Weighted Soft Condorcet Optimization

Santeri Koivula
ETH Zurich
Zurich, Switzerland
skoivula@ethz.ch

ABSTRACT

Background: Soft Condorcet Optimization (SCO) evaluates agents by approximating the Kemeny-Young voting method. While standard SCO treats all pairwise disagreements equally, practical agent evaluation often prioritizes identifying top-performing agents over accurate rankings of lower-tier candidates.

Objectives and Research Questions: We introduce Weighted SCO, which incorporates Vigna's weighted Kendall-tau distance to assign higher penalties to disagreements at the top of the ranking. We investigate whether this weighting scheme improves Condorcet winner detection and top-k ranking accuracy compared to the unweighted baseline.

Methods: We derive a differentiable loss function for weighted SCO and theoretically analyze its global optimality guarantees. We empirically evaluate the method using hyperbolic, quadratic, and logarithmic weighting schemes on both real-world PrefLib preference data and synthetic tournaments with known ground truth rankings.

Results: Theoretically, we characterize when weighted SCO preserves the Condorcet guarantee: the guarantee holds when the margin exceeds a threshold determined by the weight ratio, and only constant weights give an unconditional guarantee. Empirically, logarithmic weighted SCO discovers Condorcet winners most frequently on the full PrefLib dataset (94.8% vs 81.1% for standard SCO), while maintaining near-identical performance on Kemeny-Young metrics. On synthetic data, logarithmic weights consistently achieve the best performance across all metrics, though effect sizes are modest.

Conclusions: Logarithmic weighting proves to be the most effective scheme in the SCO optimization context. It combines the global ranking consistency of standard SCO with the top-rank identification benefits of heavier weightings. While weighted SCO lacks unconditional theoretical guarantees, top-weighted gradients appear to help the optimizer converge to a Condorcet winner faster.

KEYWORDS

agent evaluation; social choice; ranking

ACM Reference Format:
Santeri Koivula. 2026. Weighted Soft Condorcet Optimization. In *Appears at the 8th Games, Agents, and Incentives Workshop (GAIW-26). Held as part of the Workshops at the 25th International Conference on Autonomous Agents and Multiagent Systems., Paphos, Cyprus, May 2026*, IFAAMAS, 10 pages.

Appears at the 8th Games, Agents, and Incentives Workshop (GAIW-26). Held as part of the Workshops at the 25th International Conference on Autonomous Agents and Multiagent Systems., Armstrong, Curry, Hosseini, Mattei, Tsang, Wąs (Chairs), May 2026, Paphos, Cyprus.

1 INTRODUCTION

Evaluating and ranking agents across diverse tasks is an important challenge in artificial intelligence. As agents become increasingly general-purpose, comparing their performance requires aggregating results across many different contexts where scoring methods and task distributions may vary substantially [9].

Recent work has addressed this challenge by drawing on social choice theory, treating tasks as voters and agents as candidates [9]. Soft Condorcet Optimization (SCO) [10] extends this approach by formulating agent ranking as an optimization problem: finding ratings whose induced ranking minimizes the sum of Kendall-tau distances to all observed preferences. SCO provides theoretical guarantees and performs well empirically, particularly in sparse data regimes common in agent evaluation.

However, standard SCO treats all pairwise disagreements equally, regardless of where they occur in the ranking. In many evaluation settings, we care primarily about identifying top-performing agents; a disagreement about which agent ranks 1st versus 2nd matters more than a disagreement about 15th versus 16th. The standard Kendall-tau distance does not capture this asymmetry.

In this paper, we introduce *weighted Soft Condorcet Optimization*, incorporating Vigna's weighted Kendall-tau distance [18] into the SCO framework. The weighted distance assigns higher penalties to disagreements involving top-ranked positions.

We make the following contributions:

(1) We formulate weighted SCO using Vigna's weighted Kendall-tau distance and derive the corresponding soft loss function (Section 4).
(2) We characterize when weighted SCO preserves Condorcet consistency, showing it depends on the interaction between the Condorcet winner's margin strength and the weight ratio (Theorem 1). As a consequence, constant weights are the only scheme with an unconditional guarantee.
(3) We evaluate weighted SCO with hyperbolic, quadratic, and logarithmic weights on PrefLib preference data and synthetic tournaments. Logarithmic weighted SCO discovers Condorcet winners most frequently (94.8%), followed by hyperbolic (93.5%), quadratic (91.1%), and unweighted SCO (81.1%). Logarithmic weights also consistently achieve the best or tied-best performance across all synthetic metrics, though differences between weighting schemes are generally small and few comparisons reach statistical significance (Section 6).

Our results show that the choice of weighting function is crucial: logarithmic weights improve both top-k accuracy and Condorcet detection while preserving global consistency, whereas more aggressive weightings sacrifice global consistency without clear benefits.

2 BACKGROUND AND RELATED WORK

2.1 Agent Evaluation

Progress in AI agents is typically measured against standardized benchmarks. However, ranking general agents poses problems, because how agents are scored varies wildly across tasks, and data collected for evaluation may not be balanced evenly across different tasks.

Classical approaches include Elo [5], TrueSkill [6], and Bradley-Terry models [2], which model pairwise win probabilities as functions of latent skill ratings. Elo in particular has been widely adopted beyond chess, including for evaluation of large language models [20].

The Bradley-Terry model is a classic framework for pairwise comparisons. Each player i is assigned a positive skill parameter $\theta_i > 0$, and the probability that player i defeats player j is modeled as $\frac{\theta_i}{\theta_i+\theta_j}$. Equivalently, defining $s_i = \log \theta_i$, this becomes the logistic function $\frac{1}{1+e^{-(s_i-s_j)}}$. The parameters are typically estimated via maximum likelihood given observed match outcomes.

Elo [5] can be understood as an online approximation to Bradley-Terry: rather than recomputing maximum likelihood estimates after each match, Elo updates ratings incrementally using a rule that corresponds to stochastic gradient descent on the Bradley-Terry likelihood. In Elo, the predicted win probability is $\frac{1}{1+10^{(r_j-r_i)/400}}$, which is equivalent to the Bradley-Terry model under a change of base.

While this approach is simple, it has various problems. Importantly, Elo cannot handle nontransitive relationships [1], a problem considering many real-world games contain large regions of strategy space that exhibit extreme non-transitivity [4]. Elo ratings can also be manipulated by introducing agents that perform similarly to each other, as shown by Lanctot et al. [9].

An alternative approach draws on social choice theory. Lanctot et al. [9] propose "Voting-as-Evaluation" (VasE), where each task is interpreted as a voter and each agent as a candidate. VasE only requires ordinal rankings and thus does not require score normalization, and makes outcomes interpretable because voting methods can be shown to satisfy various consistency properties.

2.2 Soft Condorcet Optimization

Classic voting schemes typically assume complete data, which is often unavailable in agent evaluation. Lanctot et al. [10] address this with Soft Condorcet Optimization (SCO).

The *Kendall-tau distance* between two rankings counts the number of pairwise disagreements. The *Kemeny-Young* voting method [7, 19] returns the ranking minimizing the sum of Kendall-tau distances to all votes, but computing it exactly is NP-hard. SCO provides a differentiable approximation.

