A STATISTICAL FRAMEWORK FOR RANKING LLM-BASED CHATBOTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language models (LLMs) have transformed natural language processing, with frameworks like Chatbot Arena providing pioneering platforms for evaluating these models. By facilitating millions of pairwise comparisons based on human judgments, Chatbot Arena has become a cornerstone in LLM evaluation, offering rich datasets for ranking models in open-ended conversational tasks. Building upon this foundation, we propose a statistical framework that incorporates key advancements to address specific challenges in pairwise comparison analysis. First, we introduce a factored tie model that enhances the ability to handle ties—an integral aspect of human-judged comparisons—significantly improving the model's fit to observed data. Second, we extend the framework to model covariance between competitors, enabling deeper insights into performance relationships and facilitating intuitive groupings into performance tiers. Third, we resolve optimization challenges arising from parameter invariances by introducing novel constraints, ensuring stable and interpretable parameter estimation. Through rigorous evaluation and extensive experimentation, our framework demonstrates substantial improvements over existing methods in modeling pairwise comparison data. To support reproducibility and practical adoption, we release `leaderbot`, an open-source Python package implementing our models and analyses.

1 INTRODUCTION

The rapid advancement of large language models (LLMs) has transformed natural language processing, enabling breakthroughs across diverse tasks. As these models evolve, the need for effective evaluation methods becomes crucial for fostering innovation and ensuring that LLMs align with human preferences.

Traditional benchmarks, such as MMLU (Hendrycks et al., 2021) and HumanEval (Chen et al., 2021), play an important role in assessing specific capabilities of LLMs. However, they often fall short in capturing the nuanced, real-world interactions characteristic of open-ended conversational tasks, particularly those seen in chatbot applications.

To address this gap, crowdsourced evaluation platforms have emerged, with Chatbot Arena (Chiang et al., 2024; Zheng et al., 2023) standing out as a pioneering framework. By facilitating millions of pairwise comparisons between LLMs based on human judgments, Chatbot Arena has become one of the largest and most credible datasets (Zheng et al., 2024) for chatbot evaluation. Its design more closely reflects the open-ended nature of chatbot usage, providing unparalleled diversity and robustness in assessing model performance. In its first year, the platform amassed over two million votes across more than 150 state-of-the-art models, gaining adoption by leading institutions such as OpenAI, Google, and Hugging Face. This unparalleled scale and impact have solidified Chatbot Arena as a cornerstone in the LLM evaluation ecosystem.

The ranking methodology employed in Chatbot Arena relies on the Elo rating system (Zermelo, 1929; Bradley & Terry, 1952), which is well-suited for competitive settings with clear win-loss outcomes. However, the Elo system does not account for ties—a notable portion of human-judged comparisons—or for modeling deeper relationships between competitors, such as correlations in performance. Addressing these aspects presents an opportunity to build upon the success of Chatbot Arena, enhancing its analytical capabilities while retaining its foundational strengths.

In this paper, we propose a statistical framework that extends the foundational Elo-based approach employed in Chatbot Arena. Our contributions include:

1. **Incorporating ties:** Ties, where two competitors are judged to perform equally, are a key feature of pairwise comparisons. To model ties, we integrated well-established methods by Rao & Kupper (1967) and Davidson (1970), which define tie probabilities based on axiomatic assumptions. While these methods offer a solid foundation for paired comparisons, applying them directly to the Chatbot Arena dataset, with its extensive pairwise comparisons, highlighted the need for further adaptation: these models yielded errors exceeding 10% when applied to tie outcomes.

   To address these challenges, we propose a novel factored tie model, which generalizes these frameworks by uncovering latent structures in tie patterns across pairs of competitors. This factor analysis substantially improves model fit, reducing errors in fitting tie data by two orders of magnitude. Notably, this enhancement also extends to win/loss predictions, achieving comparable improvements. Details of tie modeling, including a background and our generalizations, are provided in Section 2.

2. **Incorporating covariance:** We extend paired comparison models by introducing Thurstonian representations to capture covariance structures between competitors (Section 2.4), enabling deeper exploration of relationships beyond rankings, such as correlations in performance. Building on our theoretical demonstration of covariance's structural non-uniqueness and its equivalency class (Appendices B and C), we show that derived metrics from covariance, such as dissimilarity metrics, are unique and provide interpretable insights. These metrics enable visualization techniques to uncover latent patterns, such as performance trends, and clustering techniques to group competitors into performance tiers based on relative strengths.

3. **Addressing optimization challenges through constraints:** Paired comparison models often exhibit parameter invariances that lead to non-unique solutions during likelihood optimization. These invariances, arising from structural symmetries in the model, result in valid but indistinguishable solutions, compromising convergence stability during parameter estimation. To resolve these issues, we introduce novel constraints that eliminate non-uniqueness by regularizing the parameter space, ensuring stable and interpretable parameter estimation (Section 2.5 and Appendix C.4).

In addition to developing our framework, we conducted comprehensive evaluations and analyses to validate its performance and interpret its results. Empirical evaluations (Section 3) demonstrate the model's fit to observed data, supported by extensive experimentation (Appendix D) on model selection, goodness-of-fit, and prediction metrics. Furthermore, detailed inference analyses (Section 4.1) explore competitor relationships, offering visual insights through clustering and ranking consistency studies (Appendices C and E).

To support reproducibility and broader adoption, we provide `leaderbot`, an open-source Python package implementing our statistical framework with tools for data processing, model fitting, and visualization. This ensures that all results in this paper are fully reproducible (Appendix G).

2 Statistical Model

2.1 Problem Statement

Consider a paired-comparison experiment involving $m \geq 2$ competitors (here, LLM chatbots), indexed by the set $V \coloneqq \{1, \ldots, m\}$. Let $E \subseteq \{\{i, j\} \mid i, j \in V\}$ denote the set of unordered pairs of competitors that have been compared in the experiment. We assume the graph $\mathcal{G}(V, E)$ is connected.

We define the $m \times m$ matrix $\mathbf{W} = [w_{ij}]$, where w_{ij} represents the frequency with which competitor i wins against competitor j, and w_{ji} represents the frequency with which i loses to j. Similarly, we define the symmetric matrix $\mathbf{T} = [t_{ij}]$, where t_{ij} denotes the frequency of ties between competitors i and j, with $t_{ij} = t_{ji}$. We set $w_{ij} = w_{ji} = t_{ij} = 0$ whenever $\{i, j\} \notin E$ to reflect the absence of comparisons between competitors i and j. The total number of comparisons between competitors i and j is denoted by n_{ij}, where $n_{ij} = w_{ij} + w_{ji} + t_{ij}$. The triple $(\mathcal{G}, \mathbf{W}, \mathbf{T})$ constitutes the input data for our problem.

Our objective is to rank the competitors based on their performance in the overall comparisons. To formalize this in a probabilistic framework, we define $P(i \succ j \,|\, \{i,j\})$ as the probability that competitor i wins against competitor j, and $P(i \sim j \,|\, \{i,j\})$ as the probability that i and j tie. For notational simplicity, we often denote these probabilities by $P_{i\succ j}$ and $P_{i\sim j}$, respectively.

A broad class of parametric models (which we will discuss in detail later) assumes the existence of a *score* array $\boldsymbol{x} = (x_1, \ldots, x_m) \in \mathbb{R}^m$, which defines the ranking. Specifically, the ranking is inferred by a bijection from V to itself that orders the scores x_i, such that $x_i > x_j$ implies i is ranked higher than j, denoted by the binary relation $i \succ j$.

