SELECTIVE RANKINGS FROM HETEROGENEOUS INCOMPLETE PREFERENCES

Anonymous authors
Paper under double-blind review

ABSTRACT

Many model alignment and recommendation systems aggregate diverse human preferences into a single ranking, concealing substantial disagreement and favoring a narrow majority. When preferences are only partially observed, even the direction of the population majority may be uncertain. We introduce *SE-RO* (**S**elective **E**vidence-based **R**ank **O**rdering), a method for inferring selective rankings from incomplete pairwise preferences. *SE-RO* constructs tiered partial orders with separate requirements for how broadly a preference must be shared and how much posterior evidence is needed to display it. Under representative sampling, we bound the risk of displaying comparisons with insufficient population support and give conditions for recovering the population ranking. We also quantify how much recording bias a fixed ranking can tolerate. We use real-world preference tasks and synthetic populations to test common ranking failures and characterize the trade-off between population support, evidence, and coverage. Our results show how selective rankings can avoid overruling divided groups while retaining comparisons with broad population support.

1 INTRODUCTION

Model alignment and recommendation systems serve populations with heterogeneous preferences. These systems often use human feedback to shape their outputs, but the raters supplying comparisons may hold conflicting values and will not always share one preference order [36]. A line of research on pluralistic alignment argues that models should accommodate this variation in preferences [32]. Preference-aggregation methods used in alignment, including Bradley–Terry and Borda, assign scores to items to produce a near-total ranking [23, 31]. The resulting order, however, does not reveal how much population support each comparison has or how strong the evidence is. Preference-based training then optimizes for higher-scoring responses despite substantial disagreement in the population. To account for this range of preferences, we may instead leave contested comparisons unresolved and train on preferences with broad population support (Fig. 1).

In practice, we must infer population support from incomplete preference data. We observe user preferences from only a sample of the target population, and each item pair may be judged by an even smaller part of that sample. A majority among just a few observed comparisons may be enough for common ranking methods to order a pair, even though its true population support remains uncertain.

We introduce *SE-RO*, which represents population preferences as a tiered partial order called a *selective ranking*. The support floor σ sets the minimum population support, while α sets the posterior probability required to meet this floor. The resulting selective ranking orders only comparisons that meet both requirements. *SE-RO* admits these comparisons pairwise and then groups the items into ordered tiers, merging any cycles that exist.

We make the following contributions.

1. We formalize the problem of recovering a selective ranking, which orders only the comparisons that a population broadly supports and withholds the rest.
2. We introduce *SE-RO* to recover a selective ranking from incomplete pairwise preferences. Its support and evidence requirements are separate, so each displayed comparison has population support of at least σ with posterior probability at least α.

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

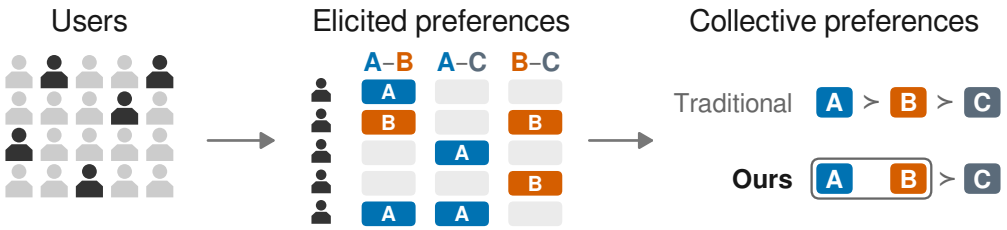


Figure 1: Each sampled user compares only a few pairs. A two-to-one split is weak evidence that most of the population prefers A to B, yet a traditional ranking orders them anyway. Ours orders a pair only given enough evidence ($\alpha = 0.70$) of support ($\sigma = 0.60$), so A and B share a tier.

3. We develop a procedure to set and audit *SE-RO* in practice. Our guarantees state what σ and α promise under representative sampling, and we show how to choose σ to protect every group above a chosen size. The audit reports how much recording bias each displayed comparison can tolerate.
4. On synthetic and real-world data, *SE-RO* avoids overruling divided groups while keeping comparisons with broad support. We provide a demonstration of how labels from *SE-RO* can train a model that withholds contested comparisons on unseen `Community Alignment` prompts.
5. We provide an open-source Python implementation of *SE-RO*, available at anonymized repository.

Related Work. *Pluralistic alignment*. Our work builds on research asking how models can serve populations with diverse values and judgments. Standard annotation and preference-learning pipelines often collapse judgments into a single label or score despite subjectivity or divergent preferences [25, 31, 37]. As a result, reward models and LLM judges fail to account for disagreement between annotators [37]. In this work, we focus on aggregation before preference-based training. We ask which comparisons have enough support to enter a selective ranking and which should remain unresolved. Displayed comparisons can serve many downstream uses, such as pairwise training labels.

Selective aggregation and partial rankings. Partial rankings leave comparisons unresolved when the evidence is uncertain or when preferences conflict. Statistical methods address uncertain evidence by inferring ties or returning confidence sets of plausible rankings [1, 20, 34]. Ranking methods with abstention declare items incomparable at a small cost rather than risk a wrong prediction [5, 17]. They abstain when a prediction is unreliable rather than when the population is divided. Selective judging and ranking apply confidence thresholds or risk controls to decide which comparisons to accept [2, 9, 12]. Other work represents conflicting preferences without forcing a total order [4, 15]. We use the graph construction and tier consolidation of *SPA* [13] and develop admission rules for incomplete pairwise observations that account for both population support and uncertainty.

Incomplete preferences. Methods that work with incomplete preferences differ in what they assume about missing comparisons. Preference-completion methods infer missing comparisons through low-rank structure across users and items [22, 24]. Score-based methods recover rankings from estimated item strengths or average win probabilities [21, 30]. Across these works, missingness is commonly assumed to be independent of preferences or to depend only on observed information [28]. We likewise assume pairs are assigned independently of users' preferences, and additionally assume users are sampled at random from a stated target population. In practice, some preferences are more likely to be recorded than others, and observed support may not reflect population support [29]. We adapt Rosenbaum's sensitivity analysis for observational studies [27] to measure how robust each displayed comparison in a fixed ranking is to recording bias.

2 Problem Statement

Our goal is to infer the population selective ranking S_σ from incomplete preference data, where S_σ is the selective ranking determined by the population support of every pair. We estimate this support from sampled users and admit a comparison only when the evidence is strong enough. We observe

pairwise preferences over n items from p users sampled from a target population. Items are indexed by $i, j \in [n]$ and sampled users by $k \in [p]$. The preference of user k between items i and j is

$$\pi_{ij}^k = \begin{cases} +1 & \text{user } k \text{ strictly prefers } i \succ j, \\ 0 & \text{user } k \text{ ties } i \text{ and } j, \\ -1 & \text{user } k \text{ strictly prefers } j \succ i, \end{cases} \qquad \text{with} \qquad \pi_{ij}^k = -\pi_{ji}^k.$$

Let $\theta_{ij} = \Pr(\pi_{ij}^k = +1)$ denote population strict support for $i \succ j$. We require each comparison in the population ranking to align with at least a fraction $\sigma \in (1/2, 1]$ of the population. We call σ the *support floor*. We denote the set of comparisons that meet this floor by

$$\mathcal{P}_\sigma = \{(i, j) : \theta_{ij} \geq \sigma\}.$$

The population selective ranking S_σ is the unique finest tiering whose comparisons meet this floor (Thm. 2).

Assumption 1 (Sampling and pairwise support model)**.**

- *Observations.* Let $R_{ij}^k = 1$ if we observe user k's preference on $\{i, j\}$, and 0 otherwise. Let R denote all such indicators for $i < j$. The observed, support, and non-support counts are

$$n_{ij} = \sum_k R_{ij}^k, \qquad s_{ij} = \sum_{k: R_{ij}^k = 1} \mathbf{1}\{\pi_{ij}^k = +1\}, \qquad \ell_{ij} = n_{ij} - s_{ij}.$$

  Observed ties and reverse preferences are included as non-support for $i \succ j$.

- *Representative pair samples.* Given the recorded users, each observed label supports $i \succ j$ with probability θ_{ij}, so $s_{ij} \mid R, \theta_{ij} \sim \text{Binomial}(n_{ij}, \theta_{ij})$ for each ordered pair. We do not assume independence across pairs.

- *Pairwise support.* Let D_{ij} denote the observed preferences for pair $\{i, j\}$. Under a uniform prior $\theta_{ij} \sim \text{Beta}(1, 1)$, the posterior for each ordered direction is

$$\theta_{ij} \mid D_{ij} \sim \text{Beta}(1 + s_{ij}, 1 + \ell_{ij}).$$

  We do not specify a joint posterior across directions or pairs.

- *Sparse assignment.* We sample p independent users from the target population and record each user's preferences on m of the $N = \binom{n}{2}$ pairs, where $m \in \{1, \ldots, N\}$. We draw each user's pairs uniformly without replacement and independently of other users. Each label supports $i \succ j$ with probability θ_{ij}, and $n_{ij} \sim \text{Binomial}(p, m/N)$.

Admission. For a support floor σ and evidence threshold $\alpha \in (\frac{1}{2}, 1)$, we admit $i \succ j$ iff

$$q_{ij}(D) := \Pr(\theta_{ij} \geq \sigma \mid D_{ij}) \geq \alpha.$$

For fixed observed preferences, raising σ requires broader population support while raising α requires stronger evidence. The conditions $\sigma > 1/2$ and $\alpha > 1/2$ ensure that at most one direction is admitted per pair (Lem. 4). We summarize our notation in Appendix A.1.

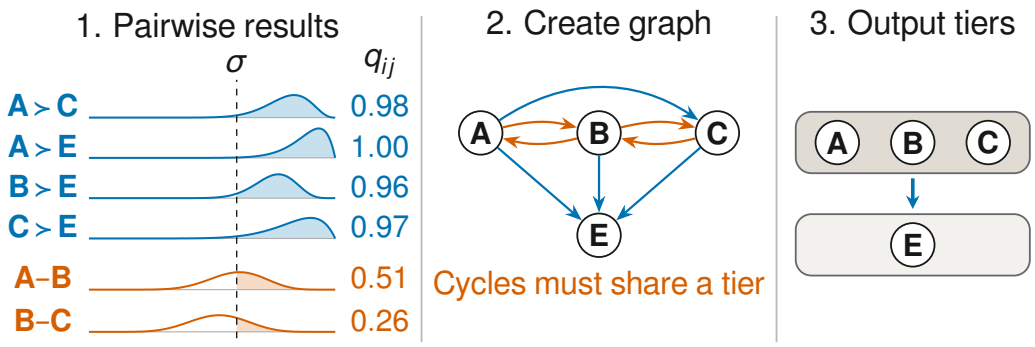


Figure 2: At $\sigma = 0.60$ and $\alpha = 0.90$, A–B and B–C remain unresolved, so A, B and C share a tier above E. The ranking withholds the admitted $A \succ C$ rather than order an unresolved pair.

3 ALGORITHM

Algorithm 1 constructs a selective ranking $S_{\sigma,\alpha}$ that estimates the population selective ranking S_σ from pairwise observations D. The first stage *admits* directions whose population support meets the floor σ with posterior probability at least α. The second stage forms tiers and *displays* admitted comparisons between them. We illustrate tier construction in Fig. 2.

Algorithm 1 *SE-RO*

Input: item set $[n]$, pairwise observations D
Input: $\sigma \in (\frac{1}{2}, 1)$, $\alpha \in (\frac{1}{2}, 1)$ — *Support floor and evidence threshold*
$V_I \leftarrow \{1, \ldots, n\}$
$A_I \leftarrow \{(i \to j) \mid i, j \in V_I,\ i \neq j\}$ — *Initialize edges for each pair*
for each ordered pair (i, j) **do**
  if $q_{ij}(D) \geq \alpha$ **then** — *Admit $i \succ j$*
    $A_I \leftarrow A_I \setminus \{(j \to i)\}$ — *Remove $j \succ i$*
  end if
end for
$V_T \leftarrow \mathsf{StronglyConnectedComponents}(V_I, A_I)$ — *Group items into tiers*
$A_T \leftarrow \{(T \to T') \mid T \neq T',\ \exists i \in T, j \in T' : (i \to j) \in A_I\}$ — *Connect tiers*
$l_1, \ldots, l_{|V_T|} \leftarrow \mathsf{TopologicalSort}(V_T, A_T)$ — *Order tiers by their constraints*
$S_{\sigma,\alpha} \leftarrow (T_{l_1}, T_{l_2}, \ldots, T_{l_{|V_T|}})$
Output: $S_{\sigma,\alpha}$ — *Selective ranking*

We initialize the graph with directed arcs in both directions between every pair of items. If $i \succ j$ is admitted, we remove $j \to i$. Once all such arcs are removed, we consolidate each remaining strongly connected component into a single tier. For fixed σ and α, computing $S_{\sigma,\alpha}$ takes $O(|D| + n^2)$ time and memory, assuming constant-time Beta-tail evaluations.

