $(k,i,j) \in \Omega_w$, the directional derivative of the structured logit difference satisfies

$$D\eta_{kij}(\theta^*)[\delta\theta] = H_{ki} - H_{kj}.$$

Equivalently,

$$\nabla_\theta \eta_{kij}(\theta^*)^\top \delta\theta = H_{ki} - H_{kj}.$$

Therefore,

$$\begin{aligned}
\delta\theta^\top I_\Theta(\theta^*)\delta\theta &= \sum_{(k,i,j)\in\Omega_w} \omega_{kij}\big(\nabla_\theta\eta_{kij}(\theta^*)^\top\delta\theta\big)^2 \\
&= \sum_{(k,i,j)\in\Omega_w} \omega_{kij}(H_{ki} - H_{kj})^2 \\
&= \langle H, I_{\mathcal{S}}(S^*)H\rangle.
\end{aligned} \tag{D.13}$$

Now suppose

$$\delta\theta^\top I_\Theta(\theta^*)\delta\theta = 0.$$

Then (D.13) and (D.11) imply that $H = 0$. That is,

$$D\Phi(\theta^*)[\delta\theta] = 0.$$

By Proposition D.1, the differential $D\Phi(\theta^*)$ is injective on $T_{\theta^*}\Theta_r$, hence $\delta\theta = 0$.

We have thus shown that

$$\delta\theta^\top I_\Theta(\theta^*)\delta\theta > 0 \qquad \forall\, 0 \neq \delta\theta \in T_{\theta^*}\Theta_r,$$

which proves positive definiteness of $I_\Theta(\theta^*)$ on the identified tangent space. Therefore $I_\Theta(\theta^*)$ is nonsingular in any local identified coordinate system. □

Remark 2 (Interpretation of the two-layer argument)**.** The proposition separates the source of nondegeneracy into two logically distinct parts.

Unlike in ordinary BTL models, graph connectivity alone is not sufficient for Fisher nondegeneracy in the structured latent-factor parameterization, due to parameter-level identifiability from local regularity of the structured map. It must be combined with local injectivity of the decomposition map, which is established in Proposition D.1.

Remark 3 (Relation to the pairwise-comparison literature)**.** The proof above is inspired by recent asymptotic-inference theory for pairwise-comparison models, where the Fisher information is represented as a graph-Laplacian quadratic form and its nondegeneracy is deduced from graph connectivity after quotienting out the additive invariance of pairwise scores [12]. The present argument extends that idea to the heterogeneous judge-aware model by treating the structured parameter space Θ_r as a smooth submanifold of the row-centered score space $\mathcal{S}$, and then pulling back the ambient Fisher information through the differential of the map $\Phi(\theta) = \gamma\mu^\top + UV^\top$.

In other words, the graph structure controls nondegeneracy in the ambient score geometry, whereas the constraint geometry of Θ_r controls whether such nondegeneracy is retained after passing to the latent parameters.

D.3 Auxiliary lemmas

Two lemmas used in the proof of Theorem 4 are established below.

Lemma D.3 (Injectivity of active score differences)**.** *Under Assumption A5, if $S, S' \in \mathcal{S}$ satisfy*

$$S_{ki} - S_{kj} = S'_{ki} - S'_{kj} \qquad \textit{for all } (k,i,j) \in \Omega_w,$$

then $S = S'$.

Proof of Lemma D.3. Fix a judge k and let $h_k := S_{k,:} - S'_{k,:} \in \mathbb{R}^N$. The assumption implies

$$h_{ki} - h_{kj} = 0 \qquad \text{for every } \{i,j\} \in \mathcal{E}_k^w.$$

Because the graph $\mathcal{G}_k^w$ is connected, h_k must be constant across items, say

$$h_k = c_k 1_N.$$

Since both S and S' lie in the centered space $\mathcal{S}$, we also have

$$h_k^\top 1_N = 0.$$

Therefore $Nc_k = 0$, so $c_k = 0$ and hence $h_k = 0$. This holds for every judge k, and thus $S = S'$. □

Lemma D.4 (Coercivity)**.** *Under Assumption A5, the population criterion M in (D.3) is coercive on $\mathcal{S}$, i.e.*

$$\|S\|_F \to \infty, \quad S \in \mathcal{S} \qquad \Longrightarrow \qquad M(S) \to \infty.$$

Moreover, with probability tending to one, the sample criterion M_n in (D.2) is also coercive on $\mathcal{S}$.

Proof of Lemma D.4. We first prove the coercivity of the population criterion.

For each fixed active triple $(k, i, j) \in \Omega_w$, define

$$\phi_{kij}(x) := -p^*_{kij}x + \log(1 + e^x).$$

Because $p^*_{kij} \in (0, 1)$, the function ϕ_{kij} is continuous, nonnegative, and satisfies

$$\phi_{kij}(x) \to \infty \qquad \text{as } |x| \to \infty.$$

Indeed, as $x \to +\infty$,

$$\phi_{kij}(x) = (1 - p^*_{kij})x + o(x) \to \infty,$$

while as $x \to -\infty$,

$$\phi_{kij}(x) = -p^*_{kij}x + o(1) \to \infty.$$

Now fix a judge k and write the kth row of S as

$$s_k := (S_{k1}, \dots, S_{kN})^\top \in \mathbb{R}^N.$$

Since $S \in \mathcal{S}$, we have $s_k^\top 1_N = 0$. Because $\mathcal{G}_k^w$ is connected and N is fixed, we have

$$\|s_k\|_2 \le \sqrt{N}\|s_k\|_1 \le 2\sqrt{N} \sum_{\{i,j\}\in\mathcal{E}_k^w} |s_{k,i} - s_{k,j}|. \tag{D.14}$$

To see this, fix a reference node in $\mathcal{G}_k^w$ and telescope along paths from the reference node to each other node; since the graph is finite and connected, all resulting path lengths are uniformly bounded.

Suppose now that $\|S\|_F \to \infty$ with $S \in \mathcal{S}$. Then for some judge k we must have $\|s_k\|_2 \to \infty$. By (D.14), this implies that for at least one active edge $\{i, j\} \in \mathcal{E}_k^w$,

$$|S_{ki} - S_{kj}| = |s_{k,i} - s_{k,j}| \to \infty.$$

Since the corresponding summand in $M(S)$ is

$$w_{kij}\phi_{kij}(S_{ki} - S_{kj}),$$

with $w_{kij} > 0$, it follows that this term diverges to $+\infty$. All other summands are nonnegative, and therefore

$$M(S) \to \infty.$$

This proves population coercivity.

We next prove sample coercivity. On the event

$$\mathcal{A}_n := \{0 < Y_{kij} < n_{kij} \text{ for all } (k, i, j) \in \Omega_w\},$$

define

$$\phi_{kij,n}(x) := -\bar{Y}_{kij}x + \log(1 + e^x).$$

Whenever $0 < \bar{Y}_{kij} < 1$, the same argument as above shows that

$$\phi_{kij,n}(x) \to \infty \qquad \text{as } |x| \to \infty.$$

Since $p^*_{kij} \in (0, 1)$ and $n_{kij} \to \infty$ for each active triple, we have

$$\mathbb{P}\big(Y_{kij} = 0 \text{ or } Y_{kij} = n_{kij}\big) \to 0,$$

and therefore, by a union bound over the finite set Ω_w,

$$\mathbb{P}(\mathcal{A}_n) \to 1.$$

On $\mathcal{A}_n$, the same graph argument used for the population criterion shows that

$$\|S\|_F \to \infty, \quad S \in \mathcal{S} \qquad \Longrightarrow \qquad M_n(S) \to \infty.$$

Thus M_n is coercive on $\mathcal{S}$ with probability tending to one. □

D.4 Partial sampling scheme

Let $\mathcal{G} := ([N], \mathcal{E})$ denote the pooled graph where $\mathcal{E} := \{\{i, j\} : (k, i, j) \in \Omega_w \text{ for some } k\}$ denotes the edge set. If the graph belonging to a judge, say, $\mathcal{G}_k$ is not connected, then the sampling is considered partial. For example, a judge may compare Item 1, 2, 3 and 4, 5, 6, but the graphs of 1–2–3 and 4–5–6 are separated. Under such a scenario, identification becomes the major obstacle to statistical analysis. The judge may assign higher scores to all 1, 2, and 3, and lower scores to all 4, 5, and 6 without breaking Equation (3) or altering the likelihood. Therefore, assumptions should be made to rule out such ambiguity: the true parameter shall lie in a region where such perturbations must make it impossible for S to still admit the structured representation (4) with unchanged r.