Let $[\succ]$ be a *preference profile*, a collection of votes where each vote $v \in [\succ]$ represents a ranking over a subset of agents. We assume all preferences are strict total orders, meaning no voter expresses indifference between alternatives. Extension to weak preferences is possible but beyond the scope of this paper. For a vote v, we write $v[i]$ to denote the agent ranked at position i, with position 0 being the highest rank. Thus, vote v can be written as $v[0] \succ v[1] \succ \ldots \succ v[|v|-1]$.

The goal is to find a ranking of agents that best agrees with the observed votes. Searching directly over rankings is computationally intractable, so instead we assign each agent a a numerical rating θ_a and obtain the ranking by sorting: agent a ranks above agent b if and only if $\theta_a > \theta_b$. This converts the discrete search into a continuous optimization problem over the rating vector $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_m)$.

Given a preference profile $[\succ]$, we define a discrete loss:

$$L([\succ], \boldsymbol{\theta}) = \sum_{v \in [\succ]} \sum_{i<j} \mathbf{1}(\theta_{v[j]} > \theta_{v[i]}) \tag{1}$$

where $\mathbf{1}(\cdot)$ is the indicator function, equal to 1 when its argument is true and 0 otherwise. Minimizing this loss corresponds to finding the Kemeny-Young optimal ranking.

> **Example 1 (Discrete Loss Calculation).** To illustrate the discrete loss $L([\succ], \theta)$ from Equation 1, consider a scenario with three agents $\mathcal{A} = \{A, B, C\}$ and a single vote v where the voter prefers A over B, and B over C ($v : A \succ B \succ C$).
>
> Suppose the current optimizer assigns the ratings $\theta_A = 5$, $\theta_B = 15$, and $\theta_C = 10$. The induced ranking is $B \succ C \succ A$ (since $15 > 10 > 5$). The loss is calculated by summing the indicator function $1(\theta_{v[j]} > \theta_{v[i]})$ for all pairs $i < j$ (where the voter preferred $v[i]$ over $v[j]$):
>
> - **Pair** (A, B)**:** Voter requires $\theta_A > \theta_B$.
>   - Observation: $\theta_A = 5$ and $\theta_B = 15$.
>   - Result: Disagreement ($5 \not> 15$). The term $1(15 > 5)$ returns **1**.
> - **Pair** (A, C)**:** Voter requires $\theta_A > \theta_C$.
>   - Observation: $\theta_A = 5$ and $\theta_C = 10$.
>   - Result: Disagreement ($5 \not> 10$). The term $1(10 > 5)$ returns **1**.
> - **Pair** (B, C)**:** Voter requires $\theta_B > \theta_C$.
>   - Observation: $\theta_B = 15$ and $\theta_C = 10$.
>   - Result: Agreement ($15 > 10$). The term $1(10 > 15)$ returns **0**.
>
> The total discrete loss is the sum of disagreements: $1 + 1 + 0 = 2$.

Since the indicator is not differentiable, SCO replaces it with a sigmoid approximation:

$$\tilde{L}([\succ], \boldsymbol{\theta}) = \sum_{v \in [\succ]} \sum_{i<j} \frac{1}{1 + e^{(\theta_{v[i]} - \theta_{v[j]})/\tau}} \tag{2}$$

where $\tau > 0$ is a temperature parameter that controls the smoothness of the approximation.

> **Example 2 (Soft Loss Calculation).** The discrete loss is discontinuous. The soft loss $\tilde{L}([\succ], \theta)$ replaces the indicator with a sigmoid. Using the same ratings from Example 1 ($\theta_A = 5, \theta_B = 15$) and temperature $\tau = 1.0$, we calculate the soft penalty for the disagreement on pair (A, B).
>
> The vote ranks A at position 0 ($v[0]$) and B at position 1 ($v[1]$). The corresponding term in Equation 2 is:
>
> $$\frac{1}{1 + e^{(\theta_{v[0]} - \theta_{v[1]})/\tau}} = \frac{1}{1 + e^{(5-15)/1.0}} = \frac{1}{1 + e^{-10}} \approx 0.9999$$

Therefore, the loss is nearly 1. Conversely, for the agreeing pair (B, C) where $\theta_B = 15$ and $\theta_C = 10$:

$$\frac{1}{1 + e^{(15-10)/1.0}} = \frac{1}{1 + e^5} \approx 0.0067$$

Here, the ratings agree, which drives the loss close to 0. SCO minimizes the sum of these soft penalties.

This loss is differentiable, so we can minimize it using stochastic gradient descent (SGD). SGD is an iterative optimization algorithm that adjusts the ratings in the direction that reduces the loss. At each iteration, the gradient is estimated using a small random subset of the votes, called a batch, making the approach computationally efficient for large preference profiles.

An agent c is a *Condorcet winner* if more voters rank c above b than vice versa, for all $b \neq c$. Kemeny-Young is *Condorcet-consistent*: if a Condorcet winner exists, it will be top-ranked.

3 MOTIVATION

The choice of ranking method has broad societal implications. The first set of implications concerns AI evaluation more generally, and how evaluation shapes which systems get deployed. The second set of implications concerns AI alignment, and particularly the question of whose preferences AI systems will be aligned to.

3.1 Societal Stakes of AI Evaluation

Progress in artificial intelligence has historically been measured through standardized benchmarks. However, the increased generality of AI models has posed new challenges for AI evaluation. Simultaneously, as AI systems are increasingly deployed across the economy, the stakes of accurately measuring capabilities are becoming more important. On one hand, a method that fails to identify the best-performing agent can lead to suboptimal deployment decisions. On the other hand, benchmarks are also used to assess the safety of AI systems, so suboptimal rankings can lead to increased risks.

The motivation behind weighted SCO is the assumption that disagreements at the top of the ranking are much more consequential than disagreements between lower rankings. This is supported by evidence from prediction markets, where many questions are specifically about the top-performing AI model [11, 16], and other sources, such as the popular AI blogger Zvi Mowshowitz emphasizing that "Claude Opus 4.5 is the best model currently available" and that "it should definitely be your daily driver." [14]

3.2 AI Alignment as Social Choice

The challenge of AI evaluation connects to a broader question in AI alignment: whose preferences should AI systems reflect? When we train AI systems using human feedback, we are implicitly aggregating the preferences of many individuals. This is fundamentally a social choice problem [3].

A common method for aligning Large Language Models to human preferences is Reinforcement Learning from Human Feedback [15]. In RLHF, human annotators compare pairs of model outputs to express preferences. A reward model is then trained on these comparisons and the language model is then fine-tuned via reinforcement learning to maximize the learned reward. Recent work has shown that standard alignment methods, such as Reinforcement Learning from Human Feedback (RLHF), effectively implement the Borda rule, which is known to violate properties such as Condorcet consistency [17]. Methods from social choice theory offer alternatives with stronger theoretical guarantees. For example, Maura-Rivero et al. [13] show that Nash Learning From Human Feedback approximates maximal lottery outcomes and thus inherits more beneficial properties than standard RLHF.

4 WEIGHTED SCO

Standard SCO treats all pairwise disagreements equally, regardless of where they occur in the ranking. However, in agent evaluation, we are often interested in identifying the top-performing agents. Discrepancies at the top of the ranking matter more than discrepancies at the bottom. To account for this, we introduce a weighted variant of SCO based on Vigna's weighted Kendall-tau distance [18].