A tiered ranking T sorts the items into ordered tiers. A comparison $i \succ j$ is displayed when i is placed in a higher tier than j. We use $\mathcal{D}(T)$ for the displayed comparisons in T. T is *feasible* for a set $\mathcal{P}$ of admitted comparisons if $\mathcal{D}(T) \subseteq \mathcal{P}$. The finest feasible ranking displays every comparison that any other feasible ranking displays. By construction, we return this ranking [13].

Theorem 2 (Graph condensation)**.** *Let $\mathcal{P}$ contain at most one direction for each unordered pair. The graph construction returns the unique finest feasible tiering for $\mathcal{P}$.*

Every comparison displayed by $S_{\sigma,\alpha}$ is therefore admitted.

Corollary 3 (Pairwise display guarantee)**.** *For each comparison $i \succ j$ displayed by $S_{\sigma,\alpha}$, $q_{ij}(D) \geq \alpha$ and $\Pr(\theta_{ij} < \sigma \mid D_{ij}) \leq 1 - \alpha$.*

On Preference-Dependent Recording. Without bounds on recording bias, observed counts alone cannot identify population support (Thm. 12). Algorithm 2 takes a fixed ranking S and reports how much recording bias each displayed comparison can tolerate (Fig. 3).

For each displayed comparison $i \succ j$, let L_{ij} be the $(1 - \alpha)$ quantile of its Beta posterior, a lower bound on support among recorded users. For example, at $\alpha = 0.90$, L_{ij} is the posterior's 10th percentile. The comparison still meets the support floor σ under recording bias up to

$$\Gamma_{ij} = \frac{L_{ij}}{1 - L_{ij}} \frac{1 - \sigma}{\sigma}.$$

The ranking $S_{\sigma,\alpha}$ tolerates recording bias up to Γ_{rank}, the smallest Γ_{ij} (Propositions 11 and 13).

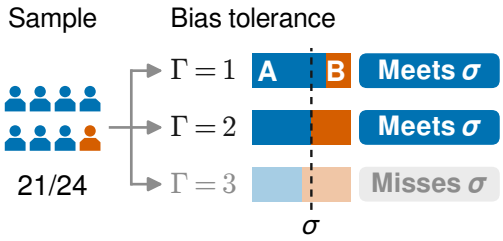


Figure 3: Recording bias can make the 75% lower bound on recorded support ($\alpha = 0.90$) overstate population support. If A's supporters are twice as likely to be recorded, population support can be as low as 60%. $A \succ B$ still meets $\sigma = 0.60$, so its tolerance is $\Gamma_{AB} = 2$. Each icon represents 3 users.

Algorithm 2 Recording Tolerance(D, S, σ, α)

Input: observations D, fixed ranking S
Input: $\sigma \in (\frac{1}{2}, 1)$, $\alpha \in (\frac{1}{2}, 1)$
for each comparison $i \succ j$ displayed by S **do**
  $L \leftarrow (1-\alpha)$ quantile of $\mathrm{Beta}(1 + s_{ij}, 1 + \ell_{ij})$ — *Lower bound on recorded support*
  $o \leftarrow L/(1-L)$ — *Recorded support ratio*
  $\Gamma_{ij} \leftarrow o\,(1-\sigma)/\sigma$ — *Bias tolerance*
end for
$\Gamma_{\mathrm{rank}} \leftarrow$ smallest Γ_{ij} computed — *Undefined if S displays nothing*
Output: each Γ_{ij} and their minimum Γ_{rank}

On Safe Displays and Exact Recovery. Over repeated samples, we bound the probability of displaying an unsupported comparison and give conditions for recovering the population ranking (Thm. 7). At any sample size, admission requires the same number of supporting observations as a one-sided exact binomial test of $\theta_{ij} \leq \sigma$ at level $1 - \alpha$, or one fewer (Prop. 5). The smaller the margin between population support and the support floor, the more observations these guarantees require. For $\sigma = 0.60$ and $\alpha = 0.90$, an exact binomial calculation shows that a direction with 90% population support is admitted with probability at least 0.90 after 14 observations, one with 70% support after 155, and one with 62% support after 3,965. We give proofs in Appendix B.

4 Experiments

On four real-world datasets and four synthetic populations, we test whether *SE-RO* leaves contested comparisons unresolved. We also test whether it recovers comparisons with broad population support, and how limited evidence constrains coverage. Unless stated otherwise, we set $\sigma = 0.60$ and $\alpha = 0.90$. We write 0.5^+ for a value just above 0.5.

4.1 Setup

Real-world datasets. We use four real-world datasets spanning sensory judgments, food and humor preferences, and electoral choices. In `EnviroSound` [38], 74 participants compare all pairs of twelve sounds by unpleasantness. `Sushi` [14] contains complete rankings of ten sushi types from 5,000 participants. From `Jester` [10], we use ratings of ten jokes from 5,000 users. `ERS Set 77` [18] contains incomplete rankings of twelve election alternatives from 3,419 voters. We convert rankings and ratings to pairwise preferences.

Synthetic populations. Each synthetic population has ten items and known support for all 45 unordered pairs. In `Acyclic`, `Cyclical`, and `Ties`, we draw each user's judgments independently across pairs, allowing intransitive preferences. `Acyclic` has three groups with 85% support for higher-group items. `Cyclical` modifies `Acyclic` by adding a three-item cycle in the first group, with 85% population support for each cycle direction. `Ties` modifies `Acyclic` by adding individual ties, with 45% population support for each strict direction and 10% for ties within groups. Across groups, the higher group has 80% population support. We use `Two Groups` to test whether *SE-RO* preserves disagreement between two groups, each unanimously ranking three items in opposite orders. The groups comprise 55% and 45% of the population. Three other pairs have 65% population support for one direction and three have 90%. We provide further details on datasets, parameter sweeps, and diagnostics in Appendix C.

For each synthetic population, we draw two independent panels and evaluate each ranking against known population support. We construct one ranking from each panel, yielding two rankings per dataset across 100 random seeds. Unless stated otherwise, each panel has $p = 250$ users with $m = 12$ assigned pairs per user. For real-world data, we split respondents into two random halves. We construct a ranking from each half and evaluate it using only comparisons from the other half. We use 100 random seeds to measure variation across sampled panels. For each seed, we pool the two outputs' metric counts and then average the seed-level metrics, giving each seed equal weight. We report each fixed ranking's recording tolerance Γ_{rank}.

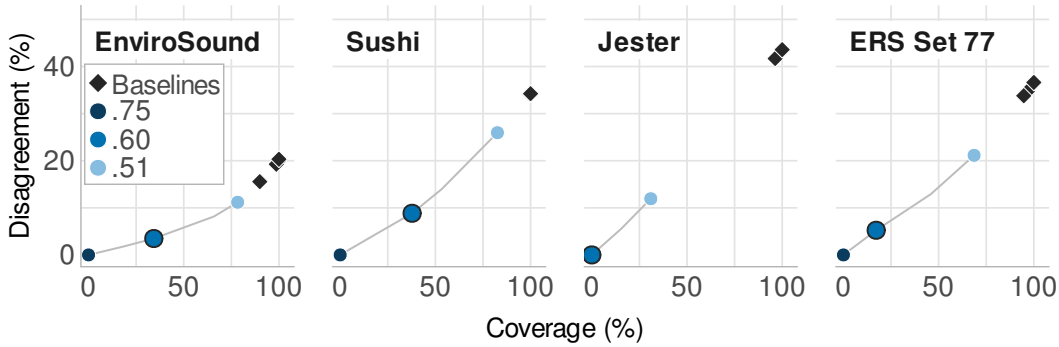


Figure 4: Lowering σ from 0.75 to 0.51 (*SE-RO*, blue dots, $\alpha = 0.90$) shows more pairs but overrules more held-out preferences. Baselines order nearly every pair and overrule 16–44%.

Methods. We compare *SE-RO* with aggregation methods from three families. We give implementation details in Appendix C.

- **Voting rules.** Voting rules rank items by aggregate pairwise scores. We use Borda and Copeland [6, 8]. Borda ranks items by their total observed wins, with ties as a half-win. Copeland ranks items by the number of opponents they beat, where a win means that more observed users prefer the item than its opponent.
- **Spectral ranking methods.** Spectral methods infer item strengths from the network of pairwise comparisons. We use Rank Centrality [21], which ranks items by how often a random walk visits them in the long run, favoring moves toward items that win their observed comparisons.
- **Probabilistic ranking.** Probabilistic methods fit a model of pairwise outcomes and rank items by their inferred strengths. We use MB–K [20], which fits a Bradley–Terry model with one strength per item group. After greedily merging groups, the method keeps the first grouping with the shortest description length and compares items across groups with different strengths.

Evaluation. We evaluate a ranking S with displayed comparisons $\mathcal{D}(S)$ against reference support h_{ij}. For synthetic populations, we set $h_{ij} = \theta_{ij}$ and $v_{ij} = 1$. For each real-data ranking, we compute s_{ij} and n_{ij} on the held-out half and set $h_{ij} = s_{ij}/n_{ij}$ and $v_{ij} = n_{ij}$. Let $V = \sum_{i<j} v_{ij}$.

Metric	Definition	Meaning
Coverage	$\lvert\mathcal{D}(S)\rvert / \binom{n}{2}$	Share of pairs displayed
Agreement	$\lvert\mathcal{D}(S)\rvert^{-1} \sum_{(i,j)\in\mathcal{D}(S)} h_{ij}$	Mean support of displayed pairs
Individual disagreement	$V^{-1} \sum_{(i,j)\in\mathcal{D}(S)} v_{ij}(1 - h_{ij})$	Share of preferences overruled

4.2 Results

Fig. 4 shows individual disagreement and coverage on the datasets. On the four synthetic populations at $\sigma = 0.60$ and $\alpha = 0.90$, *SE-RO* displays 57.4–87.6% of pairs and overrules 1.0–11.5% of preferences. Borda, Copeland, and Rank Centrality overrule more on every population (5.7–29.3%, Tables 5 and 6) because they order nearly every pair.

On Population Disagreement. A ranking that follows the overall majority can overrule a coherent minority group. The support floor can prevent this without group labels. If a strict majority of a group with population share at least g opposes a comparison, its support is below $1 - g/2$. A support floor $\sigma \geq 1 - g/2$ therefore withholds from S_σ any comparison opposed by a majority of such a group. If specific groups are known, g can be selected to match their proportions. The same σ also protects unknown groups with share at least g.

On `Two Groups`, more evidence lets *SE-RO* recover supported comparisons while preserving disagreement between groups (Fig. 5). We set $\sigma = 0.60$ and compare panels of 250 to 2,000 users. In Fig. 5, each user compares all pairs. MB–K orders more contested pairs as the panel grows, from 14.7% at 250 users to 76.0% at 2,000. When each user compares only 12 pairs, *SE-RO* retains all 39 unanimous or 90%-support comparisons in every output at 250 users. It displays 13.8% of the 65%-support directions at 250 users and 78.0% at 2,000, while all 55%/45% pairs remain unordered.

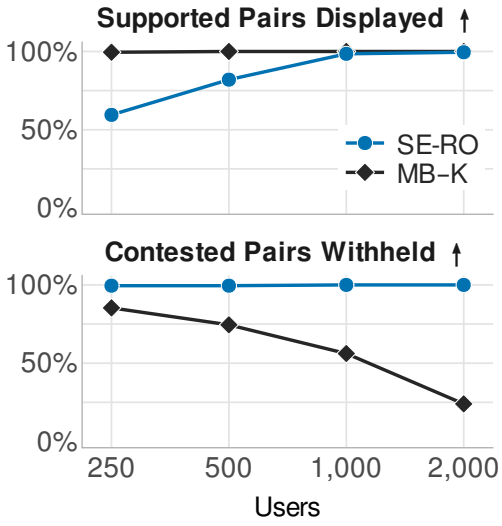


Figure 5: `Two Groups`, all pairs observed. With more users, *SE-RO* displays more of the supported (65%) pairs and still withholds nearly all contested (55%/45%) ones. MB–K orders more contested pairs, overruling one group.

We observe a similar pattern in `Sushi`, a real-world dataset. While 58.8% of all participants preferred sea urchin over egg, 63.5% of participants aged 19 or younger preferred the reverse. Among participants aged 20 or older, 60.3% preferred sea urchin over egg. Across all 200 samples, *SE-RO* withholds this pair at $\sigma = 0.60$, where its overall support falls below the floor, and displays sea urchin above egg at $\sigma = 0.55$. MB–K, Borda, Copeland, and Rank Centrality rank sea urchin above egg in every sample, overruling the younger group's majority preference. With *SE-RO*, we can choose to leave this disagreement unresolved rather than impose the overall majority's preference.