We define the structured score manifold

$$\mathcal{M}_r := \{S = \gamma\mu^\top + UV^\top : (\gamma, \mu, U, V) \in \Theta_r\} \cap \mathcal{S},$$

and let $\mathcal{T}^* := T_{S^*}\mathcal{M}_r$ be its tangent space at S^* to characterize the potential local perturbations on S^*.

Assumption A6 (Identification under partial overlap)**.** Every nonzero local perturbation of the structured score matrix changes at least one observed pairwise log-odds difference. Equivalently, if $H \in \mathcal{T}^*$ satisfies

$$H_{ki} - H_{kj} = 0 \qquad \text{for all } (k, i, j) \in \Omega_w,$$

then we must have $H = 0$.

On the other hand, we assume connectivity on the pooled graph to ensure that there are shared nodes (items) that connect the disconnected parts of each judge, so information is expected to be aligned with the help of these connecting nodes.

Assumption A7 (Connectivity of the pooled graph)**.** The pooled graph $\mathcal{G} = ([N], \mathcal{E})$ with edge set $\mathcal{E} := \{\{i, j\} : (k, i, j) \in \Omega_w \text{ for some } k\}$ is connected.

Remark 4 (How overlap helps)**.** The point of Assumptions A6 and A7 is that overlap lets different judges' local comparisons be aligned through the shared low-rank structure. For example, suppose judge 1 compares items $(1, 2)$ and judge 2 compares items $(2, 3)$, but neither judge compares all three pairs. Item 2 then serves as an overlap point linking the two local pieces of information. Assumption A6 dictates that, after imposing the structured model, such overlaps are rich enough that one cannot change the latent score matrix in a nontrivial way without altering at least one observed comparison.

Assumption A8 (Local compact parameter chart)**.** There exists a compact set $\Theta_{r,0} \subset \Theta_r$ such that θ^* lies in its relative interior after fixing the local anchor signs of the columns of U and V. And the population structured criterion

$$m(\theta) := \sum_{(k,i,j)\in\Omega_w} w_{kij}\Big[- p^*_{kij}\eta_{kij}(\theta) + \log\big(1 + \exp\{\eta_{kij}(\theta)\}\big)\Big],$$

has θ^* as its unique minimizer over $\Theta_{r,0}$.

This assumption replaces the coercivity established in Lemma D.4 under Assumption A5 (full sampling setting). As the identification constraint becomes obscured for the entire space $\mathcal{S}$, this assumption, together with Assumption A8, dictates that even though the likelihood may be flat along component-wise shifts of an unrestricted row, those shifts should either be infeasible within the fixed-rank structured model or break the $\Theta_{r,0}$ boundary that we set.

By the standard asymptotic theory of MLE, the asymptotic properties of $\widehat{\theta}$ of Theorem 4 still hold under Assumptions A4 and A6–A8.

In practice, however, we need to avoid this setting due to the difficulty in verifying Assumptions A6 and A8, unless the sample size is really limited.

E Simulation Studies

This section details Section 6.1 on the specifications and settings.

E.1 Main heterogeneous DGP

The main synthetic simulations use $N = 8$ items, $K = 4$ judges, true heterogeneity rank $r = 1$, and 50 independent repetitions for each condition. The sample-size experiment varies

$$n_{\text{cmp}} \in \{400, 800, 1200, 1600, 2000, 2500, 3000\}$$

at heterogeneity scale $h = 1$. The heterogeneity experiment fixes $n_{\text{cmp}} = 800$ and varies $h \in \{0, 0.5, 1, 2\}$. Error bars in Figure 2 report mean ± 1.96 standard errors across repetitions.

For each repetition, the true parameters are generated as follows. We draw $\widetilde{\mu}_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1)$ and center it as

$$\mu = \widetilde{\mu} - \bar{\widetilde{\mu}} 1_N,$$

so that $1_N^\top \mu = 0$. We draw judge consensus sensitivities as $\gamma = Kp$, where $p \sim \text{Dirichlet}(1_K)$, so that $1_K^\top \gamma = K$. For $r > 0$, raw item and judge factor matrices are drawn with iid standard-normal entries. The item factor is column-centered and projected off the consensus direction,

$$\widetilde{V} \leftarrow \widetilde{V} - 1_N \bar{v}^\top, \qquad \widetilde{V} \leftarrow \widetilde{V} - \mu \frac{\mu^\top \widetilde{V}}{\max(\mu^\top \mu, \varepsilon)}.$$

A QR decomposition is then applied in this constrained column space, yielding an item factor satisfying $V^\top 1_N = 0$ and $\mu^\top V = 0$. The judge factor is similarly column-centered,

$$\widetilde{U} \leftarrow \widetilde{U} - 1_K \bar{u}^\top,$$

and orthogonalized in the zero-sum judge space, so that $1_K^\top U = 0$.

The code assigns decreasing factor strengths $r, r-1, \ldots, 1$ and sets

$$V = Q_V \,\text{diag}\{\sqrt{N}\,\text{strengths}\}, \qquad U = Q_U \,\text{diag}\{\sqrt{K}\,\text{strengths}\}.$$

The item factor is then re-centered, re-projected off μ, and the columns of (U, V) are sign-canonicalized. Finally, both U and V are multiplied by $\sqrt{h}$, so the heterogeneous score component $UV^\top$ is scaled by h. These operations preserve the zero-sum and orthogonality constraints.

By construction, the generated parameters satisfy

$$1_N^\top \mu = 0, \qquad 1_K^\top \gamma = K, \qquad 1_K^\top U = 0, \qquad V^\top 1_N = 0, \qquad \mu^\top V = 0.$$

After QR normalization and canonical sign anchoring, the factors also satisfy the anchoring requirements in Condition 1. The true score matrix is then

$$S = \gamma \mu^\top + UV^\top,$$

and pairwise outcomes are generated according to the BTL model in (2). We also use an auxiliary DGP from JA-Ranking, which uses the same μ, γ draw but sets $U = V = 0$, so $S_{ki} = \gamma_k \mu_i$. The experiment specifications and results are shown in Section E.4.

E.2 Comparison design, fitting protocol, and metrics

For the design of experiment, comparisons are allocated using a near-balanced design over all judge-item-pair cells (k, i, j) with $i < j$. Let $C = KN(N-1)/2$. Each cell receives either $\lfloor n_{\text{cmp}}/C \rfloor$ or $\lceil n_{\text{cmp}}/C \rceil$ comparisons, with the remaining cells selected uniformly without replacement. For each comparison in cell (k, i, j), we draw

$$Y_{kij}^{(\ell)} \sim \text{Bernoulli}\{\sigma(S_{ki} - S_{kj})\},$$

where $Y_{kij}^{(\ell)} = 1$ means that judge k prefers item i to item j. The generated records are aggregated into upper-triangular comparison counts n_{kij} and win counts Y_{kij}.