Kumar and Vassilvitskii [8] also study weighted variants of Kendall-tau. Their position weights model the cost of adjacent transpositions at each rank, and the weight of an inversion depends on the full permutation. In contrast, Vigna's [18] approach assigns weights as a simple function of the two positions involved, making it more suitable for pairwise optimization.

Vigna [18] proposes three weighting functions for the weighted Kendall-tau distance: logarithmic, hyperbolic, and quadratic. For rank r (starting from zero), these assign weights:

$$w_{\log}(r) = \frac{1}{\ln(r + e)}$$
$$w_{\text{hyp}}(r) = \frac{1}{r + 1}$$
$$w_{\text{quad}}(r) = \frac{1}{(r + 1)^2}$$

The weight of an exchange between positions i and j is then $w(i) + w(j)$.

Vigna finds that logarithmic weights produce results almost indistinguishable from unweighted Kendall-tau, providing insufficient differentiation between top and bottom ranks. Quadratic weights exhibit the opposite problem: the weight decays so rapidly that disagreements outside the top few positions contribute negligibly to the distance, leading to results that Vigna describes as "too uncorrelated" with standard Kendall-tau.

Hyperbolic weighting represents a middle ground. It ensures that the total weight mass grows unboundedly as the number of agents increases, whereas quadratic weights converge to a finite sum. This property makes hyperbolic weighting more robust across different ranking sizes.

To empirically assess sensitivity to the choice of weighting function, we evaluate all three weighting functions in our experiments (Section 6). This allows us to determine whether findings depend on the specific weighting scheme and, as we show, reveals that logarithmic weights perform surprisingly well in the SCO optimization context.

Following Vigna's [18] definition of weighted Kendall-tau, we apply weights multiplicatively to each pairwise term. We first define

the weighted SCO discrete loss:

$$L_W([>], \theta) = \sum_{v \in [>]} \sum_{i<j} w(i,j) \cdot \mathbf{1}(\theta_{v[j]} > \theta_{v[i]}) \tag{3}$$

where $v[i]$ denotes the agent ranked at position i by voter v, and $w(i,j)$ is a weight function that assigns higher penalties to disagreements involving top-ranked positions. The positions i and j refer to positions in the voters' rankings.

To obtain a differentiable objective, we replace the indicator function with a sigmoid approximation, yielding the soft weighted loss:

$$\tilde{L}_W([>], \theta) = \sum_{v \in [>]} \sum_{i<j} \frac{w(i,j)}{1 + e^{(\theta_{v[i]} - \theta_{v[j]})/\tau}} \tag{4}$$

where $\tau > 0$ is a temperature parameter controlling the smoothness of the approximation.

Following Vigna [18], we evaluate all three weighting functions.
Logarithmic weights assign:

$$w_{\log}(i,j) = \frac{1}{\ln(i+e)} + \frac{1}{\ln(j+e)} \tag{5}$$

Hyperbolic weights assign:

$$w_{\text{hyp}}(i,j) = \frac{1}{i+1} + \frac{1}{j+1} \tag{6}$$

Quadratic weights assign:

$$w_{\text{quad}}(i,j) = \frac{1}{(i+1)^2} + \frac{1}{(j+1)^2} \tag{7}$$

All three weighting functions penalize top-rank disagreements more heavily, but with increasing aggressiveness: logarithmic weights decay slowest, while quadratic weights decay fastest, making bottom-rank disagreements nearly negligible. Setting $w(i,j) = 1$ for all pairs recovers standard (unweighted) SCO.

To illustrate the difference between weighted and standard SCO, we compute the discrete loss L_W using hyperbolic weights for a simple preference profile.

Example 3 (Weighted Discrete Loss Calculation). Consider a similar setting as Example 1: three agents $\mathcal{A} = \{A, B, C\}$ with ratings $\theta_A = 5$, $\theta_B = 15$, $\theta_C = 10$, inducing the ranking $B \succ C \succ A$. But now we have three voters:

- Voter 1: $A \succ B \succ C$
- Voter 2: $A \succ B \succ C$
- Voter 3: $C \succ B \succ A$

Step 1: Compute the weights. For a 3-agent ranking, the position pairs and their weights are:

$$w(0,1) = \frac{1}{0+1} + \frac{1}{1+1} = \frac{3}{2} \quad \text{(1st vs 2nd)}$$
$$w(0,2) = \frac{1}{0+1} + \frac{1}{2+1} = \frac{4}{3} \quad \text{(1st vs 3rd)}$$
$$w(1,2) = \frac{1}{1+1} + \frac{1}{2+1} = \frac{5}{6} \quad \text{(2nd vs 3rd)}$$

Step 2: Calculate loss contributions.
Voters 1 and 2 (both prefer $A \succ B \succ C$):

- Pair $(0,1)$: Voter ranks A first, but $\theta_B > \theta_A$. Disagreement.
- Pair $(0,2)$: Voter ranks A above C, but $\theta_C > \theta_A$. Disagreement.
- Pair $(1,2)$: Voter ranks B above C, and $\theta_B > \theta_C$. Agreement.

Each voter contributes: unweighted 2, weighted $\frac{3}{2} + \frac{4}{3} = \frac{17}{6}$.
Voter 3 (prefers $C \succ B \succ A$):

- Pair $(0,1)$: Voter ranks C first, but $\theta_B > \theta_C$. Disagreement.
- Pair $(0,2)$: Voter ranks C above A, and $\theta_C > \theta_A$. Agreement.
- Pair $(1,2)$: Voter ranks B above A, and $\theta_B > \theta_A$. Agreement.

Voter 3 contributes: unweighted 1, weighted $\frac{3}{2}$.
Step 3: Compare total losses.

	Unweighted	Weighted
Voters 1 & 2	$2+2=4$	$\frac{17}{6} + \frac{17}{6} = \frac{17}{3}$
Voter 3	1	$\frac{3}{2}$
Total	**5**	$\frac{43}{6} \approx 7.17$

All disagreements occur at high-weight positions (pairs $(0,1)$ and $(0,2)$), so the weighted loss is substantially larger than the unweighted loss. If disagreements instead occurred only at pair $(1,2)$, the weighted loss would be relatively smaller. This asymmetry causes weighted SCO to prioritize correctly identifying top-ranked agents.

4.1 Theoretical Properties

An important theoretical result for standard SCO is a guarantee regarding the global minimum: if a Condorcet winner exists, the global minimum of the sigmoid loss assigns maximum rating to that agent (Lanctot et al., Theorem 1).

However, the soft loss is nonconvex, so SGD is not guaranteed to find this global minimum in practice. The theorem nonetheless justifies the choice of loss function by showing that the optimization target is correct. A natural question is whether weighted SCO preserves this theoretical guarantee. We give a characterization: the guarantee depends on the interaction between the Condorcet winner's margin strength and the *weight ratio* $r = w_{\max}/w_{\min}$, where $w_{\max} = \max_{i<j} w(i,j)$ and $w_{\min} = \min_{i<j} w(i,j)$ are the extreme weights. The ratio $r \geq 1$ measures the dynamic range of the weighting function: $r = 1$ for constant weights, with larger values indicating more aggressive top-weighting.