On Finite Evidence. Under representative sampling, a direction with sufficient population support should still be withheld when the evidence is insufficient. At $\sigma = 0.60$ and $\alpha = 0.90$, admission requires 5 supporting observations out of 5, but 67 out of 100. As observations accumulate, the required fraction approaches the 60% population support floor.

When our task requires greater coverage, we can accept comparisons with less population support or weaker evidence. On the 250-user `Two Groups` panels with 12 pairs per user, lowering σ and α to 0.5^+ raises coverage from 87.6% to 98.8%. Individual disagreement rises from 1.0% to 5.6%, slightly lower than Borda, Copeland, and Rank Centrality at similar coverage (5.7–6.4%). MB–K is fixed between the two settings, displaying 91.6% of pairs with 2.4% disagreement.

Table 1: On `Jester`, *SE-RO* admits directions for many pairs but displays few, since every comparison between tiers must be admitted. Lowering σ trades population support for coverage ($\alpha = 0.5^+$, % of pairs). Agreement averages outputs with nonempty displays.

σ	Admitted	Coverage	Agreement
0.60	23.2	0.0	—
0.58	31.7	0.4	63.5
0.55	55.3	20.0	64.1
0.53	67.3	20.8	64.0
0.51	86.9	78.0	57.9

On Recording Bias. Our guarantees assume that supporters and nonsupporters of a comparison are equally likely to be recorded. On `Two Groups`, we instead record each supporter with twice the probability of a nonsupporter ($\rho = 2$). Recorded support for the 55%-support directions then rises to 71.0%, above the 60% floor. More observations only reinforce this bias, and *SE-RO* displays 57.3% of these directions at 250 users and 99.3% at 1,000. We therefore report Γ_{rank}, the largest recording bias under which every displayed comparison still meets the support floor. At $\sigma = 0.60$ and $\alpha = 0.90$, the median Γ_{rank} allows supporters to be 11% more likely to be recorded than nonsupporters on `EnviroSound`, 17% on `ERS Set 77`, and 20% on `Sushi`. Building the ranking at a higher σ or α never lowers a nonempty output's tolerance at $\sigma = 0.60$ and $\alpha = 0.90$ but may display fewer comparisons. Practitioners can vary both to find the balance of coverage and tolerance they are comfortable with.

On Coherence and Coverage. Common ranking methods impose a transitive total or near-total order on items, but individual preferences need not be transitive [33]. This intransitivity can even extend to population support, as in a 2021 Minneapolis city council election [19]. We test how ranking methods handle such a cycle with `Cyclical`, where any ranking that splits the three cycle items across tiers overrules 85% of the population on at least one pair. Borda and Rank Centrality display nearly every pair but overrule 24% of preferences. *SE-RO* keeps the cycle items in one tier above the other seven. At $\sigma = 0.60$, it recovers the population ranking in every output at both $\alpha = 0.70$ and $\alpha = 0.90$.

On `Jester` at $\sigma = 0.60$ and $\alpha = 0.5^{+}$, *SE-RO* admits directions for 23.2% of pairs but displays none (Table 1). Every tiered ranking of the ten jokes with two or more tiers would display some comparison without enough support or evidence. *SE-RO* therefore places all ten jokes in a single tier. Lowering σ raises coverage to 20.0% at $\sigma = 0.55$ and 78.0% at $\sigma = 0.51$. *SE-RO* displays more comparisons only when the practitioner accepts less population support.

5 Learned Withholding on Community Alignment

Preference-based training optimizes for responses that raters rank highly, even where raters disagree. *SE-RO* can instead supply training labels that leave contested comparisons unresolved. We first compare these labels with common aggregation methods. We then train a model on these labels and test whether its withholding generalizes to unseen prompts.

Setup. We use `Community Alignment`, a preference dataset collected from evaluators in five countries [36]. Each prompt is a request to a language model and comes with four candidate responses. Many evaluators see the same prompt, and each picks a top response i. Each pick gives three comparisons $i \succ j$, with the rest missing. We note that we only see $i \succ j$ when i is the top choice, so support for i over j may differ from population support θ_{ij}. We use five deterministic splits of source prompts into 60% training, 20% calibration and 20% test, keeping response sets that at least ten evaluators rated.

In each split, the methods see the votes of 70% of evaluators. We hold out the other 30% for evaluation. We report coverage and agreement from Section 4. Both metrics use only pairs that held-out evaluators voted on.

Methodology. We compare methods on two tasks (Table 2). The first task aggregates the votes on each test prompt to rank its four responses. At $\sigma = 0.60$ and $\alpha = 0.70$, *SE-RO* labels each pair that received votes with its admitted direction or *withhold*. We compare it with three rules on the same votes:

- Vote Share admits a direction when its vote share reaches a threshold.
- BT-Margin admits a direction when its gap in fitted Bradley–Terry scores [3] reaches a threshold.
- BT fits a Bradley–Terry model with a tie term [7] and orders nearly every pair.

Each threshold is tuned to match the number of calibration pairs that *SE-RO* orders.

The second task predicts labels for unseen prompts. A linear model over fixed Qwen2.5-0.5B embeddings [26] learns to predict the label of *SE-RO* for each pair. It sees only the prompt and its four responses. We pass its predictions through the graph stage of Algorithm 1 to form tiers. We compare it with a BT model trained on the same embeddings, which orders every pair. Appendix D gives rule definitions, model details and a worked example.

Results. BT orders 91.8% of pairs with 71.4% agreement (Table 2). *SE-RO* withholds more than half of the pairs and raises agreement to 81.4%. At $\sigma = 0.51$ and $\alpha = 0.5^{+}$, *SE-RO* orders 87.8% of pairs with 72.2% agreement, close to BT. Withholding more pairs buys higher agreement.

At similar coverage, *SE-RO* agrees more often with held-out voters than Vote Share and BT-Margin. To match the coverage of *SE-RO*, Vote Share must require 83–100% of votes across splits. Neither rule accounts for the uncertainty left by few votes. *SE-RO* orders a pair only when enough votes back its direction. On 598 source prompts outside the five splits, *SE-RO* again beats Vote Share by 4.8 points of agreement.

Table 2: Coverage and held-out agreement (%) at $\sigma = 0.60$ and $\alpha = 0.70$. *SE-RO* orders more pairs than Vote Share and BT-Margin and agrees more often with held-out voters.

	Aggregation				*Unseen prompts*	
	SE-RO	Vote Share	BT-Margin	BT	Learned model	BT model
Coverage	44.1	40.2	40.0	91.8	42.3	100.0
Agreement	81.4	77.0	76.6	71.4	75.2	64.7

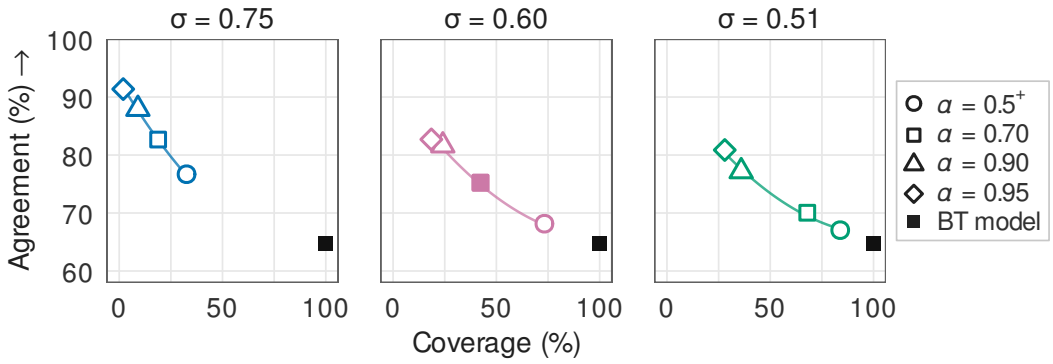


Figure 6: Coverage and held-out agreement of the learned model, averaged over five splits. Curves are quadratic fits across ten values of α. The filled pink square marks $\sigma = 0.60$ and $\alpha = 0.70$.

On unseen prompts, the learned model orders 42.3% of pairs with 75.2% agreement. The BT model orders every pair with 64.7% agreement. This 10.5-point gain matches the 10-point gain of *SE-RO* over BT. Even at $\sigma = 0.51$ and $\alpha = 0.5^+$, the learned model orders 83.9% of pairs with 67.0% agreement. The withholding of *SE-RO* therefore generalizes to unseen prompts. Raising σ or α when creating the training labels generally lowers the coverage of the resulting model and raises its agreement (Fig. 6). Practitioners can therefore decide how much the model withholds before it is trained.

6 Concluding Remarks

Incomplete preferences leave us uncertain about which comparisons a population supports and where we should abstain. *SE-RO* extends selective aggregation to this setting by constructing rankings from limited pairwise evidence. On synthetic and real-world data, *SE-RO* avoids overruling divided groups while keeping comparisons with broad support. Methods that order nearly every pair overrule more held-out preferences. Our experiments expose the trade-off between coverage, population support, and evidence. A higher support floor protects smaller groups but leaves more comparisons unresolved. We also report how sensitive a fixed ranking is to recording bias by quantifying the recording imbalance it tolerates. On `Community Alignment`, a model trained on labels from *SE-RO* withholds comparisons on unseen prompts.

Our approach assesses each pair separately under representative sampling. Γ_{rank} reports how much recording bias a ranking tolerates but cannot tell us how much is truly present. The tiers can also withhold admitted comparisons, since every comparison between tiers must be admitted.

Sharing information across pairs could reduce the observations needed to support a comparison. This requires models that estimate population support and quantify the evidence that it exceeds the chosen support floor. Future work should model when pooling improves these estimates and how errors in a shared model affect the comparisons displayed in the ranking.

AI USE STATEMENT

We used generative AI tools to discuss candidate methods for inference from sparsely sampled preferences, generate substantial portions of the codebase, run experiments and produce figures, identify existing real-world datasets, and generate synthetic datasets according to author-specified designs. We also used these tools to discuss results and their implications, review and revise mathematical arguments, suggest citations for review and highlight key passages, and assist with drafting and revising the manuscript. These tools also helped formulate mathematical claims, provide key steps in their proofs, write the proofs, design experiments and give feedback on them, and clean and reformat datasets. The authors verified AI-generated code through direct inspection and checks against simple toy examples. We inspected the data and consulted online dataset descriptions to check our understanding and use of the datasets. The authors take responsibility for the final content of this work, including its methods, claims, code, and results.

REFERENCES

[1] Angel Rodrigo Avelar Menendez, Yufeng Liu, and Xiaowu Dai. Prompt-dependent ranking of Large Language Models with uncertainty quantification. *arXiv preprint arXiv:2603.03336*, 2026. URL https://arxiv.org/abs/2603.03336.

[2] Sher Badshah, Ali Emami, and Hassan Sajjad. SCOPE: Selective conformal optimized pairwise LLM judging. In *International Conference on Machine Learning*, 2026. URL https://icml.cc/virtual/2026/poster/62431.

[3] Ralph Allan Bradley and Milton E. Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952. doi: 10.1093/biomet/39.3-4.324.

[4] Yen-Liang Chen and Li-Chen Cheng. Mining maximum consensus sequences from group ranking data. *European Journal of Operational Research*, 198(1):241–251, 2009. doi: 10.1016/j.ejor.2008.09.004.

[5] Weiwei Cheng, Michaël Rademaker, Bernard De Baets, and Eyke Hüllermeier. Predicting partial orders: Ranking with abstention. In *Machine Learning and Knowledge Discovery in Databases*, volume 6321 of *Lecture Notes in Computer Science*, pp. 215–230, 2010. doi: 10.1007/978-3-642-15880-3_20.

[6] Arthur H. Copeland. A reasonable social welfare function, 1951. Mimeographed notes from a seminar on applications of mathematics to social sciences, University of Michigan.

[7] Roger R. Davidson. On extending the Bradley-Terry model to accommodate ties in paired comparison experiments. *Journal of the American Statistical Association*, 65(329):317–328, 1970. doi: 10.1080/01621459.1970.10481082.

[8] Jean-Charles de Borda. Mémoire sur les élections au scrutin. *Histoire de l'Académie Royale des Sciences*, pp. 657–665, 1781. URL https://www.jehps.net/juin2008/Document2.pdf.

[9] Yuxuan Gao, Megan Wang, and Yi Ling Yu. Distribution-free uncertainty quantification for continuous AI agent evaluation. *arXiv preprint arXiv:2605.19779*, 2026. URL https://arxiv.org/abs/2605.19779. Accepted at the ICML 2026 Workshop on Agentic Uncertainty Quantification (AgenticUQ).