All methods are fit to the same aggregated counts. For pooled BTL, we aggregate comparisons over judges and fit a single centered score vector $\widehat{\mu}_{\mathrm{BTL}}$, then form

$$\widehat{S}_{\mathrm{BTL}} = 1_K \widehat{\mu}_{\mathrm{BTL}}^\top.$$

For JA-Ranking, we fit the sensitivity-only model

$$\widehat{S}_{\mathrm{JA}} = \widehat{\gamma}\widehat{\mu}^\top,$$

using the same centering and normalization convention as in the main text. For HJA, we fit the rank-r constrained model

$$\widehat{S}_{\mathrm{HJA}} = \widehat{\gamma}\widehat{\mu}^\top + \widehat{U}\widehat{V}^\top.$$

All score-level metrics below are computed from these fitted score matrices.

For each method, score-entry intervals are obtained by applying the corresponding asymptotic covariance estimate to the fitted score-entry functional $q(\theta) = S_{ki}$. For HJA, this uses the plug-in Fisher-information estimator from Theorem 4; for JA-Ranking and pooled BTL, we use the corresponding model-specific delta-method covariance estimates under their restricted parameterizations. For misspecified baselines, coverage is evaluated against the true heterogeneous score matrix S, so undercoverage reflects both statistical uncertainty and structural bias.

We report four groups of metrics:

(1) **Score-matrix error.** Score recovery is measured by the mean squared error

$$\mathrm{MSE} = \frac{1}{KN}\sum_{k=1}^{K}\sum_{i=1}^{N}(\widehat{S}_{ki} - S_{ki})^2.$$

(2) **Consensus ranking recovery.** Let $r_i^\star$ and $\widehat{r}_i$ denote the ranks of item i induced by the true consensus score μ and the estimated consensus score $\widehat{\mu}$, respectively, with rank 1 assigned to the largest score. We report the Spearman rank correlation

$$\rho_{\mathrm{sp}}(\mu, \widehat{\mu}) = \mathrm{corr}\big((r_1^\star, \ldots, r_N^\star), (\widehat{r}_1, \ldots, \widehat{r}_N)\big).$$

When there are no ties, this is equivalently

$$\rho_{\mathrm{sp}}(\mu, \widehat{\mu}) = 1 - \frac{6\sum_{i=1}^{N}(r_i^\star - \widehat{r}_i)^2}{N(N^2-1)}.$$

We also report NDCG@N. Let $\widehat{\pi}$ be the ordering of items sorted by decreasing $\widehat{\mu}_i$, and define the true relevance score

$$\mathrm{rel}_i = 2^{N - r_i^\star} - 1.$$

The discounted cumulative gain of the estimated ranking is

$$\mathrm{DCG}\,@N = \sum_{\ell=1}^{N}\frac{\mathrm{rel}_{\widehat{\pi}(\ell)}}{\log_2(\ell+1)}.$$

Let $\pi^\star$ be the ordering of items sorted by decreasing μ_i. The ideal discounted cumulative gain is

$$\mathrm{IDCG}\,@N = \sum_{\ell=1}^{N}\frac{\mathrm{rel}_{\pi^\star(\ell)}}{\log_2(\ell+1)}.$$

We define

$$\mathrm{NDCG}\,@N = \frac{\mathrm{DCG}\,@N}{\mathrm{IDCG}\,@N}.$$

(3) **Uncertainty calibration.** Let $C_{ki}^{0.95}$ denote the nominal 95% confidence interval for the score entry S_{ki}. Coverage is measured by

$$\mathrm{Coverage} = \frac{1}{KN}\sum_{k=1}^{K}\sum_{i=1}^{N}\mathbb{1}\{S_{ki} \in C_{ki}^{0.95}\}.$$

Note that "Unstructured BTL + SVD" is excluded from this metric as truncated SVD is a non-smooth mapping prohibiting ordinary uncertainty quantification.

(4) **Judge-specific sign accuracy.** Judge-specific pairwise behavior is measured by the fraction of item-pair signs correctly recovered:

$$\mathrm{SignAcc} = \frac{1}{K\binom{N}{2}}\sum_{k=1}^{K}\sum_{1\le i<j\le N}\mathbb{1}\left\{\mathrm{sign}(S_{ki} - S_{kj}) = \mathrm{sign}(\widehat{S}_{ki} - \widehat{S}_{kj})\right\}.$$

E.3 Convergence analysis

A run is declared converged when the relative likelihood decrease falls below the tolerance used in Algorithm 1. Figure E2 offers a typical example of the convergence of Algorithm 1. Since each iteration step calculates a minimum, its convergence rate is fast.

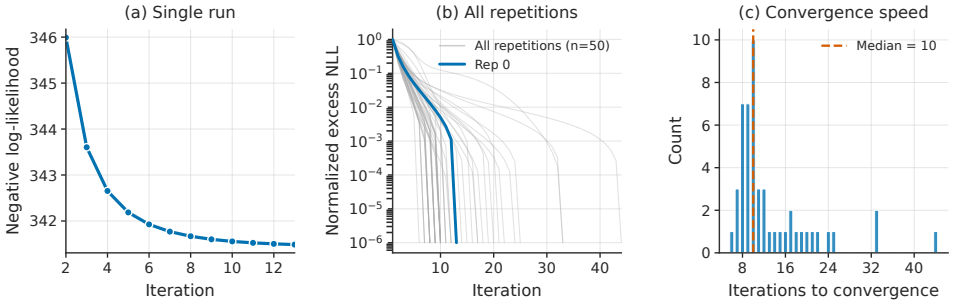


Figure E2: Convergence behavior of our proposed model. (a) Negative log-likelihood (NLL) decreases monotonically over iterations for a representative run. (b) Normalized excess NLL (log-scale) across $n = 50$ repetitions shows consistent geometric convergence regardless of random seed. (c) Histogram of convergence iterations demonstrates that the algorithm consistently converges in a small, predictable number of steps.

E.4 Auxiliary sensitivity-only DGP

We also run an auxiliary simulation under a sensitivity-only data-generating process matching the structure of JA-Ranking. The true score matrix is

$$S_{ki} = \gamma_k \mu_i,$$

with no low-rank heterogeneous component. We vary

$$n_{\text{cmp}} \in \{400, 800, 1200, 1600, 2000, 2500, 3000\},$$

and otherwise use the same comparison allocation, aggregation, and evaluation metrics as in Section E.2. HJA is still fit with rank $r = 1$, so this experiment evaluates robustness to mild over-parameterization when the true heterogeneity rank is zero.

Figure E3 reports the results. The proposed method remains stable despite fitting one redundant heterogeneous direction. In some finite-sample settings, HJA can match or slightly improve on the sensitivity-only fit. We interpret this as a numerical finite-sample effect rather than a structural advantage: when some generated γ_k values are close to zero, the log-parameterization used by the sensitivity-only baseline can become unstable, whereas the affine parameterization used by HJA does not require strict positivity of γ_k.

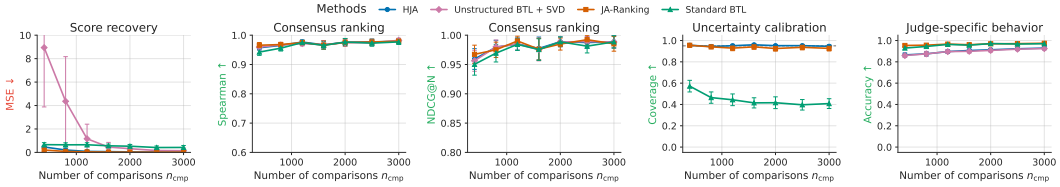


Figure E3: Synthetic performance under JA-Ranking data-generating process [1], where the true score matrix has $S_{ki} = \gamma_k \mu_i$ and no low-rank interaction term.