The score vector $\boldsymbol{x}$ forms part of the model's parameters, denoted by $\boldsymbol{\theta}$, which also includes other parameters governing the probability of each outcome. A common approach for estimating these parameters is the maximum likelihood method. The likelihood function $\mathcal{L}(\boldsymbol{\theta} \,|\, \mathcal{G}, \mathbf{W}, \mathbf{T})$ is defined as the product of multinomial distributions for each compared pair $\{i,j\} \in E$, given by

$$\mathcal{L}(\boldsymbol{\theta} \,|\, \mathcal{G}, \mathbf{W}, \mathbf{T}) = \prod_{\{i,j\}\in E} \frac{n_{ij}!}{w_{ij}!w_{ji}!t_{ij}!} P_{i\succ j}^{w_{ij}}(\boldsymbol{\theta}) P_{i\prec j}^{w_{ji}}(\boldsymbol{\theta}) P_{i\sim j}^{t_{ij}}(\boldsymbol{\theta}). \tag{1}$$

The parameter estimate $\hat{\boldsymbol{\theta}}$ is then obtained by maximizing the log-likelihood function $\ell(\boldsymbol{\theta}) \coloneqq \log \mathcal{L}(\boldsymbol{\theta})$, i.e., $\hat{\boldsymbol{\theta}} = \operatorname{argmax}_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta})$.

2.2 Probabilistic Models

A parametric model for the above probabilities must satisfy two fundamental axioms. First, by the law of total probability, we have $P_{i\succ j} + P_{i\prec j} + P_{i\sim j} = 1$. Second, the model should respect the concept of transitivity in ranking, though in a probabilistic setting. Rather than standard transitivity, where $i \succ j$ and $j \succ k$ imply $i \succ k$, we adopt the principle of *stochastic transitivity* (see, e.g., Shah et al. (2017); Shah & Wainwright (2018)).

In particular, we are interested in *strong stochastic transitivity*, which states that if $P_{i\succ j} \geq \frac{1}{2}$ and $P_{j\succ k} \geq \frac{1}{2}$, then $P_{i\succ k} \geq \max\{P_{i\succ j}, P_{j\succ k}\}$. A key sub-class of strong stochastic transitivity, and the focus of this work, is *linear stochastic transitivity*. This property is characterized by the existence of an increasing *comparison function* $F : \mathbb{R} \to [0,1]$ and a *merit function* $\zeta : \mathbb{R} \to \mathbb{R}$, such that $P_{i\succ j} = F(\zeta(x_i) - \zeta(x_j))$.

In the following subsections, we describe several common models of paired comparison that satisfy these properties.

2.2.1 Bradley-Terry Model

One of the most widely used models for paired comparison was first introduced by Zermelo (1929) and later rediscovered by Bradley & Terry (1952), leading to the model being named after them. The Bradley-Terry model forms the basis of the well-known Elo rating system, which is extensively used by the World Chess Federation. In this model, the probabilities of win and loss are given by

$$P(i \succ j \,|\, \{i,j\}) \coloneqq \frac{\pi_i}{\pi_i + \pi_j} \quad \text{and} \quad P(i \prec j \,|\, \{i,j\}) \coloneqq \frac{\pi_j}{\pi_i + \pi_j}, \tag{2}$$

where $\pi_i \coloneqq e^{x_i}$. This model assumes that $x_i - x_j$ follows a logistic distribution, as shown by

$$P(x_i - x_j > 0) = \frac{1}{1 + e^{-(x_i - x_j)}}. \tag{3}$$

Variants of this model, such as the one proposed by Glenn & David (1960), assume a standard normal distribution instead of the logistic. However, the logistic distribution is typically preferred in paired comparison settings due to its heavier tails, which provide better fit for real-world data, and its computational advantages and tractability (Böckenholt, 2001).

2.2.2 Models with Ties

The Bradley-Terry (BT) model does not account for ties (i.e., $P_{i\sim j} = 0$), making it suitable only for balanced paired comparisons, such as zero-sum games. However, in our application of ranking

LLM chatbot agents, ties frequently occur, rendering the BT model inadequate for capturing these outcomes.

A workaround used in previous work, such as Chiang et al. (2024) for ranking LLM models, is to treat a tie as halfway between a win and a loss, modifying the outcome matrix as $\mathbf{W} \leftarrow \mathbf{W} + \frac{1}{2}\mathbf{T}$. However, more sophisticated extensions of the BT model have been developed, incorporating ties based on proper axioms.

One such generalization is the model of Rao & Kupper (1967), which modifies the logistic distribution to account for ties. The resulting probabilities for win, loss, and tie are given by

$$P(i \succ j \,|\, \{i,j\}) = P(x_i - x_j > \eta) \coloneqq \frac{\pi_i}{\pi_i + \nu\pi_j}, \tag{4a}$$

$$P(i \prec j \,|\, \{i,j\}) = P(x_i - x_j < -\eta) \coloneqq \frac{\pi_j}{\nu\pi_i + \pi_j}, \tag{4b}$$

$$P(i \sim j \,|\, \{i,j\}) = P(|x_i - x_j| < \eta) \coloneqq \frac{\pi_i\pi_j(\nu^2 - 1)}{(\pi_i + \nu\pi_j)(\nu\pi_i + \pi_j)}, \tag{4c}$$

where $\nu \coloneqq e^{\eta} \geq 1$ and $\eta > 0$ is a threshold parameter to be optimized. In this model, if the difference between competitors' scores is less than the threshold η, the judge is unable to distinguish between competitors i and j, resulting in a tie. Setting $\eta = 0$ reduces this model to the standard BT model.

Another extension of the BT model, introduced by Davidson (1970) and based on the axiom of choice of Luce (1959), models ties differently:

$$P(i \succ j \,|\, \{i,j\}) = \frac{\pi_i}{\pi_i + \pi_j + \nu\sqrt{\pi_i\pi_j}}, \tag{5a}$$

$$P(i \prec j \,|\, \{i,j\}) = \frac{\pi_j}{\pi_i + \pi_j + \nu\sqrt{\pi_i\pi_j}}, \tag{5b}$$

$$P(i \sim j \,|\, \{i,j\}) = \frac{\nu\sqrt{\pi_i\pi_j}}{\pi_i + \pi_j + \nu\sqrt{\pi_i\pi_j}}, \tag{5c}$$

where $\nu \coloneqq e^{\eta}$ and $\eta \in \mathbb{R}$ is a threshold parameter for tie to be optimized. Here, setting $\eta = -\infty$ reduces this model to the BT model. Note that in both the Rao-Kupper and Davidson models, the probability of a tie increases as the difference in scores decreases.

2.3 A Generalization for Models with Ties

The original Rao-Kupper and Davidson models each employ a single parameter for ties, ν. However, as our numerical results will demonstrate, one tie parameter is insufficient to capture the complexity of ties across all pairs of competitors. To address this, we propose a generalization of these models by incorporating additional parameters, which, to our knowledge, is a novel extension.

In this generalized model, the tie parameter $\nu = e^{\eta}$ is replaced with a pair-specific parameter $\nu_{ij} = e^{\eta_{ij}}$, where $i, j \in V$, subject to the symmetry condition $\nu_{ij} = \nu_{ji}$. This modification introduces $|E|$ parameters, potentially leading to overfitting. To mitigate this, we propose a reduced model. Let the symmetric $m \times m$ matrix $\mathbf{H} = [\eta_{ij}]$ represent the pairwise tie parameters. Rather than treating all η_{ij} as independent parameters, we introduce the following factor model to construct $\mathbf{H}$ by

$$\mathbf{H}_k \coloneqq \begin{cases} \mathbf{G}\mathbf{\Phi}^{\intercal} + \mathbf{\Phi}\mathbf{G}^{\intercal}, & k \in V, \\ \eta\mathbf{J}, & k = 0, \end{cases} \tag{6}$$

where $\mathbf{G} = [g_{ij}]$ is an $n \times k$ matrix of the new parameters g_{ij}, $\mathbf{\Phi} = [\phi_{ij}]$ is an $m \times k$ constant matrix of rank k consisting of predefined basis vectors that will be discussed below, and $\mathbf{J}$ is the $m \times m$ matrix of all ones. The rank of $\mathbf{H}_k$ is $\max(1, \min(2k, m))$ containing $\max(1, mk)$ parameters, thus, choosing k allows us to strike a balance between the goodness of fit and the complexity of the model. Note that the case $k = 0$ reverts to the original Rao-Kupper or Davidson models where a single parameter $\eta_{ij} = \eta$ is used.