[10] Ken Goldberg, Theresa Roeder, Dhruv Gupta, and Chris Perkins. Eigentaste: A constant time collaborative filtering algorithm. *Information Retrieval*, 4(2):133–151, 2001. doi: 10.1023/A:1011419012209.

[11] Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979. URL https://www.jstor.org/stable/4615733.

[12] Jaehun Jung, Faeze Brahman, and Yejin Choi. Trust or escalate: LLM judges with provable guarantees for human agreement. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/08dabd5345b37fffcbe335bd578b15a0-Abstract-Conference.html. arXiv:2407.18370.

[13] Shreyas Kadekodi, Hayden McTavish, and Berk Ustun. Selective preference aggregation. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 28644–28669, 2025. URL https://proceedings.mlr.press/v267/kadekodi25a.html.

[14] Toshihiro Kamishima. Nantonac collaborative filtering: Recommendation based on order responses. In *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 583–588, 2003. doi: 10.1145/956750.956823. Data release: https://www.kamishima.net/sushi/.

[15] Dongqing Li, Geoff K. Nicholls, Jeong Eun Lee, and Chuxuan Jiang. Hierarchical partial-order models for ranking. *arXiv preprint arXiv:2606.25062*, 2026. URL https://arxiv.org/abs/2606.25062.

[16] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://arxiv.org/abs/1711.05101.

[17] Anqi Mao, Mehryar Mohri, and Yutao Zhong. Ranking with abstention. In *ICML 2023 Workshop on the Many Facets of Preference-Based Learning*, 2023. URL https://openreview.net/forum?id=B8mQ1eOTOo. arXiv:2307.02035.

[18] Nicholas Mattei and Toby Walsh. PrefLib: A library for preferences. In *Proceedings of the 3rd International Conference on Algorithmic Decision Theory (ADT 2013)*, volume 8176 of *Lecture Notes in Artificial Intelligence*, pp. 259–270. Springer, 2013. doi: 10.1007/978-3-642-41575-3_20. URL https://www.preflib.org.

[19] David McCune and Lori McCune. The curious case of the 2021 Minneapolis Ward 2 City Council election. *The College Mathematics Journal*, 57(1):16–20, 2026. doi: 10.1080/07468342.2023.2212548. URL https://doi.org/10.1080/07468342.2023.2212548.

[20] Sebastian Morel-Balbi and Alec Kirkley. Estimation of partial rankings from sparse, noisy comparisons. *Communications Physics*, 9(1), 2026. doi: 10.1038/s42005-025-02461-y. Article 30.

[21] Sahand Negahban, Sewoong Oh, and Devavrat Shah. Rank centrality: Ranking from pairwise comparisons. *Operations Research*, 65(1):266–287, 2017. doi: 10.1287/opre.2016.1534.

[22] Sahand Negahban, Sewoong Oh, Kiran K. Thekumparampil, and Jiaming Xu. Learning from comparisons and choices. *Journal of Machine Learning Research*, 19(40):1–95, 2018. URL https://jmlr.org/papers/v19/17-607.html.

[23] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. 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 F. Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pp. 27730–27744, 2022. doi: 10.52202/068431-2011.

[24] Dohyung Park, Joe Neeman, Jin Zhang, Sujay Sanghavi, and Inderjit S. Dhillon. Preference completion: Large-scale collaborative ranking from pairwise comparisons. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *PMLR*, pp. 1907–1916, 2015. URL https://proceedings.mlr.press/v37/park15.html.

[25] Barbara Plank. The "problem" of human label variation: On ground truth in data, modeling and evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 10671–10682, 2022. doi: 10.18653/v1/2022.emnlp-main.731.

[26] Qwen, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report, 2025. URL https://arxiv.org/abs/2412.15115.

[27] Paul R. Rosenbaum. *Observational Studies*. Springer, New York, 2 edition, 2002. doi: 10.1007/978-1-4757-3692-2.

[28] Donald B. Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1976. doi: 10.1093/biomet/63.3.581.

[29] Tobias Schnabel, Adith Swaminathan, Ashudeep Singh, Navin Chandak, and Thorsten Joachims. Recommendations as treatments: Debiasing learning and evaluation. In *Proceedings of the 33rd International Conference on Machine Learning*, volume 48 of *PMLR*, pp. 1670–1679, 2016.

[30] Nihar B. Shah and Martin J. Wainwright. Simple, robust and optimal ranking from pairwise comparisons. *Journal of Machine Learning Research*, 18(199):1–38, 2018. URL https://jmlr.org/papers/v18/16-206.html.

[31] Anand Siththaranjan, Cassidy Laidlaw, and Dylan Hadfield-Menell. Distributional preference learning: Understanding and accounting for hidden context in RLHF. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/7423902b5534e2b267438c85444a54b1-Abstract-Conference.html.

[32] Taylor Sorensen, Jared Moore, Jillian Fisher, Mitchell L. Gordon, Niloofar Mireshghallah, Christopher Michael Rytting, Andre Ye, Liwei Jiang, Ximing Lu, Nouha Dziri, Tim Althoff, and Yejin Choi. Position: A roadmap to pluralistic alignment. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *PMLR*, pp. 46280–46302, 2024. URL https://proceedings.mlr.press/v235/sorensen24a.html.

[33] Amos Tversky. Intransitivity of preferences. *Psychological Review*, 76(1):31–48, 1969. doi: 10.1037/h0026750.

[34] Zebin Wang, Yi Han, Ethan X. Fang, Lan Wang, and Junwei Lu. Confidence diagram of nonparametric ranking for uncertainty assessment in Large Language Models evaluation. *arXiv preprint arXiv:2412.05506*, 2024. URL https://arxiv.org/abs/2412.05506.

[35] Edwin B. Wilson. Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, 22(158):209–212, 1927. doi: 10.1080/01621459.1927.10502953.

[36] Lily Zhang, Smitha Milli, Karen Jusko, Jonathan Smith, Brandon Amos, Wassim Bouaziz, Manon Revel, Jack Kussman, Yasha Sheynin, Lisa Titus, Bhaktipriya Radharapu, Jane Dwivedi-Yu, Vidya Sarma, Kristopher Rose, and Maximilian Nickel. Cultivating pluralism in algorithmic monoculture: The community alignment dataset. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/ee208bfc04b1bf6125a6a34baa1c28d3-Abstract-Conference.html. Dataset at https://huggingface.co/datasets/facebook/community-alignment-dataset.

[37] Michael J. Q. Zhang, Zhilin Wang, Jena D. Hwang, Yi Dong, Olivier Delalleau, Yejin Choi, Eunsol Choi, Xiang Ren, and Valentina Pyatkin. Diverging preferences: When do annotators disagree and do models know? In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 76193–76212. PMLR, 2025. URL https://proceedings.mlr.press/v267/zhang25bx.html.

[38] Karin Zimmer, Wolfgang Ellermeier, and Christian Schmid. Using probabilistic choice models to investigate auditory unpleasantness. *Acta Acustica united with Acustica*, 90(6): 1019–1028, 2004. URL https://www.psychologie.tu-darmstadt.de/media/ellermeier/pubs/zimmer_ell_schmid_aaa_2004.pdf.

Supplementary Materials

Selective Rankings from Heterogeneous Incomplete Preferences

A Supplementary Material for Section 2 **15**
A.1 Symbols and observed data 15

B Proofs **16**
B.1 Pairwise admission and graph construction 16
B.2 Finite-evidence recovery 17
B.3 False-display guarantees 18
B.4 Recovery under randomized sparse assignment 19
B.5 Bounded preference-dependent recording 19
B.6 Unrestricted preference-dependent recording 20
B.7 Fixed assessment of displayed comparisons 21

C Additional Results for Section 4 **23**
C.1 Recording tolerance of fixed rankings 23
C.2 Disagreement, coverage, and stability 23
C.3 Adjusted gate under simulated recording bias 23
C.4 Populations, datasets, and baselines 24
C.5 Admission, false displays, and recovery 25
C.6 Coverage-matched comparison 27
C.7 Minimum evidence and failure frontiers 27

D Additional Results for Section 5 **29**

A SUPPLEMENTARY MATERIAL FOR SECTION 2

A.1 SYMBOLS AND OBSERVED DATA

We summarize our notation below.

Table 3: Notation.

Symbol	Meaning
$i, j \in [n]$	items ($n \geq 2$)
$k \in [p]$	sampled users, or judges ($p \geq 2$)
$N = \binom{n}{2}$	number of unordered item pairs
$\pi_{ij}^k \in \{-1, 0, +1\}$	individual preference of user k on pair $\{i, j\}$
$T = (T_1, \ldots, T_q)$	tiered ranking
$\pi_{ij}(T) \in \{-1, \perp, +1\}$	collective preference expressed by T
$R_{ij}^k \in \{0, 1\}$	indicator that π_{ij}^k is observed ($R_{ji}^k = R_{ij}^k$)
m	pairs assigned to each user
$n_{ij} = \sum_k R_{ij}^k$	users who compared i and j ($n_{ji} = n_{ij}$)
D	observed dataset (assigned labels) $\{(i, j, k, \pi_{ij}^k) : i < j,\ R_{ij}^k = 1\}$
D_{ij}	observed preferences for pair $\{i, j\}$, with its recorded-user set fixed
$\sigma \in (\frac{1}{2}, 1]$	population strict-support floor
$\alpha \in (\frac{1}{2}, 1)$	posterior evidence threshold
$\theta_{ij} = \Pr(\pi_{ij}^k = +1)$	population strict-support rate
$S_\sigma,\ S_{\sigma,\alpha}$	population and *SE-RO* selective rankings
$\mathcal{D}(T)$	comparisons displayed between tiers of T
$L_{ij,\alpha}$	lower Beta posterior bound with upper-tail mass α
Γ_{ij}	pairwise recording tolerance
$\Gamma \geq 1$	declared bound on the aggregate recording-probability ratio

B PROOFS

The recovery proofs (Thm. 7 and Cor. 9) use a sufficient observation count per pair. This threshold depends on the failure probability δ and on the margin η between population support and the floor.

For fixed n, $\alpha \in (\frac{1}{2}, 1)$, $\eta > 0$, and $\delta \in (0, 1)$, let $M = n(n-1)$ and define this threshold as

$$r_{\min}(\eta, \delta) := \left\lceil \max\left\{ \frac{2}{\eta^2} \log \frac{2M}{\delta}, 1 + \frac{2}{\eta^2} \log \frac{1}{1-\alpha} \right\} \right\rceil .$$

B.1 PAIRWISE ADMISSION AND GRAPH CONSTRUCTION

Lemma 4 (Admission antisymmetry)**.** *For every fixed dataset D, every $\sigma \in (\frac{1}{2}, 1]$, and every unordered pair $\{i, j\}$,*

$$q_{ij}(D) + q_{ji}(D) < 1.$$

Proof of Lem. 4. Fix an unordered pair $\{i, j\}$ with x strict preferences for i over j, y for j over i, and z ties, and let $r = x + y + z$. For $u \in \{0, \dots, r\}$, define

$$h_r(u) := \Pr\{\mathrm{Beta}(1 + u, 1 + r - u) \geq \sigma\} .$$

The Beta–Binomial identity gives

$$h_r(u) = \Pr\{Z \leq u\} , \qquad Z \sim \mathrm{Binomial}(r + 1, \sigma),$$

so h_r is nondecreasing. The two directional scores are $q_{ij} = h_r(x)$ and $q_{ji} = h_r(y)$. Because $y \leq r - x$,

$$q_{ji} \leq h_r(r - x).$$

Let $B \sim \mathrm{Beta}(1 + x, 1 + r - x)$. By reflection of the Beta distribution,

$$h_r(r - x) = \Pr(B \leq 1 - \sigma).$$

Since $1 - \sigma < \sigma$ and B has a positive continuous density on $(0, 1)$,

$$\Pr(B \leq 1 - \sigma) < \Pr(B < \sigma) = 1 - h_r(x).$$

Thus $q_{ij} + q_{ji} < 1$, ruling out admission of both directions at $\alpha > 1/2$. □

Uniform prior and exact binomial testing. Other priors require rechecking the admission and finite-sample guarantees (Section C.7).