E.5 Ablation Study

We introduce a variant that performs the same alternating likelihood updates but omits ReAnchor as the ablated model. We also add an unstructured estimation scheme where $\bar{S}$ (subject to (3)) is first estimated from the data and $\widehat{S}$ and $\widehat{\theta}$ are recovered via Proposition 2 and truncated SVD on $\widetilde{S}$ defined by (8). We compare the score recovery, consensus ranking, and sign prediction accuracy of

our full model, the direct-S truncated-SVD model, and the ablated model where the ReAnchor Step (Algorithm C.2) is removed from Algorithm 1. The settings are the same as in the main simulation experiment, although the plot of confidence interval coverage accuracy is removed due to the non-identifiability of the ablated model.

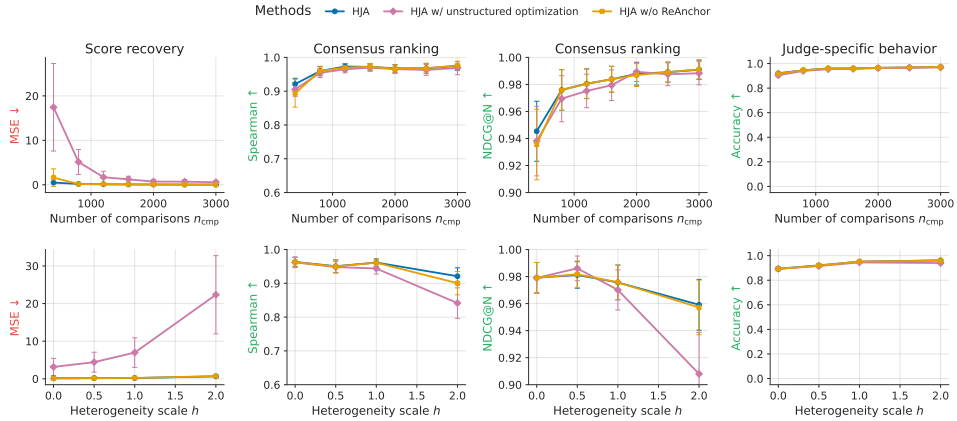


Figure E4: ReAnchor ablation under the synthetic DGP. We compare the proposed HJA estimator, the direct-S truncated-SVD estimator, and an ablated HJA variant that omits the ReAnchor step after each alternating update. Columns report score MSE, Spearman correlation, NDCG@N, and judge-specific sign accuracy; rows vary sample size and heterogeneity strength.

Figure E4 shows that the performance of the ablated models are worse than that of the full model. This shows that ReAnchor and structure are important for stabilizing the identified low-rank parameterization and improving score recovery, while the non-reanchored variant can still preserve much of the ordinal information needed for ranking and judge-specific sign prediction. In summary, the design of Algorithm 1 stabilizes score recovery while preserving the strong ranking and judge-behavior performance of the proposed estimator.

E.6 Misspecified synthetic data on latent expertise in Section 6.1(B)

We provide a hidden-domain experiment with $N = 12$ items and $K = 9$ judges to illustrate how our model can identify the potential latent areas of judges' expertise, when the synthetic data do not agree with our model.

Specifically, each comparison is a coding or maths task, with frequencies $\pi = (\pi_{\text{coding}}, \pi_{\text{math}}) = (0.65, 0.35)$. Following (7) with its $U = 0$ and d unobserved, set $\gamma_{kd} = 12$ when judge k is good at domain d and 0.005 when k is bad at d, up to small random perturbations. Hence, a judge is sensitive to a model's ability only within its capable latent domain.

Each item i has a fixed latent skill vector $a_i = (a_{i,\text{code}}, a_{i,\text{math}})$, and the pair of interest has $a_A = (1.5, 0.2)$ and $a_B = (0.2, 1.0)$, so A leads by 1.3 on coding and B leads by 0.8 on maths. Conditional on judge k, pair (i, j), and domain d, the outcome is

$$Y_{kij} \mid D = d \sim \text{Bernoulli}\left[\sigma(\gamma_{kd}(a_{id} - a_{jd}))\right].$$

We define the equal-judge, domain-weighted latent score before generating outcomes as

$$\bar{S}_i = \frac{1}{K} \sum_{k=1}^{K} \sum_{d} \pi_d \gamma_{kd} a_{id},$$

and the contribution of judge k to A versus B is $\Delta_k = \sum_d \pi_d \gamma_{kd}(a_{Ad} - a_{Bd})$. A maths-reliable judge contributes $0.65(0.005)(1.3) + 0.35(12)(-0.8) = -3.36$, while a coding-reliable judge contributes $0.65(12)(1.3) + 0.35(0.005)(-0.8) = 10.14$, hence

$$\bar{S}_A - \bar{S}_B = (6(-3.36) + 3(10.14))/9 = 1.14 > 0,$$

thus $A > B$, which an unbiased latent panel should report. However, since the domain info D is hidden, the data identify only the domain-marginal win-probability by mixing out latent domains,

i.e.,

$$p_{kij} = \sum_d \pi_d \sigma(\gamma_{kd}(a_{id} - a_{jd})) = 0.49 < 0.5. \tag{E.1}$$

So pooled binary outcomes favor $B > A$ due to info loss. The mechanism is also misspecified relative to our model because the *observed* probability p_{kij} is a *mixture model that cannot be expressed via our likelihood*. Thus, it's beyond HJA to identify or "unmix" it either, but, as shown below, $UV^\top$ can well *approximate* such mixed hidden interaction.

When models are fit, they only receive judge identities, item pairs, and comparison outcomes, and no task-domain label is supplied. Every judge contributes 4,800 comparisons. Across 50 seeds, HJA recovers $A > B$ in 48 runs, against 35/0 for the sensitivity-only/pooled-BTL model. Held-out accuracy is 0.714/0.551/0.542 for HJA/pooled-BTL/sensitivity-only, showcasing our model's performance under a more general noise paradigm. Five-fold validation selects $r = 1$ in 49 runs, agreeing with the latent contrast of coding vs maths. This experiment shows the versatility of our model when some important labels are hidden in the data.

F Real Data

This section provides three layers of real-data analysis: (i) experimental specifications and robustness benchmarks comparing HJA to baselines (Section F.1); (ii) proper interpretations of patterns of model outputs (Section F.2).

F.1 Experiment specifications for Real-data robustness evaluation of Section 6.1(C)

The overall aim of real data comparison is to evaluate whether the proposed heterogeneous judge-aware model improves both prediction and ranking sensitivity relative to JA-Ranking and Standard BTL. The experiments are organized around held-out prediction, robustness to injected noisy judges, and accuracy on near-tie item pairs.

Dataset description. We evaluate on four pairwise LLM-comparison datasets downloaded from `https://github.com/TanXZfra/A-Judge-Aware-Ranking-Framework-for-Evaluating-Large-Language-Models-without-Ground-Truth/tree/main/data`. Recall that N denotes the number of candidate models, K the number of judge models, and n_{cmp} the number of usable pairwise comparisons after preprocessing. Each comparison is represented as (k, i, j, y), where k indexes the judge, i, j index the compared candidate models, and $y \in \{0, 0.5, 1\}$ records the judge preference, with ties mapped to 0.5. We drop only records with unknown labels. For each judge-item-pair cell, n_{kij} denotes the number of observed comparisons for judge k on candidate pair (i, j); the summaries below report the average, minimum, and maximum over nonzero n_{kij} cells.

The Chatbot Arena dataset is derived from the LMSYS Chatbot Arena pairwise evaluation setting under the Apache 2.0 license, a crowdsourced battle platform for comparing chat assistants [13, 14]. The original file contains 10,000 pairwise records. After removing 63 unknown-label records, we use $n_{\text{cmp}} = 9{,}937$ comparisons, including 1,224 ties; the tie-excluded count is 8,713. The dataset has $N = 20$ candidate models and $K = 10$ judge models. Among the possible judge-item-pair cells, 1,737 of 1,900 are observed, and nonzero n_{kij} has average/minimum/maximum $5.72/1/29$.