Theorem 1 (Condorcet guarantee for weighted SCO)**.** Let $m = |\mathcal{A}| \geq 3$ be the number of agents and let

$$\tilde{L}_W([>], \theta) = \sum_{v \in [>]} \sum_{i<j} w(i,j)\, \sigma(\theta_{v[j]} - \theta_{v[i]}) \tag{8}$$

be the weighted soft loss, where $\sigma(z) = (1 + e^{-z/\tau})^{-1}$. Let $r = w_{\max}/w_{\min}$ be the weight ratio, and assume $w_{\max} = w(0,1)$. Suppose a Condorcet winner c exists in an election with n voters, and let k be c's minimum pairwise margin, i.e., the smallest number of votes by which c defeats any opponent.

(i) **Guarantee.** If $\dfrac{k}{n} > \dfrac{r-1}{r+1}$, then the global minimum of $\tilde{L}_W$ top-ranks c.

(ii) **Tightness.** For any k and all sufficiently large n (depending on k, m, and w) with $\frac{k}{n} < \frac{r-1}{r+1}$, there exists a preference profile with n voters and a Condorcet winner with margin k such that the global minimum of $\tilde{L}_W$ does not necessarily top-rank c.

Proof. We first regroup the loss to iterate over unordered pairs of agents $\{a, b\} \subseteq \mathcal{A}$. Define the *weighted preference count*

$$N_W(a,b) = \sum_{v \in [\succ]} \sum_{i<j} w(i,j) \cdot \mathbf{1}(v[i] = a \wedge v[j] = b)$$

and the *weighted margin* $M_W(a,b) = N_W(a,b) - N_W(b,a)$. Applying the sigmoid symmetry $\sigma(z) = 1 - \sigma(-z)$, the loss simplifies to

$$\tilde{L}_W = C + \sum_{\{a,b\} \subseteq \mathcal{A}} M_W(a,b)\, \sigma(\theta_b - \theta_a), \tag{9}$$

where $C = \sum_{\{a,b\}} N_W(a,b)$ is a constant independent of θ. For the loss to be strictly decreasing in the rating of a Condorcet winner c, we require $M_W(c,b) > 0$ for all $b \neq c$.

Part (i). Fix an opponent b. Let n_c be the number of voters ranking c above b and $n_b = n - n_c$. Since c is a Condorcet winner with minimum margin k, we have $n_c - n_b \geq k$. Combined with $n_c + n_b = n$, this gives $n_c \geq (n+k)/2$ and $n_b \leq (n-k)/2$. Each voter preferring c to b contributes at least $w_{\min}$ to $N_W(c,b)$; each voter preferring b to c contributes at most $w_{\max}$ to $N_W(b,c)$. Therefore:

$$M_W(c,b) \geq \frac{n+k}{2}\, w_{\min} - \frac{n-k}{2}\, w_{\max}.$$

This is strictly positive whenever $(n+k)\, w_{\min} > (n-k)\, w_{\max}$. Dividing both sides by $n\, w_{\min}$ and rearranging gives $k/n > (r-1)/(r+1)$.

Part (ii). Let $x = \lfloor k/2 \rfloor + 1$. Construct a profile over m candidates $\{c, b, d_1, \ldots, d_{m-2}\}$ with three voter groups:

- **Minority** ($\frac{n-k}{2}$ voters): $b \succ c \succ d_1 \succ \cdots \succ d_{m-2}$. The b-over-c preference is at positions $(0,1)$, contributing weight $w_{\max}$.
- **Majority-A** ($\frac{n+k}{2} - x$ voters): $d_1 \succ \cdots \succ d_{m-2} \succ c \succ b$. The c-over-b preference is at positions $(m-2, m-1)$, contributing weight $w_{\min}$.
- **Majority-B** (x voters): $c \succ d_1 \succ \cdots \succ d_{m-2} \succ b$. The c-over-b preference is at positions $(0, m-1)$, contributing weight $w(0, m-1)$.

All majority voters rank c above b, so c defeats b by margin k. For any filler d_i: minority and Majority-B voters rank c above d_i, while Majority-A voters rank d_i above c; the margin of c over d_i is $\frac{n-k}{2} + x - (\frac{n+k}{2} - x) = 2x - k \geq 1$, so c is the Condorcet winner. The weighted margin is:

$$M_W(c,b) = \Big(\frac{n+k}{2} - x\Big) w_{\min} + x\, w(0,m-1) - \frac{n-k}{2}\, w_{\max}. \tag{10}$$

If all majority voters used the Majority-A ordering, the margin would be $\frac{n+k}{2}\, w_{\min} - \frac{n-k}{2}\, w_{\max}$. The Majority-B group replaces x of these voters, upgrading their per-voter weight from $w_{\min}$ to $w(0,m-1)$. This adds $x\,(w(0,m-1) - w_{\min})$ to the margin—a quantity that depends only on k and m, not on n. Rearranging (10), $M_W(c,b) < 0$ whenever

$$\frac{k}{n} < \frac{r-1}{r+1} - \frac{2x\,(w(0,m-1) - w_{\min})}{n\,(w_{\max} + w_{\min})}.$$

The second term vanishes as $n \to \infty$, so for any fixed k with $k/n < (r-1)/(r+1)$, all sufficiently large n satisfy this bound. Since $M_W(c,b) < 0$, minimizing (9) drives $\theta_b > \theta_c$, so the global minimum does not top-rank c. □

Two corollaries follow immediately: the threshold $(r-1)/(r+1)$ vanishes if and only if $r = 1$.

Corollary 1 (Condorcet-compatible weights)**.** A weight function w preserves Condorcet consistency for all preference profiles if and only if w is constant.

Corollary 2 (Vigna weights)**.** Hyperbolic, quadratic, and logarithmic weighting schemes fail Condorcet consistency.

A non-constant weight function allows a Condorcet winner's pairwise victories to be concentrated at low-weight positions while its losses occur at high-weight positions, flipping the weighted margin. The strength of this effect is determined by r: for $m = 3$, logarithmic weights have $r \approx 5/4$, hyperbolic $r = 9/5$, and quadratic $r \approx 3.46$, requiring progressively larger margins. As m grows, $w_{\min}$ decreases while $w_{\max} = w(0,1)$ stays fixed, so r increases and the requirement becomes stricter.

Since both the standard and weighted losses are nonconvex, and neither approach guarantees that SGD finds a global minimum, the practical implications of these theoretical results are not immediately clear. We investigate this empirically in Section 6.

5 EXPERIMENTAL SETUP

5.1 Implementation Details

We implement weighted SCO by incorporating Vigna weights into the gradient descent procedure of Lanctot et al. [10], using the weighted soft loss defined in Section 4.

All experiments use stochastic gradient descent with the following hyperparameters, matching those of Lanctot et al. where applicable:

- Learning rate $\alpha = 0.01$
- Temperature $\tau = 1.0$
- Iterations $T = 10{,}000$
- Initial ratings $\theta_a = 50.0$ for all agents
- Rating bounds $[\theta_{\min}, \theta_{\max}] = [0, 100]$

5.2 PrefLib Experiments

We evaluate on strictly-ordered complete (SOC) and strictly-ordered incomplete (SOI) preference profiles from the PrefLib archive [12]. To enable comparison against Kemeny-Young optimal rankings (which require brute-force enumeration), we restrict to instances with $|A| \leq 10$ alternatives, yielding $N = 4915$ preference profiles.

For Condorcet winner detection rates, we report results on two datasets: (i) the $|A| \leq 10$ subset ($N = 4915$) for direct comparison with Kemeny-Young metrics, and (ii) the full PrefLib dataset to assess performance across all instance sizes, including large elections where weighted SCO may have different convergence properties.