Proposition 5 (Relation to exact binomial testing)**.** *Fix an observation count $r \geq 0$, a support floor $\sigma \in (1/2, 1)$, and an evidence threshold $\alpha \in (1/2, 1)$. For $s \in \{0, \dots, r\}$, define*

$$q_r(s) = \Pr\{\mathrm{Beta}(s + 1, r - s + 1) \geq \sigma\} , \qquad p_r(s) = \Pr(X \geq s), \quad X \sim \mathrm{Binomial}(r, \sigma).$$

Then

$$1 - q_r(s) = \Pr(X > s) + \sigma \Pr(X = s).$$

Let k_{B} and k_{E} be the smallest support counts satisfying $q_r(s) \geq \alpha$ and $p_r(s) \leq 1 - \alpha$, respectively. Assign either threshold the value $r + 1$ if no feasible count satisfies its condition. Then

$$k_{\mathrm{B}} \leq k_{\mathrm{E}} \leq k_{\mathrm{B}} + 1.$$

Proof of Prop. 5. The Beta–Binomial identity gives

$$1 - q_r(s) = \Pr(Z \geq s + 1), \qquad Z \sim \mathrm{Binomial}(r + 1, \sigma).$$

Write $Z = X + B$, where $B \sim \mathrm{Bernoulli}(\sigma)$ is independent of X. Conditioning on B yields

$$\begin{aligned} 1 - q_r(s) &= (1 - \sigma) \Pr(X \geq s + 1) + \sigma \Pr(X \geq s) \\ &= \Pr(X > s) + \sigma \Pr(X = s). \end{aligned}$$

Consequently,

$$p_r(s + 1) \leq 1 - q_r(s) \leq p_r(s),$$

where $p_r(r + 1) = 0$. The posterior rule and exact test are monotone in s, proving the threshold inequalities. □

Posterior rule versus exact test. The posterior rule counts a fraction σ of the probability mass at the observed count, while the exact test counts all of it. For example, at $\sigma = 0.60$ and $\alpha = 0.90$, a direction supported by all four of its observations has $q_4(4) = 1 - 0.6^5 = 0.92224$ and is admitted. Its exact binomial p-value $0.6^4 = 0.1296$ exceeds $1 - \alpha = 0.10$, so the exact test withholds it. This difference can also change the tiers.

The repeated-sampling admission probability at fixed $\theta_{ij} \leq \sigma$ can therefore exceed $1 - \alpha$. This differs from the posterior probability bounded in Cor. 3.

Proof of Thm. 2. Let $H(\mathcal{P})$ contain $i \to j$ if and only if $(j, i) \notin \mathcal{P}$. Antisymmetry makes $H(\mathcal{P})$ semicomplete. Its condensation K is an acyclic tournament, hence has a unique topological order.

If component C precedes C', no arc $j \to i$ remains for $i \in C$ and $j \in C'$. Hence $(i, j) \in \mathcal{P}$, proving $\mathcal{D}(T_{\mathcal{P}}) \subseteq \mathcal{P}$.

For any feasible tiering T with tier indices $r_T(i)$, a retained arc $i \to j$ forces $r_T(i) \leq r_T(j)$: otherwise T displays $(j, i) \in \mathcal{P}$, which would have deleted that arc. Along directed paths, this makes r_T constant on each strongly connected component and nondecreasing along K. Thus T merges consecutive tiers of $T_{\mathcal{P}}$, the unique finest feasible tiering. □

Proof of Cor. 3. Immediate from Thm. 2 and the admission rule. □

Proposition 6 (One-parameter paths)**.** *For any observed dataset D, increasing σ at fixed α does not add any displayed comparison. Increasing α at fixed σ does not add any displayed comparison.*

Proof of Prop. 6. For fixed data, the posterior tail $q_{ij}(D; \sigma)$ is nonincreasing in σ. Increasing σ at fixed α, or α at fixed σ, therefore only removes admitted directions. If $\mathcal{P}' \subseteq \mathcal{P}$ are the resulting admission sets, the ranking for $\mathcal{P}'$ is feasible for $\mathcal{P}$. By Thm. 2, it is obtained by merging consecutive tiers of the ranking for $\mathcal{P}$, which adds no displayed comparison. □

B.2 Finite-evidence recovery

Theorem 7 (Safe displays and finite-evidence recovery)**.** *Fix n, $\sigma \in (\frac{1}{2}, 1]$, $\alpha \in (\frac{1}{2}, 1)$, $\delta \in (0, 1)$, and $\eta > 0$. Under Assump. 1, condition on a pair assignment R with $n_{ij} \geq r_{\min}(\eta, \delta)$ for every pair $i < j$.*

(a) If every unsupported direction (i, j) satisfies $\theta_{ij} \leq \sigma - \eta$, then

$$\Pr(\mathcal{D}(S_{\sigma,\alpha}) \subseteq \mathcal{D}(S_\sigma) \,|\, R) \geq 1 - \delta.$$

(b) If, in addition, $\sigma < 1$, $\eta < 1 - \sigma$, and every supported direction (i, j) satisfies $\theta_{ij} \geq \sigma + \eta$, then

$$\Pr(S_{\sigma,\alpha} = S_\sigma \,|\, R) \geq 1 - \delta.$$

Section C.7 reports recovery failures in a one-observation zero-margin pilot.

Proof of Thm. 7(b). Write $t_0 = \sigma$ and condition on R. For an ordered pair, let $r = n_{ij}$, $X = s_{ij}$, and $\widehat{\theta}_{ij} = X/r$. Under Assump. 1, $X \sim \mathrm{Binomial}(r, \theta_{ij})$, so Hoeffding's inequality gives

$$\Pr\left(|\widehat{\theta}_{ij} - \theta_{ij}| \geq \frac{\eta}{2} \,\middle|\, R\right) \leq 2\exp\left(-\frac{r\eta^2}{2}\right).$$

A union bound over the $M = n(n-1)$ ordered pairs, using $r_{\min}(\eta, \delta)$ from the start of Section B, shows that

$$\mathcal{E} := \left\{|\widehat{\theta}_{ij} - \theta_{ij}| < \frac{\eta}{2} \text{ for every } (i, j)\right\}$$

holds with probability at least $1 - \delta$.

On $\mathcal{E}$, the Beta–Binomial identity gives

$$q_{ij}(D) = \Pr(Z \leq X), \qquad Z \sim \mathrm{Binomial}(r + 1, t_0).$$

If $\theta_{ij} \geq t_0 + \eta$, then $X/r \geq t_0 + \eta/2$. Hence

$$X + 1 - (r+1)t_0 \geq \frac{r\eta}{2} + 1 - t_0 \geq \frac{r\eta}{2}.$$

A one-sided Hoeffding bound and $r \geq r_{\min}(\eta, \delta)$ give

$$\begin{aligned} 1 - q_{ij}(D) &= \Pr(Z \geq X + 1) \\ &\leq \exp\left(-\frac{r^2\eta^2}{2(r+1)}\right) \\ &\leq \exp\left(-\frac{(r-1)\eta^2}{2}\right) \\ &\leq 1 - \alpha. \end{aligned}$$

Thus $q_{ij}(D) \geq \alpha$, and the direction is admitted.

If $\theta_{ij} \leq t_0 - \eta$, then $X/r \leq t_0 - \eta/2$ and

$$(r+1)t_0 - X \geq \frac{r\eta}{2} + t_0 \geq \frac{r\eta}{2}.$$

The lower-tail Hoeffding bound and $r \geq r_{\min}(\eta, \delta)$ give

$$\begin{aligned} q_{ij}(D) &= \Pr(Z \leq X) \\ &\leq \exp\left(-\frac{(r-1)\eta^2}{2}\right) \\ &\leq 1 - \alpha < \alpha. \end{aligned}$$

Thus the direction is withheld. These cases exhaust all directions, giving admitted set $\mathcal{P}_\sigma$ on $\mathcal{E}$ and $S_{\sigma,\alpha} = S_\sigma$ with probability at least $1 - \delta$. □

B.3 False-display guarantees

Corollary 8 (Pairwise false-display bound)**.** *Assume Assump. 1. Let ties count as non-support. Fix $\sigma \in (\frac{1}{2}, 1]$ and $\alpha \in (\frac{1}{2}, 1)$. For an ordered pair (i, j), suppose $\theta_{ij} \leq \sigma - \eta$ for some $\eta \in (0, \sigma]$. For every assignment R with $n_{ij} = r$, if*

$$r \geq \left\lceil 1 + \frac{2}{\eta^2} \log \frac{1}{1-\alpha} \right\rceil,$$

then

$$\Pr(S_{\sigma,\alpha} \textit{ displays } i \succ j \mid R) \leq \exp\left(-\frac{r\eta^2}{2}\right).$$

Proof of Cor. 8. By Cor. 3, displaying $i \succ j$ requires $q_{ij}(D) \geq \alpha$. Fix an assignment R with $n_{ij} = r$, and define t_0 and $X = s_{ij}$ as in the proof of Thm. 7(b). If $X/r \leq t_0 - \eta/2$, the lower-tail bound from that proof and the assumed bound on r give

$$q_{ij}(D) \leq \exp\left(-\frac{(r-1)\eta^2}{2}\right) \leq 1 - \alpha < \alpha.$$

Hence displaying $i \succ j$ requires $X/r > t_0 - \eta/2 \geq \theta_{ij} + \eta/2$, an event of probability at most $\exp(-r\eta^2/2)$ by Hoeffding's inequality. □

Proof of Thm. 7(a). Fix R satisfying the pair-count condition. Let $\mathcal{U}_\sigma = \{(i, j) : \theta_{ij} < \sigma\}$ be the unsupported directions and $M = n(n-1)$. The margin assumption, $n_{ij} \geq r_{\min}(\eta, \delta)$, and Cor. 8 give

$$\Pr(S_{\sigma,\alpha} \text{ displays } i \succ j \mid R) \leq \exp\left(-\frac{n_{ij}\eta^2}{2}\right) \leq \frac{\delta}{2M}$$

for every $(i, j) \in \mathcal{U}_\sigma$. A union bound over the at most M directions in $\mathcal{U}_\sigma$ shows that, given R, the output displays no unsupported direction with probability at least $1 - \delta/2 \geq 1 - \delta$. On this event, $\mathcal{D}(S_{\sigma,\alpha}) \subseteq \mathcal{P}_\sigma$, so $S_{\sigma,\alpha}$ is feasible for $\mathcal{P}_\sigma$. Since S_σ displays every comparison of any such tiering (Thm. 2), $\mathcal{D}(S_{\sigma,\alpha}) \subseteq \mathcal{D}(S_\sigma)$. □

B.4 Recovery under randomized sparse assignment

Under uniform assignment of m of the $N = \binom{n}{2}$ pairs per user, Cor. 9 gives the sufficient user-count rate

$$p = O\left(\frac{N}{m} \cdot \frac{1}{\eta^2} \log \frac{n}{\delta(1-\alpha)}\right).$$

The factor N/m accounts for sparse exposure. The dependence on the support margin is quadratic in $1/\eta$.

Corollary 9 (Finite-user recovery)**.** *For any fixed n, $1 \le m \le N = \binom{n}{2}$, $\sigma \in (\frac{1}{2}, 1)$, $\alpha \in (\frac{1}{2}, 1)$, and $\delta \in (0, 1)$, suppose users are sampled i.i.d. from the target population, each receives a uniform set of m pairs drawn independently across users and of preferences, and, for some $0 < \eta < 1 - \sigma$, population support stays at least η away from σ. Let D_p contain the observations from the p sampled users. If*

$$\frac{pm}{N} \ge \max\left\{2r_{\min}\left(\eta, \frac{\delta}{2}\right), 8 \log \frac{2N}{\delta}\right\},$$

then

$$\Pr(S_{\sigma,\alpha}(D_p) = S_\sigma) \ge 1 - \delta.$$

Proof of Cor. 9. Let

$$r_\delta = r_{\min}\left(\eta, \frac{\delta}{2}\right), \qquad \mu = \frac{pm}{N}.$$

Each unordered pair has $n_{ij} \sim \mathrm{Binomial}(p, m/N)$ marginally, so the multiplicative Chernoff bound and the hypotheses $\mu \ge 2r_\delta$ and $\mu \ge 8\log(2N/\delta)$ give

$$\Pr(n_{ij} < r_\delta) \le \Pr(n_{ij} \le \mu/2) \le \exp(-\mu/8) \le \frac{\delta}{2N}.$$

Let $\mathcal{E} = \{n_{ij} \ge r_\delta \text{ for all } i < j\}$. A union bound over the N pairs gives $\Pr(\mathcal{E}^c) \le \delta/2$, without requiring independent pair counts. For every assignment $R \in \mathcal{E}$, Thm. 7(b) with $\delta/2$ in place of δ gives

$$\Pr(S_{\sigma,\alpha}(D_p) \ne S_\sigma \mid R) \le \frac{\delta}{2}.$$

Averaging over R and adding $\Pr(\mathcal{E}^c)$ gives

$$\Pr(S_{\sigma,\alpha}(D_p) \ne S_\sigma) \le \Pr(\mathcal{E}^c) + \frac{\delta}{2} \le \delta.$$

□

B.5 Bounded preference-dependent recording

We guard against recording bias up to Γ by running *SE-RO* at an adjusted support floor (Prop. 11). Fix D, $\sigma \in (\frac{1}{2}, 1)$, and $\alpha \in (\frac{1}{2}, 1)$. Write $\mathrm{odds}(x) = x/(1-x)$ on $[0,1]$ with $\mathrm{odds}(1) = \infty$. For an ordered pair (i,j) and population user k, let $Y_{ij}^k = \mathbf{1}\{\pi_{ij}^k = +1\}$ indicate strict support and R_{ij}^k indicate recording. Define aggregate recording probabilities, their ratio, and recorded support by

$$e_y := \Pr\big(R_{ij}^k = 1 \mid Y_{ij}^k = y\big), \qquad \rho_{ij} := \frac{e_1}{e_0}, \qquad \tilde{\theta}_{ij} := \Pr\big(Y_{ij}^k = 1 \mid R_{ij}^k = 1\big),$$

where each probability averages over users and over every other label, so ρ_{ji} is a separate aggregate. If $\theta_{ij} \in \{0, 1\}$, one class is empty. We then set $\rho_{ij} = 1$, and $\tilde{\theta}_{ij} = \theta_{ij}$. Bayes' rule gives $\mathrm{odds}(\tilde{\theta}_{ij}) = \rho_{ij}\,\mathrm{odds}(\theta_{ij})$. Under Assump. 1, $\rho_{ij} = 1$ and $\tilde{\theta}_{ij} = \theta_{ij}$.