To ensure $\mathbf{\Phi}$ is of full rank, we design $\mathbf{\Phi}$ with orthogonal column vectors. This can be achieved, for instance, by orthogonalizing a randomly generated matrix using Gram-Schmidt orthogonalization. However, for reproducibility and to avoid using randomly generated matrices, we recommend constructing $\mathbf{\Phi}$ using discrete orthogonal polynomials with respect to a uniform weight over equally spaced points (Baik et al., 2007), which are inherently orthogonal. Such matrices can be generated, for instance, by discrete Legendre polynomials, the Hadamard transform (when m is a power of 2), discrete Chebyshev polynomials (Corr et al., 2000), or the discrete cosine transform (DCT). For simplicity, we choose the discrete cosine transform of the fourth type (DCT-IV) basis, where the elements ϕ_{ij} are given by

$$\phi_{ij} = \sqrt{\frac{2}{m}} \cos\left(\frac{\pi}{4m}(2i-1)(2j-1)\right), \quad i = 1, \ldots, m, \quad \text{and} \quad j = 1, \ldots, k. \tag{7}$$

2.4 Thurstonian Representation of Models

A fundamental approach to modeling paired comparisons was introduced by Thurstone (1927) through the laws of comparative judgment, laying the foundation for psychometric choice modeling from a statistical perspective. Here, we briefly describe Thurstone's multivariate discriminal process and apply it to the models discussed earlier.

Thurstonian models assume that the score variables $\boldsymbol{x}$ are stochastic processes defined by $\boldsymbol{x} = \boldsymbol{\mu} + \boldsymbol{\epsilon}$, where $\boldsymbol{\mu} \in \mathbb{R}^m$ is the mean, and $\boldsymbol{\epsilon}$ is a zero-mean random component with covariance $\mathbf{\Sigma} = [\sigma_{ij}]$, often referred to as the *comparative dispersion*.

The difference between scores, which is central to the previously mentioned models, also becomes a stochastic process: $x_i - x_j = \mu_i - \mu_j + \epsilon_{ij}$, where ϵ_{ij} has covariance $\mathbf{S} = [s_{ij}]$, given by $s_{ij} = \sigma_{ii} + \sigma_{jj} - 2\sigma_{ij}$.[1] This is commonly referred to as the *discriminal dispersion* (Heiser & de Leeuw, 1981).

The original Thurstonian model assumes that $\boldsymbol{x}$ follows a joint normal distribution, meaning that $x_i - x_j$ also has a normal distribution. Let $x_{ij} \coloneqq x_i - x_j$, $\mu_{ij} \coloneqq \mu_i - \mu_j$, and $y_{ij} \coloneqq (x_{ij} - \mu_{ij})/\sqrt{s_{ij}}$. It can be shown that $P_{i \succ j} = P(x_{ij} > 0) = \Phi(z_{ij})$ (Maydeu-Olivares, 1999), where Φ is the cumulative standard normal distribution, and

$$z_{ij} \coloneqq \frac{\mu_i - \mu_j}{\sqrt{s_{ij}}}. \tag{8}$$

Although the Thurstonian discriminal process is typically applied using normal distributions, we extend this process to the models previously introduced, including the Bradley-Terry model (which uses the logistic distribution) and the Rao-Kupper and Davidson models (which incorporate ties). To our knowledge, these extensions have not been explored in the literature.

To generalize the Bradley-Terry model with the discriminal process, we assume y_{ij} follows a logistic distribution, yielding

$$P(i \succ j \,|\, \{i, j\}) = \frac{1}{1 + e^{-z_{ij}}}, \quad \text{and} \quad P(i \prec j \,|\, \{i, j\}) = \frac{1}{1 + e^{-z_{ji}}}. \tag{9}$$

We can similarly extend the Rao-Kupper and Davidson models. The probabilities in equations (4) and (5) can be expressed in terms of the logit function $\log(\pi_i/\pi_j) = x_{ij}$, which represents the quantile of the logistic distribution. To incorporate the Thurstonian model, we replace x_{ij} with z_{ij}. Thus, the Rao-Kupper model becomes

$$P(i \succ j \,|\, \{i, j\}) = \frac{1}{1 + e^{-(z_{ij} - \eta_{ij})}}, \tag{10a}$$

$$P(i \prec j \,|\, \{i, j\}) = \frac{1}{1 + e^{-(z_{ji} - \eta_{ij})}}, \tag{10b}$$

$$P(i \sim j \,|\, \{i, j\}) = \frac{e^{\eta_{ij}} - 1}{\left(1 + e^{-(z_{ij} - \eta_{ij})}\right)\left(1 + e^{-(z_{ij} - \eta_{ij})}\right)}. \tag{10c}$$

[1] This is due to the fact that $\operatorname{var}(x_i - x_j) = \operatorname{var}(x_i) + \operatorname{var}(x_j) - 2\operatorname{cov}(x_i, x_j)$.

Similarly, the Davidson model becomes

$$P(i \succ j \,|\, \{i,j\}) = \frac{1}{1 + e^{-z_{ij}} + e^{-(\frac{1}{2}z_{ij} - \eta_{ij})}}, \tag{11a}$$

$$P(i \prec j \,|\, \{i,j\}) = \frac{1}{1 + e^{-z_{ji}} + e^{-(\frac{1}{2}z_{ji} - \eta_{ij})}}, \tag{11b}$$

$$P(i \sim j \,|\, \{i,j\}) = \frac{1}{1 + e^{-(\frac{1}{2}z_{ij} - \eta_{ij})} + e^{-(\frac{1}{2}z_{ji} - \eta_{ij})}}. \tag{11c}$$

These models introduce an additional $\binom{m}{2}$ covariance parameters σ_{ij}, which can lead to overparameterization. To address this, Thurstone (1927) proposed various constraints on the covariance matrix, while Takane (1989) suggested a factor model for covariance structure analysis. We adopt the factor model employed by Böckenholt (1993); Maydeu-Olivares & Böckenholt (2005), where the covariance is constructed as

$$\boldsymbol{\Sigma} = \mathbf{D} + \boldsymbol{\Lambda}\boldsymbol{\Lambda}^{\intercal}, \tag{12}$$

where $\mathbf{D} = [d_{ij}]$ is a positive-definite diagonal matrix with $d_{ii} > 0$, and $\boldsymbol{\Lambda} = [\lambda_{ij}]$ is an $m \times k$ matrix of rank $k \leq n$, containing the factor parameters λ_{ij}. By selecting k, we can balance model fit with complexity.

2.5 Imposing Constraints to Resolve Symmetries

When estimating the parameters of the models (such as $\boldsymbol{\theta} = (\boldsymbol{\mu}, \mathbf{G}, \mathbf{D}, \boldsymbol{\Lambda})$) through the maximum likelihood method described in (1), we encounter computational issues due to the non-uniqueness of the solution. This arises from the fact that the likelihood function remains invariant under certain transformations of the parameters, known as symmetries. These symmetries can result in poor optimization behavior, such as large or small parameter values, causing instability in the estimation process. To address these issues, we propose constraints on the log-likelihood function to eliminate the problematic symmetries. While one of these symmetries has been addressed in prior work, the other two have not, to the best of our knowledge.

The first symmetry relates to the fact that the models are invariant under the transformation $x_i \mapsto x_i + c$, where $c \in \mathbb{R}$ is a constant. This means that the absolute values of x_i are not identifiable, only their differences matter. As a result, the parameters can shift without changing the model's predictions. Common approaches to address this symmetry include fixing one of the score parameters, as done in Chiang et al. (2024), or imposing other constraints on the parameter values. For instance, the original Bradley-Terry model (Bradley & Terry, 1952) imposes the constraint $\sum_{i=1}^{m} \pi_i = 1$. In our work, we adopt a similar constraint by fixing the mean of the scores: $\sum_{i=1}^{m} \mu_i = 0$. This natural choice effectively eliminates the shift invariance, ensuring stable optimization of the score parameters $\boldsymbol{x}$.