To understand the optimization dynamics underlying Condorcet winner detection, we additionally analyze convergence behavior by tracking the rank assigned to the Condorcet winner throughout training. For this analysis, we stratify the PrefLib instances by the number of alternatives: small instances ($|A| \leq 10$) and large

instances ($|A| > 10$). We average the Condorcet winner's rank across 500 elections at each training iteration, using only instances where a Condorcet winner exists.

For each instance, we run SCO with batch size $|B| = 32$ and compare the resulting ranking against (i) the Kemeny-Young optimal ranking, and (ii) the Condorcet winner when one exists. We run each configuration for uniform, hyperbolic, quadratic, and logarithmic weights.

5.3 Synthetic Tournament Experiments

Following Section 4.3 of Lanctot et al. [10], we generate synthetic tournament data with known ground truth rankings. We simulate $|A| = 20$ agents with true skill ratings drawn from $\mathcal{N}(100, 30)$. Contests involve 4 agents, with outcomes determined by adding performance noise $\epsilon \sim \mathcal{N}(0, 5)$ to true ratings and sorting.

We test two contest generation distributions:

- **Uniform**: Agents sampled uniformly at random.
- **Skill-matched**: Agents selected to have similar skill levels, simulating matchmaking systems.

For each configuration (number of contests $\in \{10, 30, 50, 100\}$, distribution, weight function), we run 50 independent trials with different random seeds and report means with 95% confidence intervals. We use batch size $|B| = 16$ for synthetic experiments.

5.4 Evaluation Metrics

For PrefLib experiments without ground truth, we report:

- Normalized Kendall-tau distance to Kemeny-Young ranking
- Condorcet winner detection rate (proportion of instances where the Condorcet winner, if one exists, is top-ranked)
- Top-k overlap with Kemeny-Young ranking for $k \in \{1, 3, 5\}$

For synthetic experiments with ground truth, we additionally report:

- Kendall-tau distance to true ranking (KTD)
- Mean true rating distance of misranked pairs (MTRD)
- Top-k precision and top-k restricted Kendall-tau distance

To assess the statistical significance of differences between weighting schemes on top-k metrics, we perform pairwise Wilcoxon signed-rank tests, paired by random seed. This non-parametric test was chosen because (1) the data consists of paired observations sharing the same random seed across conditions, and (2) it makes no assumptions about the underlying distribution of the differences.

For each combination of distribution type (uniform, skill-matched) and contest count (10, 30, 50, 100), we compared all six pairs of weighting schemes on four metrics: Top-1 Accuracy, Top-3 Precision, Top-5 Precision, and Top-5 KTD. To control for multiple comparisons, we applied Bonferroni correction, adjusting the significance threshold from $\alpha = 0.05$ to $\alpha = 0.0083$ (0.05 / 6 comparisons).

Each condition included $n = 50$ independent trials with matched random seeds across weighting schemes, yielding paired samples for the statistical tests.

6 RESULTS

We evaluate weighted SCO on preference profiles from the PrefLib archive [12] and synthetic tournament data with known ground truth rankings. Code for all experiments is available at https://github.com/sanyerr/weight-sco.

Table 1: Comparison of SCO variants on PrefLib data ($N = 4915$, $|A| \leq 10$). Metrics evaluated against Kemeny-Young optimal rankings.

Metric	Unif.	Hyp.	Quad.	Log.
KT to Kemeny (mean)	0.036	0.076	0.120	0.042
Perfect Kemeny match (%)	74.0	56.4	41.9	71.6
Top-1 accuracy (%)	92.9	88.7	85.5	92.5
Top-3 overlap (%)	96.7	92.4	88.0	95.9
Top-5 overlap (%)	93.2	89.9	87.1	92.3
Condorcet detected, $\|A\| \leq 10$ (%)	99.8	96.3	93.1	99.6
Condorcet detected, all data (%)	81.1	93.5	91.1	94.8

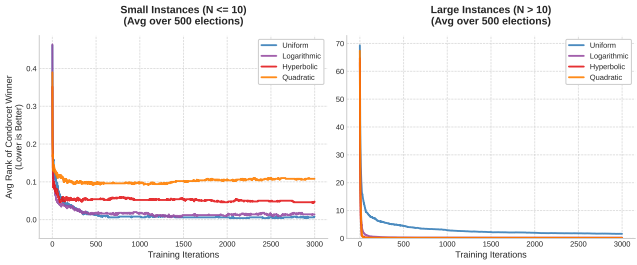

Figure 1: Convergence of the Condorcet winner's rank during optimization, averaged over 500 PrefLib elections where a Condorcet winner exists. Lower is better. Left: small instances ($|A| \leq 10$). Right: large instances ($|A| > 10$).

6.1 PrefLib Experiments

We first evaluate on preference profiles from the PrefLib archive. Since ground truth rankings are unavailable for real-world data, we use Kemeny-Young optimal rankings as a reference point. Logarithmic SCO discovers a Condorcet winner 94.8% of the time when one exists, followed by hyperbolic (93.5%), quadratic (91.1%), and standard SCO (81.1%). Logarithmic weights maintain near-identical performance to standard SCO on Kemeny-Young metrics, while hyperbolic and especially quadratic weights show greater deviation.

On the $|A| \leq 10$ subset, standard SCO discovers a Condorcet winner 99.8% of the time, followed by logarithmic (99.6%), hyperbolic (96.3%), and quadratic (93.1%). Across all PrefLib data (including larger instances), these rates reverse: logarithmic achieves 94.8%, hyperbolic 93.5%, quadratic 91.1%, and standard SCO only 81.1%. The substantial difference between small and large instances motivates our convergence analysis in Section 6.1.1.

The better performance of standard SCO on Kemeny-Young rankings is expected, as standard SCO directly optimizes for minimizing the sum of Kendall-tau distances.

6.1.1 Convergence Towards Detecting Condorcet Winners. To understand why weighted SCO achieves higher Condorcet winner detection despite the theoretical limitations identified in Theorem 1, we analyzed convergence dynamics by tracking the average rank of the Condorcet winner throughout training (Figure 1).

The results reveal a trade-off between theoretical guarantees and convergence speed that varies with problem size. On small instances ($|A| \leq 10$), standard and logarithmic SCO converge to rank 0, consistent with the global optimality guarantee for standard SCO and logarithmic weights' near-uniform behavior. Hyperbolic and quadratic SCO stabilize at slightly higher average ranks, confirming that more aggressive weightings occasionally prefer non-Condorcet winners as predicted by Theorem 1.

However, on large instances ($|A| > 10$), standard SCO converges to an average rank of approximately 1–2, while all weighted variants converge to nearly 0. This suggests that the gradient signal from top-weighted disagreements helps weighted SCO navigate the loss landscape more effectively in higher-dimensional settings, where the standard objective may have more local minima or flatter regions near the optimum.

This behavior explains the overall Condorcet detection rates. On the $|A| \leq 10$ subset, standard and logarithmic SCO achieve near-perfect detection (99.8% and 99.6%), while hyperbolic and quadratic achieve 96.3% and 93.1%. However, on the full dataset, all weighted variants substantially outperform standard SCO (94.8%, 93.5%, and 91.1% vs 81.1%), with logarithmic achieving the highest rate.