Assumption 10 (Bounded pair-level selection)**.** *Fix $\Gamma \ge 1$. For every ordered pair (i,j), $0 < \rho_{ij} \le \Gamma$, and, conditional on the set $\{k : R_{ij}^k = 1\}$ of users whose comparison of the pair is recorded, $s_{ij} \mid \{k : R_{ij}^k = 1\}, \tilde{\theta}_{ij} \sim \mathrm{Binomial}(n_{ij}, \tilde{\theta}_{ij})$. We do not assume independence across pairs.*

The $\mathrm{Beta}(1,1)$ prior and update in Section 2 now give $\tilde{\theta}_{ij} \mid D_{ij} \sim \mathrm{Beta}(1+s_{ij}, 1+\ell_{ij})$. This posterior conditions on the directional counts (s_{ij}, n_{ij}) with the recorded set fixed, and every probability given D_{ij} below means this posterior. For a declared bound Γ, the adjusted threshold and admission rule are

$$\sigma_\Gamma := \frac{\Gamma\sigma}{1-\sigma+\Gamma\sigma}, \qquad \text{admit } i \succ j \text{ iff } \Pr\big(\tilde{\theta}_{ij} \geq \sigma_\Gamma \mid D_{ij}\big) \geq \alpha,$$

so that $\mathrm{odds}(\sigma_\Gamma) = \Gamma\,\mathrm{odds}(\sigma)$. Let $\mathcal{P}_\Gamma$ be the admitted directions and S_Γ their tiering from Algorithm 1. Let $L_{ij,\alpha}$ be a lower bound on recorded support with posterior mass α above it. The tolerance is

$$\Gamma_{ij} = (1-\sigma) \cdot \frac{L_{ij,\alpha}}{\sigma(1-L_{ij,\alpha})}. \tag{1}$$

Proposition 11 (Adjusted pairwise gate)**.** *Fix D, $\sigma \in (\frac{1}{2}, 1)$, $\alpha \in (\frac{1}{2}, 1)$, and $\Gamma \geq 1$.*

(a) $\sigma_\Gamma \in [\sigma, 1)$, and the adjusted gate is the admission rule of Algorithm 1 at support floor σ_Γ, so $S_\Gamma = S_{\sigma_\Gamma,\alpha}$. For every ordered pair, $(i,j) \in \mathcal{P}_\Gamma$ if and only if $\Gamma \leq \Gamma_{ij}$. Hence $\mathcal{P}_1$ is the admitted set of $S_{\sigma,\alpha}$, $\mathcal{P}_{\Gamma'} \subseteq \mathcal{P}_\Gamma$ whenever $\Gamma \leq \Gamma'$, and $\mathcal{P}_\Gamma$ contains at most one direction per unordered pair.

(b) Under Assump. 10 with bound Γ, for every $(i,j) \in \mathcal{P}_\Gamma$ and every value $\rho_{ij} \in (0, \Gamma]$, $\Pr(\theta_{ij} \geq \sigma \mid D_{ij}) \geq \alpha$, where $\theta_{ij} = \tilde{\theta}_{ij}/[\tilde{\theta}_{ij} + \rho_{ij}(1-\tilde{\theta}_{ij})]$ carries the posterior of $\tilde{\theta}_{ij}$.

Proof of Prop. 11. (a) Since $\Gamma \geq 1$, $\sigma \leq \sigma_\Gamma < 1$. The adjusted gate is the rule of Algorithm 1 at σ_Γ, so $S_\Gamma = S_{\sigma_\Gamma,\alpha}$. The Beta posterior has a positive continuous density on $(0,1)$, so its tail at t is at least α if and only if $t \leq L_{ij,\alpha}$. Therefore

$$(i,j) \in \mathcal{P}_\Gamma \quad \Longleftrightarrow \quad \sigma_\Gamma \leq L_{ij,\alpha} \quad \Longleftrightarrow \quad \Gamma \leq \Gamma_{ij}.$$

At $\Gamma = 1$ the threshold is σ, and increasing Γ can only remove directions. Lem. 4 applies at σ_Γ.

(b) For fixed $\rho_{ij} \in (0, \Gamma]$, the inverse selection map gives $\mathrm{odds}(\theta_{ij}) = \mathrm{odds}(\tilde{\theta}_{ij})/\rho_{ij}$. On $\tilde{\theta}_{ij} \geq \sigma_\Gamma$,

$$\mathrm{odds}(\theta_{ij}) \geq \frac{\Gamma}{\rho_{ij}}\,\mathrm{odds}(\sigma) \geq \mathrm{odds}(\sigma).$$

Thus $\{\tilde{\theta}_{ij} \geq \sigma_\Gamma\} \subseteq \{\theta_{ij} \geq \sigma\}$, giving the claimed posterior bound on $\mathcal{P}_\Gamma$. □

What the bound assumes. The case $\Gamma = 1$ contains Assump. 1 and also allows supporters to be recorded less often than nonsupporters, so $\tilde{\theta}_{ij} \leq \theta_{ij}$ and Cor. 3 holds for every $\rho_{ij} \leq 1$. The bound must hold for the aggregate ratio ρ_{ij}. Groups with $\rho_{ij} = 1$ each can pool to $\rho_{ij} \approx 4.6$, so a user-level bound needs an aggregation argument.

B.6 Unrestricted preference-dependent recording

Theorem 12 (MNAR non-identification)**.** *Under unrestricted preference-dependent recording, two populations can induce the same distribution of observed counts (s_{ij}, n_{ij}) while their population selective rankings display opposite directions. Hence no rule based only on these counts identifies the population direction without an observation-model restriction.*

Proof of Thm. 12. Fix $\sigma \in (1/2, 1)$ and take $n = 2$ items without ties. Users are sampled independently from either process below. Let $Y = 1$ indicate strict support for i over j and $O = 1$ indicate

recording. Choose $a, b \in (0, 1-\sigma)$ and the processes

$$\begin{aligned}
\mathsf{P}_+ : \quad & \Pr(Y=1) = \sigma, \\
& \Pr(O=1 \mid Y=1) = \frac{a}{\sigma}, \\
& \Pr(O=1 \mid Y=0) = \frac{b}{1-\sigma}; \\
\mathsf{P}_- : \quad & \Pr(Y=1) = 1-\sigma, \\
& \Pr(O=1 \mid Y=1) = \frac{a}{1-\sigma}, \\
& \Pr(O=1 \mid Y=0) = \frac{b}{\sigma}.
\end{aligned}$$

Both processes give observed support, observed non-support, and missing labels with probabilities a, b, and $1-a-b$. Thus (s_{ij}, n_{ij}) has the same distribution at every sample size.

The first process has $\theta_{ij} = \sigma$, so the inclusive target displays i over j. The second has $\theta_{ij} = 1-\sigma$ and $\theta_{ji} = \sigma$, so the target displays the opposite direction. For $\sigma = 1$, the same conclusion follows by comparing populations with strict support rates one and zero when every preference is missing. □

B.7 Fixed assessment of displayed comparisons

We compare rankings built at different settings using common population support and evidence requirements (σ_0, α_0), data, and prior. For $\mathcal{D}(S) \neq \varnothing$, Algorithm 2 run at (σ_0, α_0) returns

$$\Gamma_{\mathrm{rank}}(S) = \min_{(i,j)\in\mathcal{D}(S)} (1-\sigma_0) \cdot \frac{L_{ij,\alpha_0}}{\sigma_0(1-L_{ij,\alpha_0})}. \tag{2}$$

Proposition 13 (Fixed assessment of a tiered ranking)**.** *Fix finite observed data D, $\sigma_0 \in (\frac{1}{2}, 1)$, and $\alpha_0 \in (\frac{1}{2}, 1)$. For $\Gamma \geq 1$, let*

$$\sigma_{0,\Gamma} = \frac{\Gamma\sigma_0}{1-\sigma_0+\Gamma\sigma_0}, \qquad \mathcal{P}^0_\Gamma = \left\{(i,j) : i \neq j,\ \Pr(\tilde{\theta}_{ij} \geq \sigma_{0,\Gamma} \mid D_{ij}) \geq \alpha_0\right\},$$

where $\tilde{\theta}_{ij} \mid D_{ij} \sim \mathrm{Beta}(1+s_{ij}, 1+n_{ij}-s_{ij})$.

(a) For every tiered ranking S with $\mathcal{D}(S) \neq \varnothing$,

$$\Gamma \leq \Gamma_{\mathrm{rank}}(S) \quad \Longleftrightarrow \quad \mathcal{D}(S) \subseteq \mathcal{P}^0_\Gamma.$$

(b) The conditions in part (a) hold if and only if, for each $(i,j) \in \mathcal{D}(S)$ and every fixed $r \in (0, \Gamma]$,

$$\Pr\left(\frac{\tilde{\theta}_{ij}}{\tilde{\theta}_{ij} + r(1-\tilde{\theta}_{ij})} \geq \sigma_0 \,\middle|\, D_{ij}\right) \geq \alpha_0.$$

Under Assump. 10, this includes the actual recording ratio $r = \rho_{ij}$ of each displayed comparison.

(c) If $\varnothing \neq \mathcal{D}(S') \subseteq \mathcal{D}(S)$, then $\Gamma_{\mathrm{rank}}(S') \geq \Gamma_{\mathrm{rank}}(S)$. Identical displayed sets have identical tolerances. The one-tier ranking satisfies the inclusion in (a) for every Γ, but its tolerance is undefined.

Proof of Prop. 13. By Prop. 11(a), $(i,j) \in \mathcal{P}^0_\Gamma$ if and only if $\Gamma \leq \mathrm{odds}(L_{ij,\alpha_0})/\mathrm{odds}(\sigma_0)$. Minimizing over $\mathcal{D}(S)$ proves (a), including equality.

For (b), $g_r(x) = x/[x + r(1-x)]$ decreases in $r > 0$ on $[0,1]$, and $g_\Gamma(x) \geq \sigma_0$ if and only if $x \geq \sigma_{0,\Gamma}$. The adjusted gate therefore gives the posterior bound for every $r \leq \Gamma$; choosing $r = \Gamma$ gives the converse. Each probability uses D_{ij}, and under Assump. 10, $g_{\rho_{ij}}(\tilde{\theta}_{ij}) = \theta_{ij}$.

Part (c) follows because restricting a nonempty finite minimum cannot decrease it. □

Choosing construction settings. By Prop. 6 and part (c), raising either construction parameter with the other fixed cannot lower Γ_{rank} while comparisons remain displayed.

Data-dependent rankings. The equivalences hold for every fixed D and tiered S, including rankings chosen after seeing the data. If $\Gamma_{\mathrm{rank}}(S) < 1$, some displayed comparison fails assessment at $\Gamma = 1$. Part (b) gives pairwise posterior statements based on D_{ij}. A joint probability requires a common joint posterior.

C Additional Results for Section 4

C.1 Recording tolerance of fixed rankings

Table 4 reports Γ_{rank} for rankings constructed and assessed at $\sigma = \sigma_0 = 0.60$ and $\alpha = \alpha_0 = 0.90$ with a $\mathrm{Beta}(1, 1)$ prior. None of our datasets supplies a known bound on recording bias. The tolerances are sensitivity values under the $\mathrm{Beta}(1, 1)$ prior, not bounds on recording bias.

Table 4: Nonempty rankings tolerate some recording imbalance. Entries give median Γ_{rank} among nonempty repeated-panel outputs. Counts give nonempty/total outputs. A dash means all outputs withhold every comparison.