The MT-Bench dataset follows the LMSYS MT-Bench setting (Apache 2.0 license), a controlled multi-turn benchmark with 80 questions. The original file contains 10,000 pairwise records comparing $N = 6$ candidate models, evaluated by $K = 20$ judge models [14]. After removing 294 unknown-label records, we use $n_{\text{cmp}} = 9{,}706$ comparisons, including 756 ties; the tie-excluded count is 8,950. All 300 possible judge-item-pair cells are observed, with nonzero n_{kij} average/minimum/maximum $32.35/10/48$.

The UltraFeedback dataset is derived from UltraFeedback-style instruction-response comparisons (under MIT license), based on a large-scale AI-feedback resource built from diverse prompts and model responses [15]. The original file contains 10,000 pairwise records. After removing 274 unknown-label records, we use $n_{\text{cmp}} = 9{,}726$ comparisons, including 761 ties; the tie-excluded count is 8,965. The dataset contains $N = 17$ candidate models and $K = 20$ judge models. We observe 2,379 of 2,720 possible judge-item-pair cells, and nonzero n_{kij} has average/minimum/maximum $4.09/1/15$.

The in-house dataset is a private pairwise comparison set over recent LLMs collected by [1]. The original file contains 36,000 pairwise records. After removing 161 unknown-label records, we use $n_{\text{cmp}} = 35{,}839$ comparisons, including 1,782 ties; the tie-excluded count is 34,057. This is the largest dataset in our experiments, with $N = 45$ candidate models and $K = 18$ judge models. We observe 15,374 of 17,820 possible judge-item-pair cells, with nonzero n_{kij} average/minimum/maximum $2.33/1/9$.

Experiment details. Stability and near-tie experiments use a record-level 80/20 train/test split. Our proposed model selects rank by 5-fold validation on negative log likelihood on the training split or full sample, depending on the workflow. Candidate ranks are $0, \ldots, \min(K-1, N-2)$. For the stability experiment, we fit each method on the training records and report held-out pairwise accuracy on the test records. For the robustness experiment, we fit base rankings on the original data, then add noisy judges sequentially. The saved default noisy-judge grid is 1 through 10 added noisy judges; the main table reports steps 1, 5, and 10 as low, mid, and high perturbation levels, with the numbers indicating exact-ranking accuracy versus the base ranking after noisy judges. For the near-tie slice, we select up to 20 item pairs from the training split with at least 20 training records per pair, sorted by empirical win-rate distance to 0.5 and ordered into low/mid/high tertiles, and evaluate

the selected near-tie test records separately from all other test records after skipping tie-labeled test records.

For the **stress test** in Section 6.1(D), we define the favored cluster as the bottom $\lfloor N/3 \rfloor$ candidates under the clean HJA consensus. One injected source copies the candidate-pair and question coverage of all raw records: 10,000 for each public dataset and 36,000 for in-house, approximately matching the entire clean panel's comparison volume. It deterministically selects the cluster candidate when exactly one belongs to the cluster; otherwise, it selects either candidate with equal probability, without ties. Across 20 label seeds (103–122), we vary only these random choices, holding the clean fits, cluster, injection volume, and clean-selected rank ($r = 1$) fixed. We compute the Spearman correlation between each contaminated consensus and that method's clean consensus, reporting the mean and sample standard deviation across repetitions. This variability reflects injected-label randomness conditional on the clean panel. HJA uses $\tau = 40$, tolerance 5×10^{-5}, at most 300 alternating steps, and 2,000 inner iterations. Sensitivity-only uses a fixed 100-iteration budget, while pooled BTL allows 800 iterations. All runs are retained: HJA and pooled BTL meet their convergence criteria throughout, whereas sensitivity-only reaches its iteration limit without meeting its convergence criterion. Its results therefore describe performance under this computational budget.

All experiments were run on a CPU server with 12 cores, the type being Intel(R) Xeon(R) Gold 6126 CPU. One repetition of a whole Robustness experiment takes 45 minutes, while other experiments take minutes for one repetition. The optimization parameter $\tau = 30$ is used by default, while it is finetuned to be 5 and 300 for the Robustness experiment on UltraFeedback (excluding low noisy injection) and the Near-tie experiment in in-house data. For each convex optimization task, i.e., Line 4–5 in Algorithm 1, we use L-BFGS-B method [16] after reparameterization based on Condition 1 (i)-(ii).

F.2 Interpretation

Outlier judge identification. We first show the composite analysis for MT-Bench and UltraFeedback in Figure F5. In both experiments, `Llama-4-Scout-17B-16E-In` appears noisy and not sensitive to the consensus of other judges.

However, in order to interpret fitted judge behavior, we need to distinguish consensus sensitivity from residual disagreement. A large value of $\widehat{\gamma}_k$ indicates that judge k responds strongly to the consensus ranking direction $\widehat{\mu}$, but it does not by itself imply that the judge is globally reliable. We therefore also examine the residual leverage

$$\widehat{h}_k = \|(\widehat{U}\widehat{V}^\top)_{k,:}\|_2,$$

or its relative version

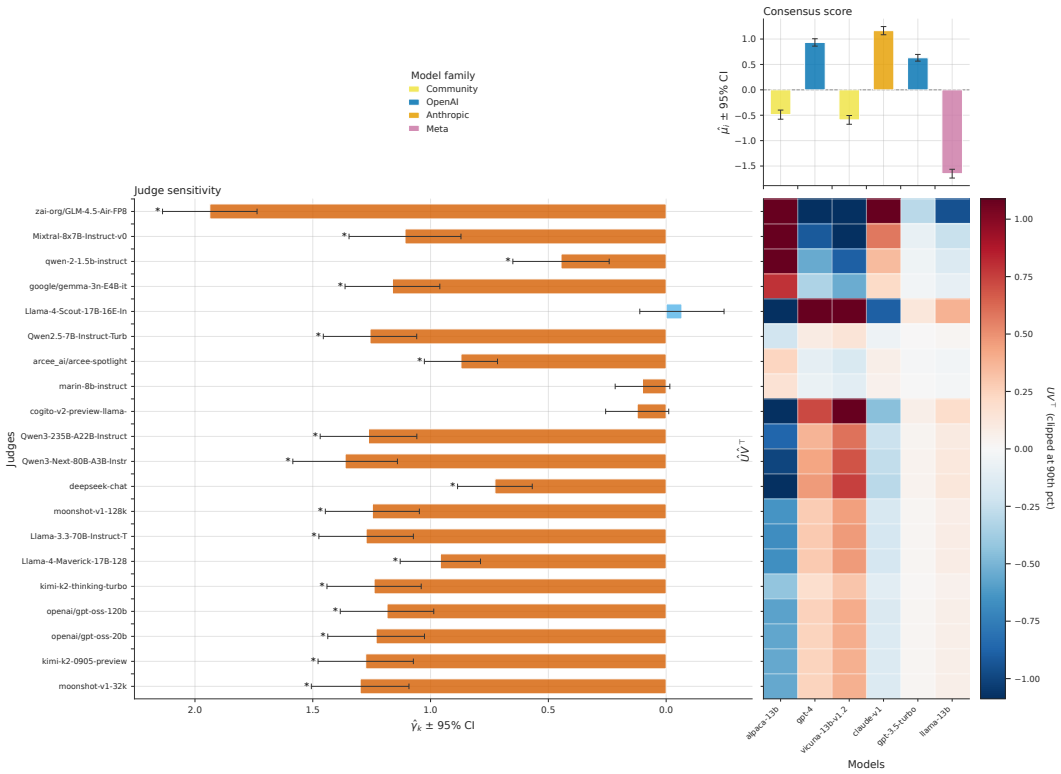
$$\widehat{\rho}_k = \frac{\|(\widehat{U}\widehat{V}^\top)_{k,:}\|_2}{|\widehat{\gamma}_k|\|\widehat{\mu}\|_2 + \varepsilon}.$$

Judges with large $\widehat{\gamma}_k$ and small $\widehat{h}_k$ are strongly aligned with the consensus signal, whereas judges with both large $\widehat{\gamma}_k$ and large $\widehat{h}_k$ are high-leverage judges whose evaluations combine sharp consensus discrimination with substantial off-consensus structure. In the rank-one fits used here, the sign of $\widehat{U}_{k1}$ further indicates whether this residual disagreement is aligned with or opposed to the dominant judge-side pattern.