6.2 Synthetic Data Experiments

To obtain a fair comparison between standard and weighted SCO, we generate synthetic tournament data where ground truth skill ratings are known. Following the setup of Lanctot et al. [10], we simulate 20 agents with true skill ratings drawn from $\mathcal{N}(100, 30)$. Contests between 4 agents are generated, and outcomes are determined by adding noise $\mathcal{N}(0, 5)$ to true ratings. We test two contest generation distributions: uniform (random sampling) and skill-matched (agents matched by similar skill).

6.2.1 Global Metrics. We measure Kendall-tau distance (KTD) and mean true rating distance (MTRD) between the predicted and true rankings. MTRD captures the average skill difference between misranked pairs.

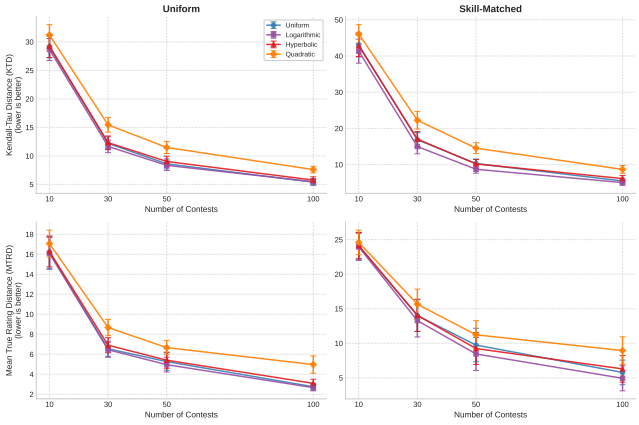

Figure 2: Global ranking metrics on synthetic data. KTD: Kendall-tau distance to ground truth. MTRD: Mean True Rating Distance of misranked pairs. Error bars show 95% CI over 50 trials. Full numerical results in Appendix A.

Results show that logarithmic weights consistently achieve the lowest KTD and MTRD across nearly all configurations, while quadratic weights perform worst. Uniform and hyperbolic weights perform similarly to each other.

6.2.2 Top-k Metrics. Since weighted SCO is designed to prioritize accuracy at the top of the ranking, we also evaluate top-k metrics: top-1 accuracy, top-3 and top-5 precision, and Kendall-tau distance restricted to the top 5 agents. Under uniform matchups, no comparisons survived Bonferroni correction, suggesting weighting scheme choice has limited impact when matchups are uniformly distributed. We focus on skill-matched results here (Figure 3); full results including uniform matchups are in Appendix A.

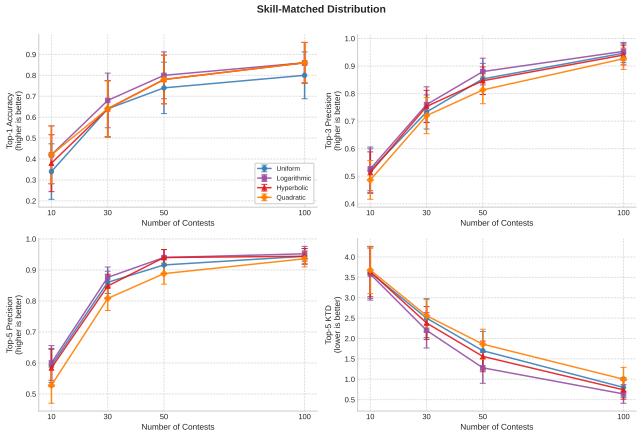

Figure 3: Top-k metrics on synthetic data under skill-matched matchups. Top-1 Acc: top-1 accuracy. Top-3/5 P: top-3/5 precision. Top-5 KTD: Kendall-tau on top 5. Error bars show 95% CI over 50 trials.

Under skill-matched matchups, logarithmic weights achieve the best top-k performance across all metrics. Logarithmic SCO significantly outperforms quadratic on top-5 precision at 10, 30, and 50 contests and on top-5 KTD at 30, 50, and 100 contests (all $p < 0.0083$, Bonferroni-corrected). Logarithmic also significantly outperforms uniform weights on top-5 KTD at 30 and 50 contests ($p < 0.0083$). Top-1 accuracy showed no significant differences between any weighting schemes at any contest count.

Given the number of conditions tested, these results should be interpreted cautiously. The clearest patterns are that quadratic weights impair top-5 identification in skill-matched settings with limited data, and that logarithmic weights consistently achieve the best or tied-best top-k performance.

7 DISCUSSION

Weighted SCO demonstrates improved Condorcet winner detection on the full PrefLib dataset, and logarithmic weights additionally yield improvements in top-k accuracy on synthetic data. However, there is a key limitation to the weighted approach: in real-world settings without access to ground truth rankings, it is difficult to evaluate how weighted SCO compares to standard SCO. Using Kemeny-Young as a proxy may unfairly penalize weighted SCO,

since standard SCO directly optimizes for this criterion. Developing appropriate evaluation metrics for weighted ranking methods in the absence of ground truth remains an open challenge.

7.1 Choice of Weighting Function

We evaluated all three of Vigna's weighting functions. As predicted by Vigna [18], quadratic weights proved too aggressive, degrading global metrics (41.9% Kemeny-Young match vs 71.6% for logarithmic and 74.0% for uniform) without providing advantages in top-k accuracy. However, contrary to Vigna's finding that logarithmic weights are "almost indistinguishable" from unweighted Kendall-tau, logarithmic weights proved to be the most effective scheme in the SCO optimization context. Logarithmic SCO achieves the highest Condorcet detection rate (94.8%), the best global ranking metrics on synthetic data, and improvements in top-k accuracy under skill-matched matchups, while maintaining near-identical Kemeny-Young performance to standard SCO (71.6% vs 74.0% perfect match). This suggests that the gentle top-weighting of logarithmic weights provides sufficient gradient signal to improve top-rank identification without the distortions introduced by more aggressive schemes.

7.2 Theoretical Guarantees vs. Empirical Performance

The superior performance of weighted SCO in identifying Condorcet winners may appear surprising given the theoretical properties derived in Section 4. However, while Theorem 1 shows that no non-constant weighting preserves the Condorcet guarantee unconditionally, the guarantee does hold for Condorcet winners with sufficiently large margins.

SGD is not guaranteed to find a global minimum in practice. Our convergence analysis (Figure 1) provides insight into this apparent paradox. On small instances ($|A| \leq 10$), standard and logarithmic SCO converge to rank 0, while hyperbolic and quadratic SCO stabilize at slightly higher average ranks due to the mechanism identified in Theorem 1. However, on larger instances ($|A| > 10$), this relationship reverses: all weighted variants converge more reliably to the Condorcet winner than standard SCO.

We hypothesize that the amplified gradient signal from top-rank disagreements helps weighted SCO escape local minima or navigate flat regions in the loss landscape that trap standard SCO in higher-dimensional settings. Furthermore, Condorcet winners in real-world elections may win by comfortable margins, well above the thresholds identified in Theorem 1. Since large instances dominate the PrefLib dataset and show the most substantial performance differences (99.8% vs 99.6% on $|A| \leq 10$ instances for standard vs logarithmic, but 81.1% vs 94.8% overall), the weighted variants' advantage on large instances drives their superior overall Condorcet detection rates.