Dataset	Panel	Nonempty / total	Tolerance
`Acyclic`	Repeated panels	199/200	1.381
`Cyclical`	Repeated panels	200/200	1.416
`Ties`	Repeated panels	170/200	1.129
`Two Groups`	Repeated panels	200/200	2.485
`EnviroSound`	Half panels	153/200	1.106
`Sushi`	Half panels	200/200	1.201
`Jester`	Half panels	0/200	—
`ERS Set 77`	Half panels	200/200	1.171

C.2 Disagreement, coverage, and stability

Table 5: Individual disagreement and coverage of *SE-RO* at $\alpha = 0.90$ (first two column pairs) and $\sigma = \alpha = 0.5^{+}$ (last pair). Values are percentages. Real-data disagreement uses the opposite half.

	$\sigma = 0.60$		$\sigma = 0.5^{+}$		$\sigma = \alpha = 0.5^{+}$	
Dataset	**Disagr.**	**Cov.**	**Disagr.**	**Cov.**	**Disagr.**	**Cov.**
`EnviroSound`	3.5	34.3	11.2	78.3	18.9	97.7
`Sushi`	8.8	37.8	29.7	90.8	34.1	99.7
`Jester`	0.0	0.0	28.2	67.8	39.6	92.1
`ERS Set 77`*	5.2	17.2	22.6	72.2	30.6	88.6
`Acyclic`	10.9	73.0	11.3	73.9	18.7	88.7
`Cyclical`	11.0	73.3	11.2	73.6	16.4	84.0
`Ties`	11.5	57.4	14.7	73.2	15.7	75.3
`Two Groups`	1.0	87.6	3.0	93.1	5.6	98.8

*Incomplete ERS ballots use all observed evaluation-half votes.

Stability across panels. Few reversals need not mean stable displays: at $\alpha = 0.90$, *SE-RO* has none on `EnviroSound` but 31.2% relation change (Table 7). Reversals count opposite orders over all unordered pairs. Relation change also counts display and withhold changes.

C.3 Adjusted gate under simulated recording bias

Setting $\Gamma = 2$ in the adjusted gate of Section B.5 removes the below-floor displays that recording bias causes on `Two Groups`, but withholds some supported directions. In the recording-bias experiment of Section 4.2, each user's preference on each pair is recorded independently. Supporters are $\rho = 2$ times as likely as nonsupporters to be recorded, and each user records 12 of the 45 pairs on average. *SE-RO* with the unadjusted gate ($\Gamma = 1$) and $\alpha = 0.90$ displays an average of 57.3% of the three 55%-support directions at 250 users and 99.3% at 1,000 users. With $\Gamma = \rho = 2$, it displays 0% at every tested user count. With 1,000 users, it keeps every unanimous and 90%-support direction but only 41.7% of the 65%-support directions, for 89.4% coverage overall.

Table 6: Baselines display more pairs and disagree more than *SE-RO* at $\sigma = 0.60$, $\alpha = 0.90$. Values are percentages. Real-data disagreement uses the opposite half.

	MB–K		Borda		Copeland		Rank Centrality	
Dataset	**Disagr.**	**Cov.**	**Disagr.**	**Cov.**	**Disagr.**	**Cov.**	**Disagr.**	**Cov.**
EnviroSound	15.6	89.9	19.9	99.8	19.3	98.6	20.3	100.0
Sushi	34.2	99.8	34.2	100.0	34.1	99.8	34.2	100.0
Jester	—	—	43.6	100.0	41.7	96.3	43.6	100.0
ERS Set 77*	33.8	94.7	36.5	100.0	34.7	96.6	36.6	100.0
Acyclic	11.3	73.8	24.1	99.5	20.9	93.2	24.3	100.0
Cyclical	11.2	73.8	24.2	99.5	18.6	88.5	24.3	100.0
Ties	—	—	29.2	99.8	25.4	92.8	29.3	100.0
Two Groups	2.4	91.6	6.4	99.7	5.7	99.0	6.2	100.0

—: MB–K excludes recorded ties. *As in Table 5.

Table 7: *SE-RO* rarely reverses pairs across panels. Percentages average 100 panel pairs at $\sigma = 0.60$ and $\alpha = 0.90$. Jester displays nothing at this setting.

Dataset	**Reversal**	**Relation change**
Acyclic	0.0	0.7
Cyclical	0.0	0.0
Ties	0.0	23.5
Two Groups	0.0	1.6
EnviroSound	0.0	31.2
Sushi	0.0	0.0
Jester	0.0	0.0
ERS Set 77	0.0	1.1

C.4 Populations, datasets, and baselines

Synthetic populations. Users are independent population draws. Acyclic, Cyclical, and Ties have ordered blocks $B_1 = \{1, 2, 3\}$, $B_2 = \{4, 5, 6\}$, and $B_3 = \{7, 8, 9, 10\}$, with judgments independent across pairs. In Acyclic, earlier-block items beat later-block items with probability 0.85 and lose with probability 0.15. Within blocks, each direction has probability 0.50, with no ties. Cyclical changes only $1 \succ 2$, $2 \succ 3$, and $3 \succ 1$ to support 0.85, with 0.15 for each reverse. In Ties, cross-block win, loss, and tie probabilities are 0.80, 0.10, and 0.10. Within blocks, each strict direction has probability 0.45 and ties 0.10.

Two Groups draws one of eight complete rankings, placing $\{1, 2, 3\}$ above $\{4, 5, 6\}$, then $\{7, 8, 9\}$, then item 10. We independently choose within-block orders $1 \succ 2 \succ 3$, $4 \succ 5 \succ 6$, and $7 \succ 8 \succ 9$ with probabilities 0.90, 0.65, and 0.55, reversing them otherwise. This gives 36 unanimous pairs and three pairs each with support 90%/10%, 65%/35%, and 55%/45%. The two groups differ only on items 7–9. At the 60% floor, the population ranking displays 42 of 45 relations, withholding these three pairs.

We assign each user the first m pairs of a uniform random permutation of all 45 pairs, independent across users and of preferences. The nested 12-pair and 45-pair settings share users, outcomes, and permutations.

Released datasets. For each seed, we randomly split respondents into exhaustive, disjoint, near-equal halves (Table 8). Methods use all observed labels in one half and reference support from the other, then swap halves. Ties count as non-support. EnviroSound has complete strict labels. Sushi rankings yield 45 preferences each, linked to participant information by ID. For Jester, we use the first 5,000 users and ten gauge jokes of Dataset 1 from the converted mldata mirror with equal ratings as ties, yielding 224,910 observations. The target is this mirror's gauge-user data, not

the official Berkeley release. `ERS Set 77` uses Set 77 from the official PrefLib release and labels only jointly ranked alternatives, yielding 110,320 observations.

Table 8: Exposure varies most on `ERS Set 77`. **Exposure** is the minimum/maximum fraction of respondents labeling a pair. **Conflict** is the pooled less-frequent-direction share among strict labels.

Dataset	Domain	n	p	Exposure	Conflict	Preference evidence
`EnviroSound` [38]	Auditory perception	12	74	1.000/1.000	0.19	Complete pairwise judgments
`Sushi` [14]	Food preferences	10	5,000	1.000/1.000	0.34	Complete rankings
`Jester` [10]	Joke ratings	10	5,000	0.999/1.000	0.43	Rating-derived pair labels; 1.7% ties
`ERS Set 77` [18]	Elections	12	3,419	0.409/0.654	0.36	Strict incomplete ballots

Settings and intervals. We reuse panels and halves across $\alpha \in \{0.5^+, 0.70, 0.90, 0.95\}$. Pointwise 95% bootstrap intervals use 1,000 resamples of 100 seeds, preserving paired outputs and nested exposures.

Baselines. Baselines use the same labels as *SE-RO* (Section 4.1). Missing labels contribute nothing. Rank Centrality forms a chain on the observed pair graph. Off-diagonal transitions equal the observed share preferring the destination, counting ties as half, divided by the maximum graph degree. Diagonals supply the remaining mass. We mix in uniform transitions with weight 10^{-6}, start uniformly, and iterate until the largest coordinate change is at most 10^{-10} or we reach 500 steps, normalizing scores to mean one. Borda, Copeland, and Rank Centrality order scores decreasingly, grouping matches to each tier's first score within relative tolerance 10^{-10} or absolute tolerance 10^{-12}. They display all between-tier comparisons. MB–K uses `partial-rankings` with solver tolerance 10^{-6}, greedily merging neighboring-strength groups from singletons to one group. Equal-strength groups remain incomparable. Binary winner–loser counts are required, excluding `Ties` and `Jester`.

C.5 Admission, false displays, and recovery

Tier construction hides a larger share of all pairs on the released datasets than on the synthetic populations (Table 9). On `Cyclical`, every output recovers the population tiers at both $\alpha = 0.70$ and $\alpha = 0.90$, with the cycle items sharing the top tier.

Table 9: Tier construction withholds admitted pairs (**Hidden**). Percentages use all pairs at $\sigma = 0.60$, averaging two panels or halves per seed over 100 seeds. Synthetic panels use $p = 250$, $m = 12$.

	$\alpha = 0.70$			$\alpha = 0.90$		
Dataset	**Admitted**	**Hidden**	**Displayed**	**Admitted**	**Hidden**	**Displayed**
`Acyclic`	74.2	0.8	73.4	73.4	0.5	73.0
`Cyclical`	80.5	7.2	73.3	80.0	6.7	73.3
`Ties`	73.3	1.7	71.6	72.4	14.9	57.4
`Two Groups`	91.2	1.7	89.4	88.7	1.1	87.6
`EnviroSound`	82.4	9.9	72.5	71.2	37.0	34.3
`Sushi`	60.0	22.1	37.9	58.8	21.0	37.8
`Jester`	19.9	19.9	0.0	16.4	16.4	0.0
`ERS Set 77`	59.7	38.9	20.8	53.6	36.4	17.2

Below-floor displays. At $\alpha = 0.90$, no display in the four fixed-setting synthetic experiments falls below 60% population support. Below 60% held-out support are 29 of 4,525 `EnviroSound` displays (Table 10), 7 of 2,270 on `ERS Set 77`, and none on `Sushi`. `Jester` displays nothing. In

Table 10: Stronger evidence removes synthetic below-floor displays in these runs. Counts at $\sigma = 0.60$ sum over both outputs of 100 seeds. `EnviroSound` uses opposite-half reference support. For synthetic 0/100 entries, the exact one-sided 95% upper bound on the per-seed rate is 2.95%.

α	Displayed	Below floor	Seeds with any
`Acyclic`			
0.70	6,606	6	3/100
0.90	6,567	0	0/100
`Cyclical`			
0.70	6,600	0	0/100
0.90	6,600	0	0/100
`Ties`			
0.70	6,447	0	0/100
0.90	5,169	0	0/100
`Two Groups`			
0.70	8,050	6	3/100
0.90	7,883	0	0/100
`EnviroSound`			
0.70	9,573	313	98/100
0.90	4,525	29	22/100

the recovery study below, no `Acyclic` output or sparse ($m = 12$) `Two Groups` output displayed a below-floor direction. For dense ($m = 45$) `Two Groups`, one of 100 seeds did at $p = 125$ and one at $p = 250$, with none at $p = 500$.

Recovery with more users and pairs. Exact tier recovery improves with more users and pairs (Table 11). Supported-direction recall can be high before exact recovery. We draw fresh `Acyclic` and `Two Groups` panels with $p = 125, 250, 500$, nesting $m = 12$ within $m = 45$ assignments under the main study's population and assignment laws.

Table 11: Recovery at $\sigma = 0.60$, $\alpha = 0.90$, averaging two independent panels per seed over 100 seeds. Recall measures population-supported directions. Pair counts average within-panel fifth percentiles. Brackets give 95% seed-cluster intervals for exact recovery, omitting point intervals. Population targets display 33 `Acyclic` relations and 42 `Two Groups` relations.

	$m = 12$ **pairs per user**				$m = 45$ **pairs per user**			
Users p	**Exact tiers (%)**	**Recall (%)**	**Displays**	**Pair count (5th pct.)**	**Exact tiers (%)**	**Recall (%)**	**Displays**	**Pair count (5th pct.)**
Population: `Acyclic`								
125	45.5 [38.5, 53.0]	60.8	20.1	25.7	100.0	100.0	33.0	125.0
250	99.0 [97.5, 100.0]	99.3	32.8	56.0	100.0	100.0	33.0	250.0
500	100.0	100.0	33.0	117.9	100.0	100.0	33.0	500.0
Population: `Two Groups`								
125	3.0 [1.0, 5.5]	93.4	39.2	25.7	38.5 [31.5, 45.5]	95.6	40.2	125.0
250	7.0 [4.0, 10.5]	93.9	39.4	56.1	58.5 [51.5, 65.0]	97.1	40.8	250.0
500	14.5 [10.0, 19.5]	95.0	39.9	118.2	87.5 [83.0, 91.5]	99.1	41.6	500.0

C.6 COVERAGE-MATCHED COMPARISON

Matching coverage on one panel need not preserve disagreement on larger panels.