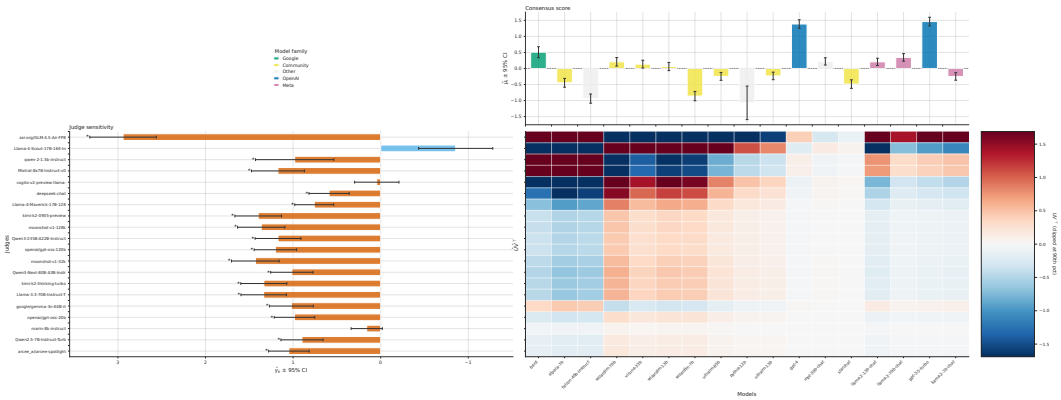
Figure F6 plots $\widehat{h}_k$ against $\widehat{\gamma}_k$. Because only the outliers matter under this metric, we only mark the outlier judge.

Under this diagnostic, `zai-org/GLM-4.5-Air-FP8` is not best interpreted as simply a highly reliable judge. Although it has the largest estimated consensus sensitivity, it also has unusually large residual leverage, with a residual direction opposed to the main judge cluster. Thus, GLM-4.5 behaves as a high-sensitivity but high-bias judge: it sharply separates models along the consensus direction while simultaneously imposing a strong idiosyncratic preference component.

In light of this, we rerun our model on MT-Bench and UltraFeedback without GLM 4.5 and visualize the results in Figure F7. From these results, we find that `Llama-4-Scout-17B-16E-In` is no longer a pure noise, but rather a judge with moderate sensitivity. This shows that the extreme leverage of GLM 4.5 distorts the fitted judge behavior of other judges, and its removal allows a more nuanced interpretation of the remaining judges.



(a) MT-Bench



(b) UltraFeedback

Figure F5: The composite figures showing the judge sensitivity, model rankings, and judge-model interaction by our proposed model. Error bars are given with near-normality assumption.

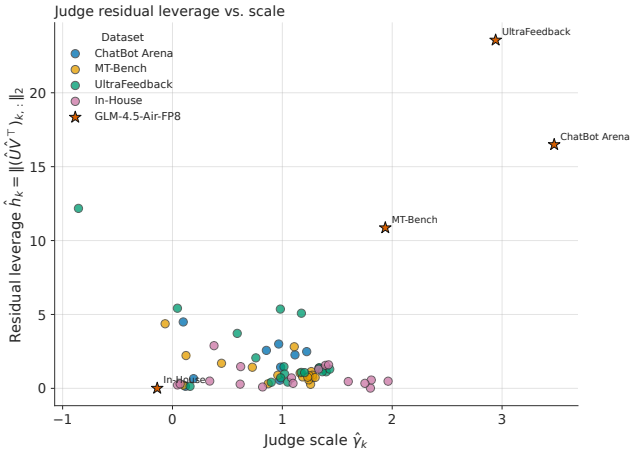


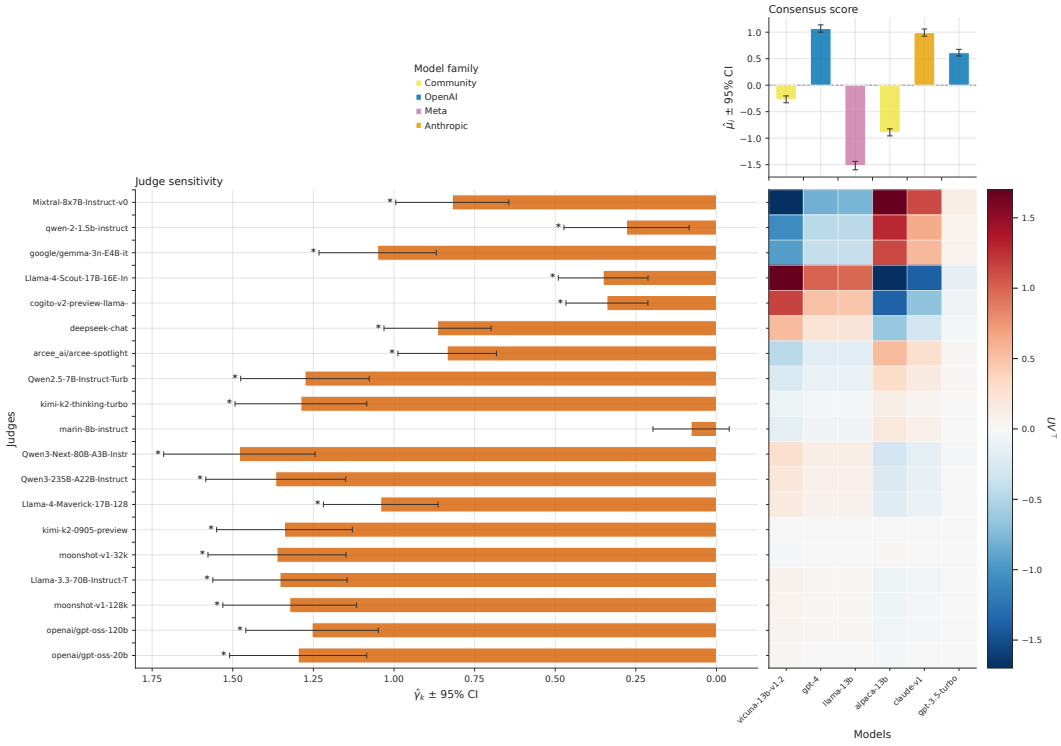
Figure F6: The scatter plot of leverages, where the outlier model is explicitly annotated.

Self-evaluation bias. A well-documented phenomenon in automated evaluation is self-evaluation bias (or self-preference), where an LLM judge systematically assigns higher scores to responses generated by itself or models from the same provider family. While this bias poses a significant threat to naive aggregation methods when a judging panel is monopolized by a single model family, its impact is fundamentally mitigated in a diverse, multi-judge setting. However, apart from selecting a dataset containing diversified judges and models, our formulation can also absorb a judge's systematic self-preference into the orthogonal residual disagreement term $UV^\top$. The heatmap of $UV^\top$ should show bright "hotspots" exactly where a judge prefers the output of its own or its family members. This mechanism ensures that μ retains a high-fidelity representation of consensus quality without the need to completely discard a biased judge.