7.3 Limitations and Future Work

Our experimental setup follows Lanctot et al. [10] to enable direct comparison with prior work. However, their synthetic tournaments were designed to evaluate the accuracy of global rankings, not top-k performance specifically. Since weighted SCO prioritizes resolving disagreements at the top of the ranking, its potential benefit is largest when top-rank identification is difficult. A more targeted evaluation would examine settings where this ambiguity is present, such as profiles with tightly clustered top-agent skills.

Regarding implementation, our multiplicative weighting follows Vigna's definition of weighted Kendall-tau. An alternative approach would be to incorporate weights into the temperature parameter τ. This would affect the sharpness of the sigmoid function rather than the penalty magnitude, changing gradient dynamics in ways that warrant future investigation. More broadly, Lanctot et al. [10] note that differentiable approximations to other distance functions, such as Spearman's footrule, may be worthwhile. Weighted variants of these alternatives present additional directions for future work.

8 CONCLUSION

We introduced weighted Soft Condorcet Optimization, extending the SCO framework with Vigna's weighted Kendall-tau distance to prioritize accuracy at the top of the ranking. We evaluated all three of Vigna's weighting functions and found that logarithmic weights are the most effective in the SCO context. Logarithmic SCO achieves the highest Condorcet winner detection rate on the full PrefLib dataset (94.8% vs 81.1% for standard SCO), maintains near-identical Kemeny-Young performance to standard SCO, and yields best performance across synthetic metrics.

Theoretically, we characterized when the global minimum in weighted SCO is guaranteed to top-rank a Condorcet winner (Theorem 1): the guarantee holds when the Condorcet winner's margin exceeds a threshold determined by the weight ratio. As a consequence, constant weights are the only scheme with an unconditional guarantee.

Despite lacking an unconditional theoretical guarantee, weighted SCO empirically finds Condorcet winners more frequently than standard SCO. The success of logarithmic weights highlights that even mild top-weighting can substantially improve optimization dynamics.

ACKNOWLEDGMENTS

We thank Barton Lee for his supervision and feedback, and the anonymous reviewers of the Games, Agents, and Incentives Workshop at AAMAS 2026 for their comments on an earlier version of this paper.

REFERENCES

[1] David Balduzzi, Karl Tuyls, Julien Perolat, and Thore Graepel. 2018. Re-evaluating evaluation. In *Advances in Neural Information Processing Systems*, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (Eds.), Vol. 31. Curran Associates, Inc.

[2] R. A. Bradley and M. E. Terry. 1952. Rank Analysis of Incomplete Block Designs I: The Method of Paired Comparisons. *Biometrika* 39, 3/4 (1952), 324–345.

[3] Vincent Conitzer, Rachel Freedman, Jobst Heitzig, Wesley H. Holliday, Bob M. Jacobs, Nathan Lambert, Milan Mossé, Eric Pacuit, Stuart Russell, Hailey Schoelkopf, Emanuel Tewolde, and William S. Zwicker. 2024. Position: Social Choice Should Guide AI Alignment in Dealing with Diverse Human Feedback. In *Proceedings of the 41st International Conference on Machine Learning*. PMLR, 9346–9360. https://proceedings.mlr.press/v235/conitzer24a.html

[4] Wojciech M. Czarnecki, Gauthier Gidel, Brendan Tracey, Karl Tuyls, Shayegan Omidshafiei, David Balduzzi, and Max Jaderberg. 2020. Real world games look like spinning tops. In *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin (Eds.), Vol. 33. Curran Associates, Inc., 17443–17454.

[5] Arpad E. Elo. 1978. *The Rating of Chessplayers, Past and Present.* Arco Publishing, New York.

[6] Ralf Herbrich, Tom Minka, and Thore Graepel. 2006. TrueSkill: A Bayesian Skill Rating System. In *Advances in Neural Information Processing Systems*, B. Schölkopf, J. Platt, and T. Hoffman (Eds.), Vol. 19. MIT Press.
[7] J. Kemeny. 1959. Mathematics Without Numbers. *Daedalus* 88, 4 (1959), 577–591.
[8] Ravi Kumar and Sergei Vassilvitskii. 2010. Generalized Distances between Rankings. In *Proceedings of the 19th International Conference on World Wide Web (WWW '10)*. ACM, Raleigh, North Carolina, USA, 571–580. https://doi.org/10.1145/1772690.1772749
[9] Marc Lanctot, Kate Larson, Yoram Bachrach, Luke Marris, Zun Li, Avishkar Bhoopchand, Thomas Anthony, Brian Tanner, and Anna Koop. 2023. Evaluating Agents using Social Choice Theory. *arXiv preprint arXiv:2312.03121* (2023). Presented at the AAMAS 2024 Games, Agents, and Incentives Workshop.
[10] Marc Lanctot, Kate Larson, Michael Kaisers, Quentin Berthet, Ian Gemp, Manfred Diaz, Roberto-Rafael Maura-Rivero, Yoram Bachrach, Anna Koop, and Doina Precup. 2025. Soft Condorcet Optimization for Ranking of General Agents. In *Proceedings of the 24th International Conference on Autonomous Agents and Multiagent Systems*. International Foundation for Autonomous Agents and Multiagent Systems, 1253–1262. https://dl.acm.org/doi/10.5555/3709347.3743757 Best Paper Award.
[11] Manifold Markets. 2025. *Which company has best AI model end of 2025? (Chatbot Arena Leaderboard)*. https://manifold.markets/Bayesian/which-company-has-best-ai-model-end-QcqLtdnQnN Prediction market. Resolution based on LMSYS Chatbot Arena Leaderboard.
[12] Nicholas Mattei and Toby Walsh. 2013. PrefLib: A Library for Preferences. In *Proceedings of the 3rd International Conference on Algorithmic Decision Theory*.
[13] Roberto-Rafael Maura-Rivero, Marc Lanctot, Francesco Visin, and Kate Larson. 2025. Jackpot! Alignment as a Maximal Lottery. https://doi.org/10.48550/arXiv.2501.19266 arXiv:2501.19266 [cs].
[14] Zvi Mowshowitz. 2025. *Claude Opus 4.5 Is The Best Model Available*. https://thezvi.substack.com/p/claude-opus-45-is-the-best-model Substack newsletter.
[15] Long Ouyang, Jeff 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 Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems* 35 (2022), 27730–27744.
[16] Polymarket. 2025. *Which company has the best AI model end of March?* https://polymarket.com/event/which-company-has-the-best-ai-model-end-of-march-751 Prediction market with $778,230 trading volume. Resolution based on LMSYS Chatbot Arena Leaderboard.
[17] Ariel D. Procaccia, Benjamin Schiffer, and Shirley Zhang. 2025. Clone-Robust AI Alignment. https://doi.org/10.48550/arXiv.2501.09254 arXiv:2501.09254 [cs] version: 1.
[18] Sebastiano Vigna. 2015. A Weighted Correlation Index for Rankings with Ties. In *Proceedings of the 24th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 1166–1176. https://doi.org/10.1145/2736277.2741088
[19] H. P. Young and A. Levenglick. 1978. A Consistent Extension of Condorcet's Election Principle. *SIAM J. Appl. Math.* 35, 2 (1978), 285–300.
[20] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. *arXiv preprint arXiv:2306.05685* (2023).