The second symmetry, which has not been addressed in previous studies, involves scaling both the score and covariance parameters. Specifically, the models remain invariant under the transformation $(x_i, \sigma_{ij}) \mapsto (tx_i, t^2\sigma_{ij})$ for any positive constant t, because the ratio z_{ij} remains unchanged. This symmetry can lead to parameters collapsing to very small values, causing numerical instability or underflow during optimization. To resolve this, we introduce the following constraint

$$\operatorname{trace}(\tilde{\boldsymbol{\Sigma}}) = 1, \tag{13}$$

where $\tilde{\boldsymbol{\Sigma}} \coloneqq \mathbf{P}\boldsymbol{\Sigma}\mathbf{P}$, and $\mathbf{P} \coloneqq \mathbf{I} - \frac{1}{m}\mathbf{1}\mathbf{1}^{\intercal}$ is the centering operator with $\mathbf{I}$ as the identity matrix and $\mathbf{1} \coloneqq (1, \ldots, 1)^{\intercal}$ is a column vector of ones. This constraint ensures proper scaling of the covariance parameters and avoids collapse during optimization. A detailed explanation of this constraint is provided in Appendix C.4.

The third symmetry pertains to the factor model parameters λ_{ij}. Using the factor model for covariance in (12), we can express the elements of the matrix $\mathbf{S}$ as

$$s_{ii} = 0, \quad \text{and} \quad s_{ij} = d_{ii} + d_{jj} + \|\boldsymbol{\lambda}_i - \boldsymbol{\lambda}_j\|_2^2, \quad i \neq j,$$

where $\boldsymbol{\lambda}_i \coloneqq (\lambda_{i1}, \ldots, \lambda_{ik})$ is the i-th row of the matrix $\boldsymbol{\Lambda}$, and $\|\cdot\|_2$ denotes the Euclidean norm. This expression reveals an invariance under translation $\lambda_{ij} \mapsto \lambda_{ij} + c'$, which has not been addressed

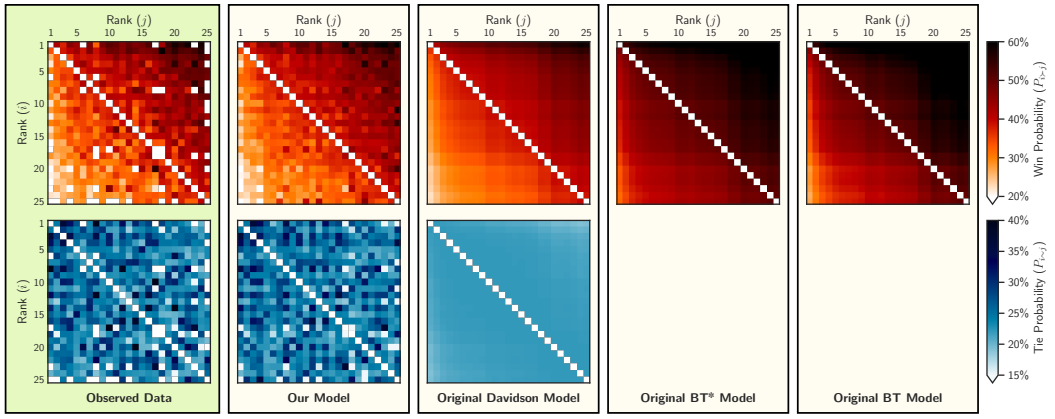

Figure 1: Comparison of pair-specific win (first row) and tie (second row) probabilities among 25 competitors between observed data (first column) and model predictions: our generalized Rao-Kupper with factored tie model (second column), original Rao-Kupper with ties (third column), Bradley-Terry with ties as half win/loss (fourth column, (Chiang et al., 2024)), and original Bradley-Terry without ties (fifth column). The ordinate and abscissa are shared across panels, shown only for the left-most and top-most panels. Each row shares the same color range, with a single colorbar per row.

in the literature. To eliminate this translation symmetry, we impose the constraint

$$\|\mathbf{\Lambda}^\intercal \mathbf{1}\|_2^2 = 0. \tag{14}$$

This constraint fixes the column-wise mean of $\mathbf{\Lambda}$ to zero, preventing the factor parameters from arbitrarily shifting.

3 Empirical Evaluation of Statistical Models

In this section, we apply the statistical models introduced earlier to the dataset from Chatbot Arena. As of September 2024, the dataset consists of $m = 129$ competitors, with $|E| = 3455$ unique pairs that have been compared. The total number of comparisons across all pairs is $\sum_{\{i,j\}\in E} n_{ij} = 1,374,996$, distributed as follows: 43.3% wins, 36.2% losses, and 20.4% ties.

3.1 Evaluating the Prediction of Win/Loss and Tie Matrices

We visualize a subset of the win/loss matrix $\mathbf{W}$ and tie matrix $\mathbf{T}$ corresponding to the top 25 models ranked by Model 18. Figure 1 presents these visualizations: the first and second rows show the win/loss matrix $\mathbf{W}$, while the third row shows the tie matrix $\mathbf{T}$. The leftmost column contains the matrices derived from observed data, while the second and third columns show matrices predicted by Models 18, 7, 4, and 1. Note that since the BT models (Models 4 and 1) do not account for ties, their corresponding tie matrix $\mathbf{T}$ is absent.

We visualize a subset of the win/loss matrix $\mathbf{W}$ and tie matrix $\mathbf{T}$ corresponding to the top 25 models ranked by our generalized Rao-Kupper model with factored ties (second column in Figure 1), the original Rao-Kupper model with ties (third column), Bradley-Terry with ties treated as half win/loss (Chiang et al., 2024) (fourth column), and the original Bradley-Terry model without ties (fifth column). The first row shows win probabilities, while the second row shows tie probabilities. The leftmost column contains the matrices derived from observed data. Note that since the Bradley-Terry models do not account for ties, their corresponding tie probabilities are not shown.

We observe that the generalized Rao-Kupper model with factored ties ($k = 20$) demonstrates a strong resemblance between the predicted and observed matrices for both win/loss and tie outcomes. By contrast, the original Rao-Kupper model with a single tie parameter performs reasonably well in predicting the win matrix but lacks accuracy in the tie matrix. Meanwhile, the Bradley-Terry

models (both with ties treated as half win/loss and without ties) produce noticeably different win/loss matrices, as they fail to account for ties, leading to discrepancies in their predictions relative to the observed data.

4 RANKING AND COMPARISON OF LLM-BASED CHATBOTS

In this section, we examine how the proposed models provide insights into the competitive performance of chatbots. In Section 4.1, we focus on analyzing the ranking of chatbots, while in Section 4.2 we explore the correlations between their performances.

4.1 RANKING

In pairwise comparison methods, the score parameters $\boldsymbol{x}$ are used to rank competitors. As an example, Figure E.1 in Appendix E illustrates the score parameters for the top 50 chatbots, ranked according to Model 30 (see Table D.1), which is based on the Davidson model with a tie factor model of rank $k = 20$.

An important question is how different statistical models affect the chatbot rankings. To explore this, we analyzed rankings produced by a selection of 12 models from Table D.1. Figure E.2 in Appendix E presents a bump chart that visually compares these rankings. In the chart, each column corresponds to a statistical model, and each row represents a chatbot's ranking position. The chatbots, listed by their abbreviated names on the left, are ranked by Model 30 in the leftmost column. Each line in the chart tracks how a chatbot's ranking changes across different models, with colors used to distinguish individual paths.

The 12 models are arranged in increasing order of complexity, moving from right to left. In the first tier, Models 1, 4, 7, and 19 (shown in the rightmost columns) represent the original BT, Rao-Kupper, and Davidson models without any of our proposed generalizations. In the second tier, Models 3 and 6 extend the basic BT models by incorporating Thurstonian covariance with $k = 0$. Models 10 and 22 further extend the Rao-Kupper and Davidson models by introducing tie factor parameters with $k = 20$. Models 14 and 26 include a Thurstonian covariance factor with $k = 3$, while the most complex models, 18 and 30, combine $k = 3$ and $k = 20$ for covariance and tie factors, respectively.