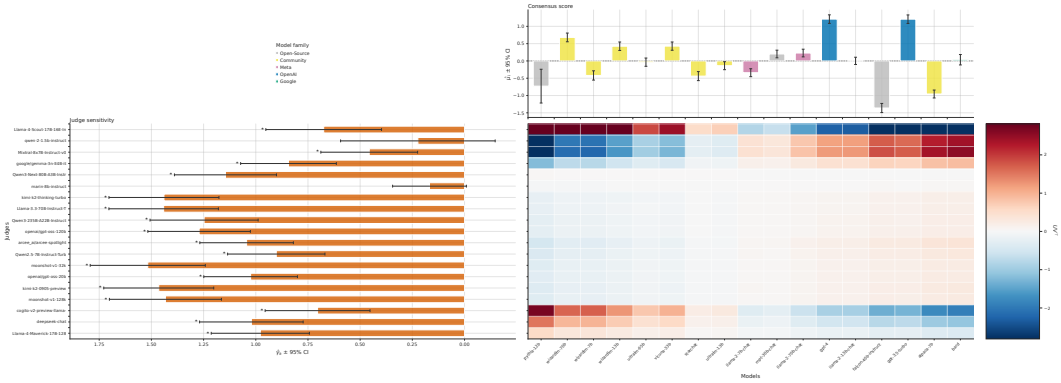
As an example, we plot the heatmap of $UV^\top$ to show the heterogeneous interaction between judges and the items on ChatBot Arena after the removal; see Figure F8, which illustrates the interaction between the judges and the models.

Figure F8 shows that although different judges indeed prefer the styles from different models, there is no observed self-preferencing bias, possibly due to the double-blindness in the design and collection of these data.

We also try a biclustering analysis [17] on the estimated $\widehat{S}$ of in-house data (with GLM 4.5) removed to mine some patterns; see Figure F9. We analyze in-house data because it contains the most diversified models and judges and is hence the most suitable for biclustering analysis. We also include the biclustering results for the other three datasets in Figure F10. The block structures validates HJA's decomposition: judges from the same family converge on similar preference profiles, suggesting that systematic disagreement is not random noise but reflects coherent, interpretable differences in evaluation philosophy. This makes the residual component $UV^\top$ a meaningful diagnostic tool rather than an unexplained variance sink.



(a) MT-Bench



(b) UltraFeedback

Figure F7: The composite figures showing the judge sensitivity, model rankings, and judge-model interaction by our proposed model with GLM 4.5 removed. Error bars are given with near-normality assumption.

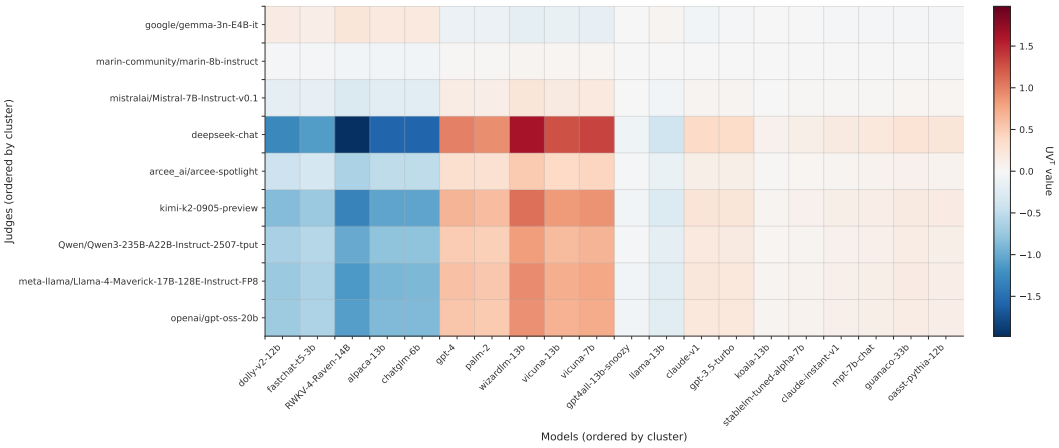


Figure F8: Judge-model heterogeneous interaction terms $UV^\top$ for ChatBot Arena clustered by judge-model affinity (extreme judge removed). The first three judges on the y-axis (`gemma-3n-E4B-it`, `marin-8b-instruct`, and `Mistral-7B-Instruct-v0.1`) show comparatively weak heterogeneous preferences, indicating closer alignment with the consensus ranking. Most disagreement happens mainly on the first ten models shown on the x-axis, with the majority of judges preferring model 6–10 than 1–5 on the x-axis.

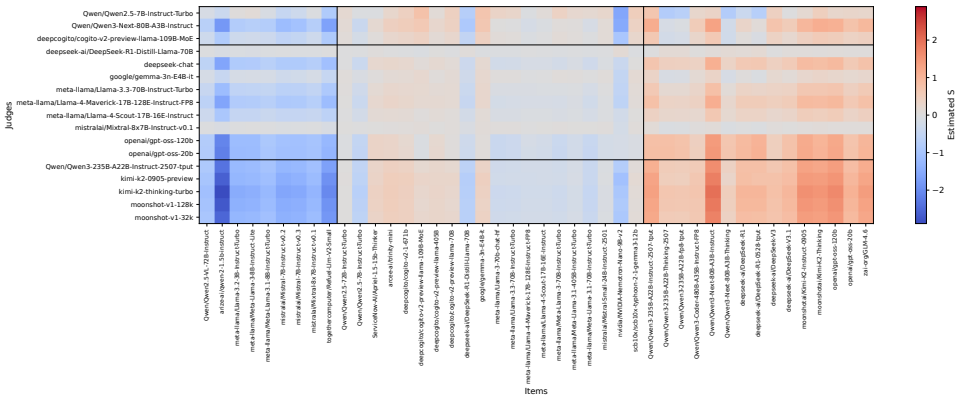
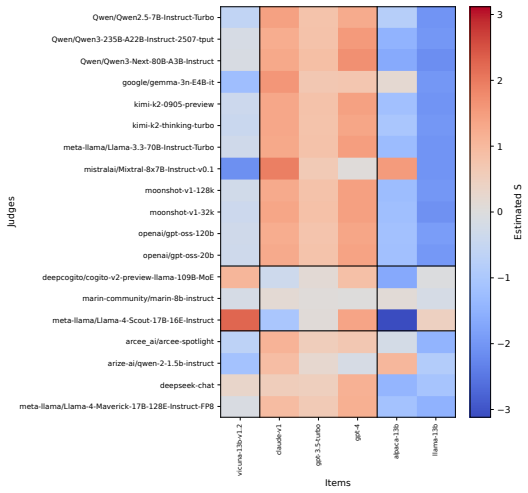
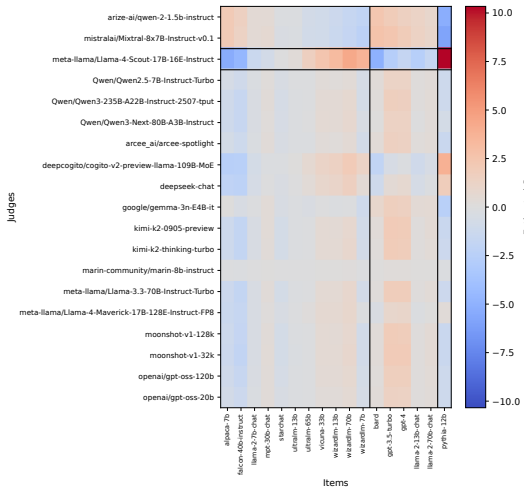