A FULL SYNTHETIC EXPERIMENT RESULTS

Table 2: Global ranking metrics on synthetic data. KTD: Kendall-tau distance to ground truth. MTRD: Mean True Rating Distance of misranked pairs. Lower is better. Bold indicates best in group. Mean ± 95% CI over 50 trials.

Contests	Weight	KTD	MTRD
	Uniform Distribution		
10	uniform	29.10 ± 1.91	16.21 ± 1.61
10	hyperbolic	29.26 ± 1.99	16.28 ± 1.53
10	quadratic	31.20 ± 1.82	17.07 ± 1.35
10	logarithmic	**28.66 ± 1.94**	**16.09 ± 1.59**
30	uniform	12.20 ± 1.12	6.55 ± 0.75
30	hyperbolic	12.32 ± 1.17	6.90 ± 0.75
30	quadratic	15.44 ± 1.28	8.67 ± 0.81
30	logarithmic	**11.66 ± 1.09**	**6.45 ± 0.76**
50	uniform	8.60 ± 0.87	5.27 ± 0.87
50	hyperbolic	9.02 ± 0.92	5.41 ± 0.80
50	quadratic	11.48 ± 1.05	6.67 ± 0.72
50	logarithmic	**8.34 ± 0.90**	**4.96 ± 0.74**
100	uniform	**5.40 ± 0.56**	2.75 ± 0.33
100	hyperbolic	5.76 ± 0.57	3.09 ± 0.41
100	quadratic	7.60 ± 0.58	4.97 ± 0.87
100	logarithmic	5.46 ± 0.53	**2.66 ± 0.32**
	Skill-Matched Distribution		
10	uniform	43.06 ± 3.33	24.08 ± 1.92
10	hyperbolic	42.84 ± 2.99	24.14 ± 1.91
10	quadratic	46.08 ± 2.58	24.60 ± 1.77
10	logarithmic	**41.34 ± 3.33**	**23.95 ± 1.96**
30	uniform	16.88 ± 1.95	14.00 ± 2.30
30	hyperbolic	17.04 ± 2.10	14.04 ± 2.33
30	quadratic	22.26 ± 2.44	15.65 ± 2.19
30	logarithmic	**14.98 ± 1.98**	**13.28 ± 2.37**
50	uniform	10.28 ± 1.26	9.74 ± 2.43
50	hyperbolic	10.20 ± 1.23	9.24 ± 2.31
50	quadratic	14.56 ± 1.50	11.22 ± 2.04
50	logarithmic	**8.72 ± 1.08**	**8.46 ± 2.39**
100	uniform	5.48 ± 0.75	5.78 ± 1.78
100	hyperbolic	6.14 ± 0.84	6.29 ± 1.93
100	quadratic	8.68 ± 1.08	8.95 ± 1.99
100	logarithmic	**5.04 ± 0.76**	**4.93 ± 1.81**

Table 3: Top-k metrics on synthetic data. T1: top-1 accuracy. T3/T5 P: top-3/5 precision. T5 KTD: Kendall-tau on top 5. Bold indicates best in group. Mean ± 95% CI over 50 trials.

n	Weight	T1	T3 P	T5 P	T5 KTD
		Uniform Distribution			
10	unif.	**0.44 ± 0.14**	**0.64 ± 0.06**	0.73 ± 0.05	3.44 ± 0.55
10	hyp.	0.42 ± 0.14	0.63 ± 0.06	0.73 ± 0.04	3.42 ± 0.53
10	quad.	0.42 ± 0.14	**0.64 ± 0.06**	0.73 ± 0.04	**3.28 ± 0.53**
10	log.	0.42 ± 0.14	0.63 ± 0.06	**0.74 ± 0.05**	3.44 ± 0.56
30	unif.	0.58 ± 0.14	0.83 ± 0.05	0.87 ± 0.03	1.78 ± 0.41
30	hyp.	0.60 ± 0.14	**0.84 ± 0.05**	**0.88 ± 0.03**	**1.66 ± 0.36**
30	quad.	**0.62 ± 0.14**	0.81 ± 0.05	**0.88 ± 0.03**	1.74 ± 0.38
30	log.	0.60 ± 0.14	0.83 ± 0.05	**0.88 ± 0.03**	1.68 ± 0.36
50	unif.	0.72 ± 0.13	**0.89 ± 0.04**	**0.92 ± 0.03**	1.26 ± 0.32
50	hyp.	**0.80 ± 0.11**	0.87 ± 0.05	0.91 ± 0.03	1.16 ± 0.29
50	quad.	0.76 ± 0.12	0.86 ± 0.05	0.90 ± 0.03	1.24 ± 0.28
50	log.	**0.80 ± 0.11**	**0.89 ± 0.04**	**0.92 ± 0.03**	**1.08 ± 0.28**
100	unif.	0.80 ± 0.11	**0.91 ± 0.04**	**0.95 ± 0.02**	0.92 ± 0.24
100	hyp.	**0.82 ± 0.11**	**0.91 ± 0.04**	0.94 ± 0.03	**0.88 ± 0.25**
100	quad.	**0.82 ± 0.11**	**0.91 ± 0.04**	0.94 ± 0.03	1.00 ± 0.26
100	log.	0.80 ± 0.11	**0.91 ± 0.04**	**0.95 ± 0.02**	0.96 ± 0.25
		Skill-Matched Distribution			
10	unif.	0.34 ± 0.13	0.52 ± 0.08	0.59 ± 0.05	3.62 ± 0.64
10	hyp.	0.38 ± 0.14	0.51 ± 0.08	0.58 ± 0.06	3.64 ± 0.62
10	quad.	**0.42 ± 0.14**	0.49 ± 0.07	0.53 ± 0.06	3.68 ± 0.58
10	log.	**0.42 ± 0.14**	**0.53 ± 0.08**	**0.60 ± 0.06**	**3.58 ± 0.64**
30	unif.	0.64 ± 0.13	0.73 ± 0.06	0.86 ± 0.04	2.50 ± 0.48
30	hyp.	0.64 ± 0.13	0.75 ± 0.06	0.85 ± 0.04	2.38 ± 0.40
30	quad.	0.64 ± 0.13	0.72 ± 0.07	0.81 ± 0.04	2.56 ± 0.39
30	log.	**0.68 ± 0.13**	**0.76 ± 0.06**	**0.88 ± 0.03**	**2.20 ± 0.43**
50	unif.	0.74 ± 0.12	0.85 ± 0.06	0.92 ± 0.03	1.70 ± 0.47
50	hyp.	0.78 ± 0.12	0.85 ± 0.05	**0.94 ± 0.03**	1.56 ± 0.38
50	quad.	0.78 ± 0.12	0.81 ± 0.05	0.89 ± 0.03	1.86 ± 0.37
50	log.	**0.80 ± 0.11**	**0.88 ± 0.05**	**0.94 ± 0.03**	**1.28 ± 0.38**
100	unif.	0.80 ± 0.11	**0.95 ± 0.03**	0.94 ± 0.03	0.80 ± 0.24
100	hyp.	**0.86 ± 0.10**	0.94 ± 0.04	0.94 ± 0.03	0.74 ± 0.23
100	quad.	**0.86 ± 0.10**	0.93 ± 0.04	0.94 ± 0.03	1.00 ± 0.29
100	log.	**0.86 ± 0.10**	**0.95 ± 0.03**	**0.95 ± 0.02**	**0.64 ± 0.23**