The bump chart reveals a high degree of consistency among all models for the highest-ranked chatbots, indicating that these models, regardless of complexity, produce stable rankings for top performers. However, discrepancies become more apparent at the lower end of the rankings, where models diverge more significantly. Notably, similar models—especially those with comparable complexity levels—tend to produce more consistent rankings across the chart, though exceptions do occur.

Further analyses of the rankings, including Kendall's rank correlation and additional visualizations, are provided in Appendix E in the appendix. These analyses explore how ranking outcomes vary across models and identify key model parameters that influence these variations.

4.2 EXPLORING COMPETITOR CORRELATIONS

One of the advantages of incorporating Thurstonian models into our extended framework is the ability to extract insights beyond simple rankings, particularly by revealing correlations between competitors. While these models significantly enhance fit, they also allow us to explore relationships that would otherwise remain hidden. However, caution must be exercised when interpreting these covariance parameters, as their individual values do not offer straightforward insights.

To clarify, just as the absolute values of the score parameters x_i hold no direct meaning—their differences between competitors being the key factor—so too is the case for the covariances. Recall from the Thurstonian representation that $x_i = \mu_i + \epsilon_i$, where ϵ_i represents the stochastic component with covariance σ_{ij}. The value of σ_{ii} alone does not convey any specific uncertainty about x_i. Instead, the interpretable quantity is the covariance of the difference $x_{ij} = x_i - x_j = \mu_i - \mu_j + \epsilon_{ij}$, which is given by $s_{ij} = \sigma_{ii} + \sigma_{jj} - 2\sigma_{ij}$. The matrix $\mathbf{S}$, consisting of these pairwise covariances s_{ij}, reveals the relationships between competitors and is commonly interpreted in the paired comparison literature (Maydeu-Olivares & Böckenholt, 2005; Böckenholt, 2006).

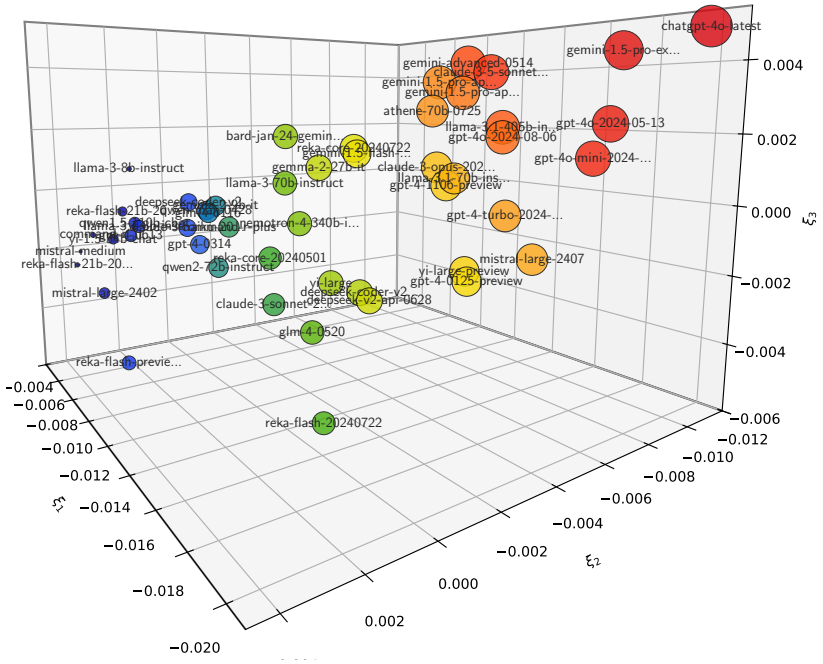

Figure 2: Kernel PCA projection of the distance matrix $\mathbf{Z}$ onto three dimensions, focusing on the top 40 competitors ranked by Model 18 of Table D.1. In the scatter plot, circle size and color are proportional to the competitors' scores.

To visualize these relationships, we use kernel PCA to project the data into a three-dimensional space, enabling more effective interpretation of the distances between competitors. The matrix $\mathbf{Z} = [z_{ij}]$, where z_{ij} is defined by (8), serves as the distance metric, normalizing score differences by $\sqrt{s_{ij}}$. We apply a squared exponential kernel, $\rho_{ij} = \exp(-\gamma z_{ij}^2)$, with $\gamma = 10^{-4}$, and project the data into three dimensions. Figure 2 shows a scatter plot of the top 40 chatbots, ranked by model 18, with circle size and color proportional to their scores. In this plot, the relative distances between points are meaningful, rather than their specific coordinates or orientation. This spatial configuration reveals how closely related the chatbots are, offering insights beyond ranking alone.

It is important to note that such visual representations of chatbot relationships are only possible with models that include Thurstonian representations and covariance structures. These models provide a deeper understanding of how competitors are related, extending beyond traditional ranking to offer insights into their underlying correlations.

5 Conclusion

Effectively evaluating large language models (LLMs) in pairwise comparison settings is crucial for understanding their strengths and limitations in real-world applications. This paper presents a novel statistical framework that extends traditional approaches, incorporating ties, covariance, and advanced optimization techniques to enhance interpretability, stability, and accuracy in LLM evaluation.

Our generalized tie model addresses long-standing limitations of Rao-Kupper and Davidson methods, reducing prediction errors by two orders of magnitude and improving fit across win, loss, and tie outcomes. By introducing covariance structures, our framework uncovers latent relationships among competitors, enabling clustering into performance tiers and enriching interpretability beyond rankings. Additionally, we resolve optimization challenges through novel constraints, ensuring stable and unique parameter estimation.

A key insight from our analysis is that covariance structures not only enrich interpretability but also shape ranking consistency, underscoring their critical role in capturing nuanced relationships in

pairwise comparisons. This finding highlights the broader utility of covariance-based approaches in complex evaluation tasks.

To validate our framework, we conducted rigorous empirical evaluations, demonstrating significant improvements in model fit and interpretability compared to existing methods. To support reproducibility and broader adoption, we provide `leaderbot`, an open-source Python package implementing our statistical framework with tools for data processing, model fitting, and visualization.

Our work not only advances LLM evaluation but also provides a robust foundation for analyzing pairwise comparison data in diverse domains. By bridging theoretical rigor with practical applicability, this framework opens new avenues for ranking, inference, and interpretability in complex comparison tasks.

REFERENCES

Baik, J., Kriecherbauer, T., McLaughlin, K. T.-R., & Miller, P. D. (2007). *Discrete Orthogonal Polynomials. (AM-164): Asymptotics and Applications (AM-164)*. Princeton University Press.

Bar-Joseph, Z., Gifford, D. K., & Jaakkola, T. S. (2001). Fast optimal leaf ordering for hierarchical clustering. *Bioinformatics*, 17(suppl_1), S22–S29.

Böckenholt, U. (1993). Applications of thurstonian models to ranking data. In M. A. Fligner & J. S. Verducci (Eds.), *Probability Models and Statistical Analyses for Ranking Data* (pp. 157–172). New York, NY: Springer New York.

Böckenholt, U. (2001). Thresholds and intransitivities in pairwise judgments: A multilevel analysis. *Journal of Educational and Behavioral Statistics*, 26(3), 269–282.

Böckenholt, U. (2006). Visualizing individual differences in pairwise comparison data. *Food Quality and Preference*, 17(3), 179–187. Seventh Sensometrics Meeting, Davis, USA, 28–30 July 2004.

Bradley, R. A. & Terry, M. E. (1952). Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3-4), 324–345.