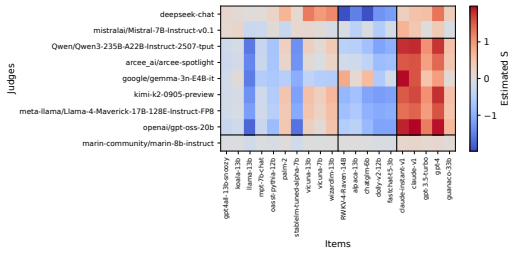
Figure F9: Biclustering of the estimated judge-specific score matrix $\widehat{S}$ on the in-house dataset after removing the high-leverage GLM-4.5 judge. Rows and columns are reordered by biclustering, so contiguous blocks indicate judges and models with similar fitted preference profiles. On the item side, the clustering separates earlier/smaller Qwen–Llama–Mistral-style models from a middle group of DeepCogito, DeepSeek-distilled, Gemma, Llama, and Mistral/Nemotron models, and a rightmost block dominated by recent Qwen3, DeepSeek, Kimi/Moonshot, OpenAI `gpt-oss`, and GLM models. On the judge side, Qwen-family judges show similar scoring behavior, Kimi and Moonshot judges form a closely related cluster, and DeepSeek, Gemma, Llama, Mistral, and OpenAI judges occupy a more moderate shared cluster. The block structure indicates that HJA recovers family-level affinity patterns in judge-specific rankings, rather than only isolated judge or model effects.



(a) MT-Bench



(b) UltraFeedback



(c) Chatbot Arena

Figure F10: Biclustering of estimated $\widehat{S}$ across (a) MT-Bench, (b) UltraFeedback, and (c) Chatbot Arena, all with GLM-4.5 removed. From (c), we find that Claude and GPT are preferred by most judges, and indeed, they have a high consensus score. Such a concordance pattern also appears in (a).

References

[1] Mingyuan Xu, Xinzi Tan, Jiawei Wu, and Doudou Zhou. A judge-aware ranking framework for evaluating large language models without ground truth. In *Forty-third International Conference on Machine Learning*, 2026.

[2] Arun Rajkumar and Shivani Agarwal. When can we rank well from comparisons of $O(n \log(n))$ non-actively chosen pairs? In *29th Annual Conference on Learning Theory*, volume 49 of *Proceedings of Machine Learning Research*, pages 1376–1401. PMLR, 2016.

[3] Tao Jin, Pan Xu, Quanquan Gu, and Farzad Farnoud. Rank aggregation via heterogeneous Thurstone preference models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(04):4353–4360, 2020.

[4] Rui Wu, Jiaming Xu, Rayadurgam Srikant, Laurent Massoulié, Marc Lelarge, and Bruce Hajek. Clustering and inference from pairwise comparisons. In *Proceedings of the 2015 ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Systems*, pages 449–450, 2015.

[5] Fan R. K. Chung. *Spectral graph theory*, volume 92 of *CBMS Regional Conference Series in Mathematics*. American Mathematical Society, 1997.

[6] Paul Erdős. Graph theory and probability. *Canadian Journal of Mathematics*, 11:34–38, 1959.

[7] Steven G. Self and Kung-Yee Liang. Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *Journal of the American Statistical Association*, 82(398):605–610, 1987.

[8] Jean-Marc Robin and Richard J. Smith. Tests of rank. *Econometric Theory*, 16(2):151–175, 2000.

[9] Hamed Karimi, Julie Nutini, and Mark Schmidt. Linear convergence of gradient and proximal-gradient methods under the Polyak–Łojasiewicz condition. In *Joint European conference on machine learning and knowledge discovery in databases*, pages 795–811. Springer, 2016.

[10] Boris Teodorovich Polyak. Gradient methods for minimizing functionals. *Zhurnal vychislitel'noi matematiki i matematicheskoi fiziki*, 3(4):643–653, 1963.

[11] Hédy Attouch, Jérôme Bolte, Patrick Redont, and Antoine Soubeyran. Proximal alternating minimization and projection methods for nonconvex problems: An approach based on the Kurdyka–Łojasiewicz inequality. *Mathematics of Operations Research*, 35(2):438–457, 2010.

[12] Ruijian Han, Wenlu Tang, and Yiming Xu. Statistical inference for pairwise comparison models. *arXiv preprint arXiv:2401.08463v3*, 2025.

[13] Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Banghua Zhu, Hao Zhang, Michael Jordan, Joseph E. Gonzalez, and Ion Stoica. Chatbot Arena: An open platform for evaluating LLMs by human preference. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 8359–8388. PMLR, 2024.

[14] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems*, volume 36, pages 46595–46623, 2023.

[15] Ganqu Cui, Lifan Yuan, Ning Ding, Guanming Yao, Bingxiang He, Wei Zhu, Yuan Ni, Guotong Xie, Ruobing Xie, Yankai Lin, Zhiyuan Liu, and Maosong Sun. ULTRAFEEDBACK: Boosting language models with scaled AI feedback. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 9722–9744. PMLR, 21–27 Jul 2024.

[16] Richard H Byrd, Peihuang Lu, Jorge Nocedal, and Ciyou Zhu. A limited memory algorithm for bound constrained optimization. *SIAM Journal on Scientific Computing*, 16(5):1190–1208, 1995.

[17] John A Hartigan. Direct clustering of a data matrix. *Journal of the American Statistical Association*, 67(337):123–129, 1972.

NeurIPS Paper Checklist

1. **Claims**
   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?
   Answer: [Yes]
   Justification: The abstract and Introduction is carefully aligned with the content of our work.
   Guidelines:
   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.
2. **Limitations**
   Question: Does the paper discuss the limitations of the work performed by the authors?
   Answer: [Yes]
   Justification: In Section 7, limitations are discussed.
   Guidelines:
   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.
3. **Theory assumptions and proofs**
   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?
   Answer: [Yes]
   Justification: All assumptions are relegated to the Appendix but are briefly introduced in the main text. All the proofs are checked and appear in the Appendix to save space. All theorems stated have referenced all of the respective assumptions needed for proof.
   Guidelines:
   - The answer [N/A] means that the paper does not include theoretical results.

- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**
   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?
   Answer: [Yes]
   Justification: Experiment setups and specifications are detailed in Sections E and F. Coding files can be downloaded from the link on the first page of Appendix.
   Guidelines:
   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**
   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?
   Answer: [Yes]
   Justification: We gave a GitHub URL where we download the datasets. Credits are also appropriately given.
   Guidelines:
   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.

- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**
   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?
   Answer: [Yes]
   Justification: The settings and details are elaborated along with the specifications (see reproducibility part).
   Guidelines:
   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.
7. **Experiment statistical significance**
   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?
   Answer: [Yes]
   Justification: All plots have error bars, and numbers in tables are accompanied by quantified uncertainty. We use "95%" language when the sample size is large and these data are obtained via random repititions, or when we use the asymptotic normality derived for the scenario. Otherwise, we state the original definition of such intervals.
   Guidelines:
   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
   - If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**
   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?
   Answer: [Yes]
   Justification: In Section F. But the time of execution varies and is only approximate because most fits take less than a minute.
   Guidelines:
   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).
9. **Code of ethics**
   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?
   Answer: [Yes]
   Justification: Checked and we conform with the Code.
   Guidelines:
   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).
10. **Broader impacts**
    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?
    Answer: [Yes]
    Justification: Broader impacts are briefly included in Section 7 as consequences of the limitations.
    Guidelines:
    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).
11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?
Answer: [N/A]
Justification: We propose a new model that analyzes existing datasets, and hence poses no such risks.
Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**
Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?
Answer: [Yes]
Justification: They are introduced in Section F.
Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**
Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?
Answer: [N/A]
Justification: We do not release new assets.
Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**
Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?
Answer: [N/A]

Justification: Not involved.
Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?
Answer: [N/A]
Justification: Not involved.
Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.
Answer: [N/A]
Justification: LLM is only used for polishing language and facilitating standard experiments.
Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.