Setup. On the 250-user `Two Groups` panels with 12 pairs per user (Table 5), *SE-RO* at $\sigma = 0.60$ and $\alpha = 0.90$ displays 87.6% of pairs. We fix each method at the grid setting closest to 87.6% coverage. We then draw 2,000-user panels from the same population with the same seeds and pair assignment. We measure displays of the three 55%/45% pairs in either direction and recall of the 42 population comparisons. These 12-pair assignments explain why the MB–K rates differ from Fig. 5.

Methods. All admitted directions pass through the graph stage of Algorithm 1. One-sided *exact binomial tests* of $\theta_{ij} \leq \sigma$ or $\theta_{ij} \leq \frac{1}{2}$ admit directions at the stated Holm-adjusted level across $n(n-1)$ directions [11]. MB–K has no setting and uses its native output.

Table 12: Direction-only methods display more contested pairs as panels grow. We report contested displays and recall (%) on `Two Groups`, with 12 pairs per user and $\sigma = 0.60$. Settings approximately match coverage at $p = 250$, then stay fixed. Entries average two panels per seed over 100 seeds.

	Coverage	**Contested displayed**		**Recall**	
Method and setting	$p = 250$	250	2,000	250	2,000
SE-RO, $\alpha = 0.90$	87.6	0.0	0.0	93.8	98.4
Exact test vs. σ, Holm 0.50	86.7	0.0	0.0	92.9	95.1
Exact test vs. $\frac{1}{2}$, Holm 0.10	87.5	0.0	11.7	93.7	100.0
MB–K	91.6	1.7	28.3	98.0	100.0

Results. The exact test against $\frac{1}{2}$ and MB–K increasingly order the contested pairs as panels grow (Table 12). The exact test against $\frac{1}{2}$ starts within 0.1 points of the coverage of *SE-RO* with no contested pair. At 2,000 users, contested displays reach 11.7%. MB–K rises from 1.7% to 28.3% on these sparse panels. Coverage matching is approximate: MB–K has no setting and starts 4.0 points above *SE-RO*.

SE-RO withholds contested pairs at 250 and 2,000 users while recall rises from 93.8% to 98.4% as its required share approaches σ. At 125 users, it displays contested pairs in 0.3% of cases. The exact test against σ also withholds them, but Holm's adjustment across 90 directions limits recall to 95.1%.

C.7 MINIMUM EVIDENCE AND FAILURE FRONTIERS

Setup. Ten items have population tiers of sizes $3/3/4$, displaying 33 relations at $\sigma = 0.70$. Cross-tier support is $0.70 + \eta$, with $\eta \in \{0.02, 0.05, 0.10, 0.20\}$. Within-tier support is 0.50 either way. From each pair's outcome law, we draw n_{pair} observations and use the $\text{Beta}(1, 1)$ posterior-tail rule at confirmatory $\alpha = 0.95$ and diagnostic $\alpha = 0.5^{+}$.

Structural usefulness requires at least two tiers, precision ≥ 0.95, recall ≥ 0.50 against the population ranking, and largest tier $< 80\%$ of items. Exact recovery means matching ordered population tiers. For either outcome, n_{80} is the smallest evaluated n_{pair} with success-probability lower bound ≥ 0.80. Selected positive cells have 400 independent runs. One-sided Wilson bounds [35] are Bonferroni-adjusted across tested sample sizes within each population, method, threshold, and outcome, for nominal simultaneous 95% confidence. These grid frontiers are not exact thresholds.

Smaller margins require more evidence. Usefulness at $\alpha = 0.95$ needs two orders of magnitude more observations per pair as η falls from 0.20 to 0.02 (Table 13). Exact recovery needs 11,200 at the smallest margin. At $\eta = 0.10$, 331 users who each compare all 45 pairs give every pair 331 observations. With one pair per user, 17,411 users give each of the 33 required relations at least 331 observations with probability at least 0.95. This bounds allocation, not admission. The gap reflects the N/m factor in the user-count rate of Section B.4.

Table 13: Smaller margins need more evidence for tiers of sizes 3, 3, and 4, at $\sigma = 0.70$, $\alpha = 0.95$. Pairwise n_{90} is the smallest count with admission probability $\geq 90\%$ at that and every larger count through 11,200. Graph frontiers use simultaneous one-sided 95% lower bounds.

Margin η	**Given pair** n_{90}	**Structural** n_{80}	**Exact tiers** n_{80}
0.02	4,476	8,960	11,200
0.05	708	1,403	1,750
0.10	170	331	331
0.20	36	69	69

The next three checks each relax one assumption behind the recovery results.

Zero-margin boundary. Thm. 7(b) excludes zero margins. The zero-margin population has win, loss, and tie probabilities $(0.70, 0.20, 0.10)$ across tiers and $(0.45, 0.45, 0.10)$ within. Its target includes relations exactly at σ. A one-tier binary control has support 0.50 either way. At $n_{\text{pair}} = 1$, $\alpha = 0.5^+$ produces at least two tiers in 11 of 100 zero-margin pilot runs (one-sided 95% Wilson upper bound 0.172). Exact recovery fails in most pilot runs. The one-tier control splits incorrectly once in 100 runs (upper bound 0.044 by the same method).

Uneven exposure increases full abstention. Cor. 9 assumes uniform pair assignment. At $\eta = 0.10$ and $\alpha = 0.5^+$, 126 users compare 15 pairs each. Five items chosen per seed get weight four, others one. Pair weights are products of item weights. Each 15-pair subset's probability is proportional to the product of its pair weights. Full abstention is the fraction of one-tier rankings. Against uniform assignment, full abstention rises 17.8 percentage points (paired 95% normal-approximation interval 12.0–23.5, 400 seeds).

Prior sensitivity. Our results use the uniform prior (Section B.1). Against $\text{Beta}(1, 1)$, we test Beta priors of strength 10 centered at θ (oracle) or $2\sigma - \theta$, across 90 combinations: support $\{0.60, 0.68, 0.70, 0.72, 0.80\}$, sample size $\{0, 1, 2, 5, 10, 20, 40, 80, 160\}$, and $\alpha \in \{0.5^+, 0.95\}$. Each frontier is the smallest tested sample size with admission probability $\geq 80\%$ at that and every larger tested size. The oracle lowers a frontier only far from the floor, at $\theta = 0.80$ and $\alpha = 0.5^+$, from 20 observations to zero. A misleading prior can double false admission. At $\theta = 0.60$, $\alpha = 0.5^+$, and five observations, the reflected prior raises false admission to 0.683 versus 0.337 under $\text{Beta}(1, 1)$.

D ADDITIONAL RESULTS FOR SECTION 5

Measurement. Held-out voters disagree on a pair when its margin $m_{ij} = |n_i - n_j|/(n_i + n_j)$ is at most 0.20 and agree when it is at least 0.60. Here n_i counts held-out votes for i. These comparisons use pairs with at least four held-out votes. A withholding rate is the share of such pairs that a method leaves unordered (Table 14). Disagreement is 100% minus agreement.

Aggregation rules. Every rule other than BT passes the directions it admits through the same graph stage (Algorithm 1). Vote Share admits a direction when its share of the pair's votes reaches a threshold. BT fits a score u_i to each response and a tie parameter λ to each prompt's votes. A direction's logit is the score of its winner, and a tie's logit is $\lambda + (u_i + u_j)/2$. This is the tie model of Davidson [7]. The fit orders every pair except exact score ties. BT-Margin admits a direction only when the fitted score gap reaches a threshold. BT bootstrap refits the scores on 200 resamples of the prompt's evaluators. It admits a direction that wins in at least a share γ of resamples. Each threshold makes its rule display about as many calibration pairs as *SE-RO*.

Table 14: We report every rule at $\sigma = 0.60$ and $\alpha = 0.70$. The last two columns give pairs withheld where voters disagree and where they agree.

Method	Cov.	Agr.	Disagree	Agree
SE-RO	44.1	81.4	60.7	31.8
Vote Share	40.2	77.0	74.4	51.8
BT-Margin	40.0	76.6	75.2	52.4
BT bootstrap	45.4	80.5	58.9	31.8
BT	91.8	71.4	6.3	1.4
Learned model	42.3	75.2	59.6	41.4
BT model	100.0	64.7	0.0	0.0

Aggregation results. BT bootstrap has 1.3 points more coverage than *SE-RO* and 0.9 points less agreement. Both paired intervals exclude zero.

Replication. On 598 further source prompts, *SE-RO* raises agreement over Vote Share by 4.8 points. The 95% bootstrap interval is 2.7 to 7.0 points. Coverage rises by 3.3 points, with an interval of 1.2 to 5.5 points. Every rule keeps its thresholds from the original calibration prompts. The evaluators overlap those of the main experiment.

Learned model. Labels use pairwise admission before tier construction (Algorithm 1), a $\mathrm{Beta}(1,1)$ prior, and ties as non-support. We left-truncate each combined prompt and response at 1,024 tokens, embed it with frozen Qwen2.5-0.5B, pool the last non-padding token, and normalize to unit length. From response embeddings $\mathbf{e}_i, \mathbf{e}_j \in \mathbb{R}^{896}$, the model computes antisymmetric direction and symmetric commitment scores,

$$d_{ij} = \mathbf{w}_d^\top(\mathbf{e}_i - \mathbf{e}_j), \qquad c_{ij} = \mathbf{w}_c^\top \begin{bmatrix} |\mathbf{e}_i - \mathbf{e}_j| \\ \mathbf{e}_i \odot \mathbf{e}_j \\ (\mathbf{e}_i + \mathbf{e}_j)/2 \end{bmatrix} + b.$$

Logits for $i \succ j$, $j \succ i$, and WITHHOLD are $(c_{ij} + d_{ij}/2,\ c_{ij} - d_{ij}/2,\ 0)$. We fit $896 + 3(896) + 1 = 3{,}585$ parameters by unweighted softmax cross-entropy with full-batch AdamW [16] (learning rate and weight decay 0.001, chosen on training data). Training lasts 500–10,000 steps, stopping after ten consecutive 50-step checks with relative loss improvement below 10^{-7}. We train the learned model at 60 settings, with $\sigma \in \{0.51, 0.55, 0.60, 0.65, 0.70, 0.75\}$ and $\alpha \in \{0.5^+, 0.55, \ldots, 0.95\}$. The model predicts the more probable direction if directional probabilities sum to at least κ, otherwise withholding. Calibration maximizes macro-F1 over $\kappa \in \{0.00, 0.01, \ldots, 1.00\}$ (Section 5), breaking ties toward lower thresholds.

BT model. The BT model is a linear score with no intercept on the same embeddings. It fits a Bradley–Terry objective (ℓ_2 regularization 0.001, at most 500 iterations) to unit-weight labels from per-prompt BT fits to training votes, excluding score differences within 10^{-12}. We reuse it across evidence settings.

What the learned model withholds. At $\sigma = 0.60$ and $\alpha = 0.70$, the learned model withholds 59.6% of pairs where voters disagree and 41.4% where they agree. These rates pool pair occurrences across the five splits. The 18.2-point difference is positive in every split and has a 95% bootstrap

interval of 8.5 to 27.3 points. The interval uses 5,000 resamples of the 866 test source prompts, with models and evaluator panels fixed. Each resample keeps a source's panels and split occurrences together. Among pairs with exactly four held-out votes, the learned model withholds 64.3% where voters disagree and 38.2% where they agree. Among pairs with five to seven, it withholds 55.7% and 43.3%. Its coverage ranges from 37.0% to 45.1% across the five splits, and held-out disagreement from 21.0% to 28.6%. Across the full 60-setting grid, 16 of the 208 coverage and agreement comparisons between neighboring settings move against the trend reported in Section 5.

Worked example. Fig. 7 shows one test prompt. A rule fixed in advance picked it to show *SE-RO* withholding pairs where voters disagree, so it is not typical.

Prompt: I'm trying to reduce my energy consumption, what are some great eco-friendly alternatives to traditional light bulbs?

BT model	D > C > A > B
SE-RO	{A, C, D} > B
Learned model	{A, B, C, D}
Held-out votes	A: 9 B: 0 C: 7 D: 6

Figure 7: *SE-RO* leaves $\{A, C, D\}$ unordered, and held-out voters indeed disagree on them. The learned model withholds every pair. Braces group unordered responses. The held-out votes add evaluators of these same four responses who gave no aggregation votes.

Label and position checks. At $\sigma = 0.60$, $\alpha = 0.90$, removing WITHHOLD labels alone selects $\kappa = 0$ on all five splits: every pair is ordered, with 33.5% held-out disagreement. With models and thresholds fixed, each of the nine orderings that move all four responses to new slots raises disagreement from 18.4% to 52.6%–70.5% against the original held-out votes. Display rates are unchanged.