Chen, M., Tworek, J., Jun, H., Yuan, Q., de Oliveira Pinto, H. P., Kaplan, J., Edwards, H., Burda, Y., Joseph, N., Brockman, G., Ray, A., Puri, R., Krueger, G., Petrov, M., Khlaaf, H., Sastry, G., Mishkin, P., Chan, B., Gray, S., Ryder, N., Pavlov, M., Power, A., Kaiser, L., Bavarian, M., Winter, C., Tillet, P., Such, F. P., Cummings, D., Plappert, M., Chantzis, F., Barnes, E., Herbert-Voss, A., Guss, W. H., Nichol, A., Paino, A., Tezak, N., Tang, J., Babuschkin, I., Balaji, S., Jain, S., Saunders, W., Hesse, C., Carr, A. N., Leike, J., Achiam, J., Misra, V., Morikawa, E., Radford, A., Knight, M., Brundage, M., Murati, M., Mayer, K., Welinder, P., McGrew, B., Amodei, D., McCandlish, S., Sutskever, I., & Zaremba, W. (2021). Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*.

Chiang, W. L., Zheng, L., Sheng, Y., Angelopoulos, A. N., Li, T., Li, D., Zhu, B., Zhang, H., Jordan, M., Gonzalez, J. E., & Stoica, I. (2024). Chatbot Arena: An open platform for evaluating LLMs by human preference. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research* (pp. 8359–8388).: PMLR.

Corr, P., Smith, F., Hanna, P., Stewart, D., & Ming, J. (2000). Discrete Chebyshev transform - a natural modification of the DCT. In *Pattern Recognition, International Conference on*, volume 3 (pp. 7154). Los Alamitos, CA, USA: IEEE Computer Society.

Davidson, R. R. (1970). On extending the Bradley-Terry model to accommodate ties in paired comparison experiments. *Journal of the American Statistical Association*, 65(329), 317–328.

Elo, A. E. (1978). *The Rating of Chessplayers, Past and Present*. New York: Arco Pub.

Epoch AI (2022). Parameter, compute and data trends in machine learning. Accessed: 2024-11-22.

Gibbons, J. & Chakraborti, S. (2003). *Nonparametric Statistical Inference*. Statistics, textbooks and monographs. Marcel Dekker Incorporated, fourth edition: revised and expanded edition.

Glenn, W. A. & David, H. A. (1960). Ties in paired-comparison experiments using a modified Thurstone-Mosteller model. *Biometrics*, 16(1), 86–109.

Glickman, M. E. (1995). A comprehensive guide to chess ratings. *American Chess Journal*, 3(1), 59–102.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, 2 edition.

Heiser, W. J. & de Leeuw, J. (1981). Multidimensional mapping of preference data. *Mathématiques et Sciences Humaines*, 73, 39–96.

Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., & Steinhardt, J. (2021). Measuring massive multitask language understanding. In *International Conference on Learning Representations*.

Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., de Las Casas, D., Hendricks, L. A., Welbl, J., Clark, A., Hennigan, T., Noland, E., Millican, K., van den Driessche, G., Damoc, B., Guy, A., Osindero, S., Simonyan, K., Elsen, E., Vinyals, O., Rae, J. W., & Sifre, L. (2022). Training compute-optimal large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22 Red Hook, NY, USA: Curran Associates Inc.

Hopkins, D. J. & Noel, H. (2022). Trump and the shifting meaning of "Conservative": Using activists' pairwise comparisons to measure politicians' perceived ideologies. *American Political Science Review*, 116(3), 1133–1140.

Hunter, D. R. (2004). MM algorithms for generalized Bradley-Terry models. *The Annals of Statistics*, 32(1), 384 – 406.

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., & Amodei, D. (2020). Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.

Karthik, S., Coskun, H., Akata, Z., Tulyakov, S., Ren, J., & Kag, A. (2024). Scalable ranked preference optimization for text-to-image generation. *arXiv preprint arXiv:2410.18013*.

Kendall, M. G. (1938). A new measure of rank correlation. *Biometrika*, 30(1/2), 81–93.

Kendall, M. G. (1945). The treatment of ties in ranking problems. *Biometrika*, 33(3), 239–251.

Liu, J., Ge, D., & Zhu, R. (2024). Reward learning from preference with ties. *arXiv preprint arXiv:2410.05328*.

Loewen, P. J., Rubenson, D., & Spirling, A. (2012). Testing the power of arguments in referendums: A Bradley–Terry approach. *Electoral Studies*, 31(1), 212–221. Special Symposium: Germany's Federal Election September 2009.

Luce, R. D. (1959). *Individual choice behavior.* Individual choice behavior. Oxford, England: John Wiley.

Mardia, K., Kent, J., & Bibby, J. (1979). *Multivariate Analysis*. Probability and Mathematical Statistics : a series of monographs and textbooks. Academic Press.

Maydeu-Olivares, A. (1999). Thurstonian modeling of ranking data via mean and covariance structure analysis. *Psychometrika*, 64(3), 325–340.

Maydeu-Olivares, A. & Böckenholt, U. (2005). Structural equation modeling of paired-comparison and ranking data. *Psychological Methods*, 10(3), 285–304.

Newman, M. E. J. (2023). Efficient computation of rankings from pairwise comparisons. *Journal of Machine Learning Research*, 24(238), 1–25.

Rafailov, R., Sharma, A., Mitchell, E., Ermon, S., Manning, C. D., & Finn, C. (2024). Direct preference optimization: your language model is secretly a reward model. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23 Red Hook, NY, USA: Curran Associates Inc.

Rao, P. V. & Kupper, L. L. (1967). Ties in paired-comparison experiments: A generalization of the Bradley-Terry model. *Journal of the American Statistical Association*, 62(317), 194–204.

Rothenberg, T. J. (1971). Identification in parametric models. *Econometrica*, 39(3), 577–591.

Schölkopf, B., Smola, A. J., & Müller, K.-R. (1999). *Kernel principal component analysis*, (pp. 327–352). MIT Press: Cambridge, MA, USA.

Seber, G. & Wild, C. (2005). *Nonlinear Regression*. Wiley Series in Probability and Statistics. Wiley.

Shah, N. B., Balakrishnan, S., Guntuboyina, A., & Wainwright, M. J. (2017). Stochastically transitive models for pairwise comparisons: Statistical and computational issues. *IEEE Trans. Inf. Theor.*, 63(2), 934–959.

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

Söderström, T. & Stoica, P. (1989). *System Identification*. Prentice-Hall Software Series. Prentice Hall.

Takane, Y. (1989). Analysis of covariance structures and probabilistic binary choice data. In G. de Soete, H. Feger, & K. C. Klauer (Eds.), *New Developments in Psychological Choice Modeling*, volume 60 of *Advances in Psychology* (pp. 139–160). North-Holland.

Thurstone, L. L. (1927). A law of comparative judgement. *Psychological Review*, 34, 278–286.

Williams, C. (2000). On a connection between kernel PCA and metric multidimensional scaling. In T. Leen, T. Dietterich, & V. Tresp (Eds.), *Advances in Neural Information Processing Systems*, volume 13: MIT Press.

Wu, T., Zhu, B., Zhang, R., Wen, Z., Ramchandran, K., & Jiao, J. (2023). Pairwise proximal policy optimization: Harnessing relative feedback for LLM alignment. *arXiv preprint arXiv:2310.00212*.

Wu, T., Zhu, B., Zhang, R., Wen, Z., Ramchandran, K., & Jiao, J. (2024). Pairwise proximal policy optimization: Language model alignment with comparative RL. In *First Conference on Language Modeling*.

Zermelo, E. (1929). Die berechnung der turnier-ergebnisse als ein maximumproblem der wahrscheinlichkeit-srechnung. *Mathematische Zeitschrift*, 29(1), 436–460.

Zheng, L., Chiang, W.-L., Sheng, Y., Li, T., Zhuang, S., Wu, Z., Zhuang, Y., Li, Z., Lin, Z., Xing, E., Gonzalez, J. E., Stoica, I., & Zhang, H. (2024). LMSYS-Chat-1M: A large-scale real-world LLM conversation dataset. In *The Twelfth International Conference on Learning Representations*.

Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E., et al. (2023). Judging LLM-as-a-judge with MT-bench and Chatbot Arena. *Advances in Neural Information Processing Systems*, 36, 46595–46623.

Appendices

CONTENTS

APPENDIX A BROADER IMPLICATIONS OF PAIRED COMPARISON METHODS

Paired comparison frameworks are foundational in many fields and have been widely applied in classical and modern domains. In sports analytics, they are used to predict match outcomes and assess player performance in games such as chess (Elo, 1978), tennis (Glickman, 1995), and soccer. Marketing applications include optimizing product offerings, advertisement placements, and analyzing

consumer preferences. In psychometrics and behavioral studies, paired comparisons assess perception and attitudes in response to visual or auditory stimuli. Similarly, in election studies and political science, they are employed to rank candidates, analyze voting behavior, test referendum arguments (Loewen et al., 2012), and measure perceived political ideologies (Hopkins & Noel, 2022). Clinical research also uses paired comparisons to evaluate treatments and interventions in clinical trials and epidemiological studies. These diverse applications illustrate the versatility of paired comparison frameworks in extracting meaningful inferences from comparative data.

Recent advancements have extended paired comparison methods to machine learning, where they play a pivotal role in preference modeling and optimization. For instance, Reinforcement Learning with Human Feedback (RLHF) uses paired comparisons to fine-tune large language models (LLMs) by ranking outputs based on human preferences, often employing the Bradley-Terry model for preference quantification (Rafailov et al., 2024; Karthik et al., 2024). Direct Preference Optimization (DPO) further refines this approach by aligning model outputs directly with human preferences without relying on scalar reward models (Wu et al., 2023). Additionally, methodologies like Pairwise Proximal Policy Optimization (P3O) leverage relative feedback to enhance LLM alignment (Wu et al., 2024). Innovations such as the integration of Rao-Kupper models have enabled paired comparison frameworks to incorporate ties, capturing ambiguous or neutral preferences in RLHF settings (Liu et al., 2024). These developments highlight the growing influence of paired comparison methods in machine learning and underscore the potential of our generalizations to enhance these frameworks further.

APPENDIX B UNIDENTIFIABILITY OF PARAMETERS IN PAIRED COMPARISON MODELS

In this section, we discuss the challenge of estimating the uncertainties of scores x_i in paired comparison models. Previous works, such as Chiang et al. (2024), have introduced methods for computing confidence intervals for scores using empirical approaches like bootstrapping. While these methods are valuable in practice, we demonstrate that this problem is intrinsically ill-posed due to the unidentifiability of the score parameters. Specifically, the likelihood function depends only on score differences $x_i - x_j$, rendering individual score estimates invariant under certain transformations. This invariance introduces non-uniqueness in the quantification of confidence intervals.

This issue arises not from specific modeling choices but from a structural characteristic of models based on strong stochastic transitivity, where probabilities take the form $F(x_i - x_j)$ (see Section 2.2). To analyze this limitation rigorously, we examine the Fisher Information Matrix (FIM) of the likelihood function. In Appendix B.1, we provide a background on the FIM and its role in parameter identifiability. In Appendix B.2, we show that the score parameters are unidentifiable due to the singularity of the FIM. Finally, in Appendix B.3, we propose reparameterizations that result in identifiable quantities.

B.1 FISHER INFORMATION AND IDENTIFIABILITY: BACKGROUND

The Fisher Information Matrix (FIM) quantifies the amount of information that the likelihood function carries about the parameters of interest. It can be derived from the gradient of the log-likelihood function, known as the informant vector,[2] or equivalently, from the negative Hessian matrix of the log-likelihood:

$$\mathbf{F}(\boldsymbol{\theta}) \coloneqq \mathbb{E}\left[\nabla_{\boldsymbol{\theta}}\ell(\boldsymbol{\theta}) \otimes \nabla_{\boldsymbol{\theta}}\ell(\boldsymbol{\theta}) \,|\, \boldsymbol{\theta}\right] = -\mathbb{E}\left[\nabla_{\boldsymbol{\theta}}\nabla_{\boldsymbol{\theta}}^{\intercal}\ell(\boldsymbol{\theta}) \,|\, \boldsymbol{\theta}\right]. \tag{B.1}$$

The FIM measures the curvature of the likelihood function around the estimated parameters, reflecting the precision of the parameter estimates. A sharper likelihood function implies higher confidence in the parameter estimates. Formally, the Cramér-Rao bound establishes a theoretical lower bound for the covariance of the parameter estimates (see e.g., Söderström & Stoica (1989)):

$$\operatorname{cov}(\boldsymbol{\theta}) \geq \mathbf{F}^{-1}(\boldsymbol{\theta}). \tag{B.2}$$

[2] Commonly referred to as the *score* but this term is avoided here to prevent confusion with the score parameters x_i.

This lower bound is often used to derive estimates of parameter uncertainty. For example, assuming the approximation $\mathrm{var}(\theta_i) \approx [\mathbf{F}^{-1}]_{ii}$, the confidence interval for θ_i can be estimated as:

$$\Delta\theta_i = t_{\alpha,n-m}\sqrt{\mathrm{var}(\theta_i)}, \tag{B.3}$$

where $t_{\alpha,n-m}$ is the critical value from the Student's t-distribution for a confidence level $\alpha \in [0,1]$ with $n-m$ degrees of freedom, n being the number of data points and m the number of parameters. This approach yields a conservative estimate of the variance of θ_i, reflecting the Cramér-Rao bound. Alternative methods, such as bootstrapping, may provide more practical confidence intervals in certain cases.

When the FIM is ill-conditioned or singular, however, the parameter estimation problem becomes ill-posed. In such cases, the uncertainty bounds $\Delta\theta_i$ become unbounded or undefined. This occurs when the likelihood function exhibits invariance under certain transformations of the parameters, leading to parameter redundancy. We formally define parameter identifiability and its connection to the FIM below.

Definition B.1 ((Rothenberg, 1971, Definitions 1, 2, and 3))**.** Two parameter vectors $\boldsymbol{\theta}$ and $\boldsymbol{\theta}'$ are said to be *observationally equivalent* if $\ell(\boldsymbol{\theta}) = \ell(\boldsymbol{\theta}')$. A parameter vector $\boldsymbol{\theta}$ is *locally identifiable* if there exists an open neighborhood around $\boldsymbol{\theta}$ containing no other $\boldsymbol{\theta}'$ that is observationally equivalent to $\boldsymbol{\theta}$. If $\boldsymbol{\theta}$ is not observationally equivalent to any other parameter vector in the entire domain of the likelihood function, it is said to be *globally identifiable*.

Theorem B.1 ((Rothenberg, 1971, Theorem 1))**.** *Let $\boldsymbol{\theta}^*$ be a regular point of the FIM $\mathbf{F}(\boldsymbol{\theta})$. Then, $\boldsymbol{\theta}^*$ is locally identifiable if and only if $\mathbf{F}(\boldsymbol{\theta})$ is non-singular.*

The above theorem establishes that the FIM plays a central role in determining parameter identifiability (see also (Seber & Wild, 2005, Sections 3.4 and 8.4)).

B.2 Unidentifiability of Score Parameters

We now focus on the identifiability of the score parameters, $\boldsymbol{x}$, in paired comparison models. For simplicity, we limit the analysis to $\boldsymbol{x}$, though the results extend naturally to other parameters. We prove that the FIM for $\boldsymbol{x}$ is singular for likelihood functions satisfying the shift invariance property. While the invariance property trivially implies unidentifiability by definition, analyzing the FIM reveals deeper insights into the parameter space. Specifically, it identifies the null space causing unidentifiability and highlights subspaces suitable for well-defined reparametrizations, as explored in the next subsection.

Proposition B.1. *Let the log-likelihood function $\ell \in C^2(\mathbb{R}^m, \mathbb{R})$ satisfy the shift invariance property*

$$\ell(\boldsymbol{x} + c\mathbf{1}) = \ell(\boldsymbol{x}), \quad c \in \mathbb{R}. \tag{B.4}$$

Then, the corresponding Fisher Information Matrix $\mathbf{F}(\boldsymbol{x})$ is singular where $\mathrm{rank}(\mathbf{F}(\boldsymbol{x})) \leq m-1$, with $\mathbf{1}$ (the vector of ones) in its null space.

Proof. From (B.4) we have

$$\frac{\partial \ell(\boldsymbol{x} + c\mathbf{1})}{\partial c} = \sum_{j=1}^{m} \frac{\partial \ell(\boldsymbol{x})}{\partial x_j} = 0. \tag{B.5}$$

On the other hand, summing over all columns of the Hessian, $\mathbf{H} := \nabla_{\boldsymbol{x}} \nabla_{\boldsymbol{x}}^{\intercal} \ell(\boldsymbol{x})$, and using (B.5) yields

$$\sum_{j=1}^{m} H_{ij} = \sum_{j=1}^{m} \frac{\partial^2 \ell(\boldsymbol{x})}{\partial x_i \partial x_j} = \frac{\partial}{\partial x_i}\left(\sum_{j=1}^{m} \frac{\partial \ell(\boldsymbol{x})}{\partial x_j}\right) = 0, \qquad \forall\, i = 1, \ldots, m. \tag{B.6}$$

Hence, $\mathbf{H}$ has a zero row sum, implying $\mathbf{H}\mathbf{1} = \mathbf{0}$. Therefore, $\mathbf{1}$ lies in the null space of $\mathbf{H}$, and by extension, $\mathbf{F}(\boldsymbol{x})$ is singular with a rank of at most $m-1$. □

The singularity of $\mathbf{F}(\boldsymbol{x})$ confirms the unidentifiability of the score parameters $\boldsymbol{x}$, rendering the quantification of their uncertainties fundamentally ill-posed. As a result, the lower bounds from the Cramér-Rao inequality in (B.2) become unbounded, making the confidence interval such as in (B.3) undefined. In the next section, we analyze the structure of $\mathbf{F}(\boldsymbol{x})$ to identify subspaces where meaningful parameter estimation is possible.

B.3 IDENTIFIABLE PARAMETRIZATION

We now address which quantities are identifiable through the FIM. Specifically, any reparameterization within the range of the FIM is identifiable. Let $\mathcal{N}_{\boldsymbol{\theta}}$ denote the null space of the FIM and $\mathcal{N}_{\boldsymbol{\theta}}^{\perp}$ its orthogonal complement. Suppose $\boldsymbol{\theta} = \boldsymbol{\theta}^*$ is a local minima of the likelihood function. The FIM, when restricted to $\mathcal{N}_{\boldsymbol{\theta}^*}^{\perp}$, is positive definite, and any parameterization within this subspace is identifiable.

In the case of pairwise comparison with the optimal solution $\boldsymbol{x} = \boldsymbol{x}^*$, assuming $\operatorname{rank}(\mathbf{F}(\boldsymbol{x}^*)) = m - 1$, we have $\mathcal{N}_{\boldsymbol{x}^*} \coloneqq \operatorname{span}(\mathbf{1})$. The projection operator onto $\mathcal{N}_{\boldsymbol{x}^*}^{\perp}$ is given by

$$\mathbf{P} = \mathbf{I} - \frac{1}{m}\mathbf{1}\mathbf{1}^{\intercal}, \tag{B.7}$$

which is the centering matrix that converts $\boldsymbol{x}^*$ to the mean-zero vector $\tilde{\boldsymbol{x}}^* \coloneqq \mathbf{P}\boldsymbol{x}^* \in \mathcal{N}_{\boldsymbol{x}^*}^{\perp}$. A representation of this reparameterization can be expressed using the $(m-1)\times m$ *forward differencing matrix* $\mathbf{A} : \mathbb{R}^m \to \mathcal{N}_{\boldsymbol{x}}^{\perp}$, defined as $A_{i,i} = 1$, $A_{i,i+1} = -1$, and zero otherwise. Specifically, $\boldsymbol{y}^* \coloneqq \mathbf{A}\boldsymbol{x}^*$, where $y_i^* = x_i^* - x_{i+1}^*$. This reparameterization lies entirely in $\mathcal{N}_{\boldsymbol{x}^*}^{\perp}$, making $\boldsymbol{y}^*$ identifiable and allowing its uncertainty to be meaningfully quantified.

Thus, in paired comparison models we consider, only differences in scores provide meaningful inference. In the next section, we explore this in the context of Thurstonian covariance parameters.

APPENDIX C COVARIANCE MODEL

In Section 2.4 we expand the inclusion of covariance via Thurstonian model. We recall that, in Thurstonian model, the score parameters are assumed to be stochastic with $x_i = \mu_i + \epsilon_i$ where ϵ_i is the stochastic component with the covariance $\boldsymbol{\Sigma} = [\sigma_{ij}]$ where $\sigma_{ij} \coloneqq \operatorname{cov}(\epsilon_i, \epsilon_j)$. Furthermore, we defined the matrix $\mathbf{S} = [s_{ij}]$ where $s_{ij} = \sigma_{ii} + \sigma_{jj} - 2\sigma_{ij}$ representing the covariance of $x_i - x_j$.

In this section, we provide a detailed analysis of the covariance matrix $\boldsymbol{\Sigma}$ and its associated matrix $\mathbf{S}$. In particular, in Appendix C.1, we explore the identifiability, particularly how $\boldsymbol{\Sigma}$ is inherently non-unique while $\mathbf{S}$ remains unique and identifiable. In Appendix C.2 we present how to interpret and visualize these matrices. In Appendix C.3, we examine hierarchical clustering of competitors based on the dissimilarity matrix, uncovering performance tiers and relationships. Finally, in Appendix C.4 we discuss constraints that allow stable identification of covariance during optimization of likelihood.

C.1 NON-UNIQUENESS AND EQUIVALENCE CLASS OF COVARIANCE

We begin by noting that the likelihood function in the Thurstonian models we presented depends on the function of z_{ij} defined in (8), which itself depends on s_{ij}. That is, $\mathbf{S}$ is an observable quantity, while $\boldsymbol{\Sigma}$ is a latent variable. Below, we formalize the relationship between these two matrices and the equivalence class of covariance matrices that share the same $\mathbf{S}$.

Let $\mathbb{S}^m$ denote the space of symmetric $m \times m$ matrices and $\mathbb{S}_\circ^m$ be the the space of hollow symmetric matrices where all diagonal elements are zero. Define the map $\mathcal{S} : \mathbb{S}^m \to \mathbb{S}_\circ^m$ that associates a covariance matrix $\boldsymbol{\Sigma}$ with the matrix $\mathbf{S}$, given by

$$\mathbf{S} = \mathcal{S}(\boldsymbol{\Sigma}) = \operatorname{diag}(\boldsymbol{\Sigma})\mathbf{1}^{\intercal} + \mathbf{1}\operatorname{diag}(\boldsymbol{\Sigma})^{\intercal} - 2\boldsymbol{\Sigma}, \tag{C.1}$$

where $\operatorname{diag}(\boldsymbol{\Sigma})$ is a vector containing the diagonal elements of $\boldsymbol{\Sigma}$. This relation corresponds to $s_{ij} = \sigma_{ii} + \sigma_{jj} - 2\sigma_{ij}$ in matrix form.

As we will show momentarily, the map $\mathcal{S}$ is non-injective, as for each $\mathbf{S}$, there exist non-unique covariance matrices $\boldsymbol{\Sigma}$ differing by elements in the kernel of $\mathcal{S}$, all of which map to the same $\mathbf{S}$. Consequently, the preimage of $\mathcal{S}$ defines the equivalence class of covariance matrices producing the same $\mathbf{S}$, given by

$$[\boldsymbol{\Sigma}] = \mathcal{S}^{-1}(\mathbf{S}) = \left\{\boldsymbol{\Sigma}' \in \mathbb{S}^m \;\middle|\; \mathcal{S}(\boldsymbol{\Sigma}') = \mathbf{S}\right\}. \tag{C.2}$$

This equivalence class partitions $\mathbb{S}^m$ modulo the kernel of $\mathcal{S}$, denoted as $\mathbb{S}^m / \ker(\mathcal{S})$. We now formalize this structure.

Proposition C.1 (Equivalence Class of Covariance). *The map $\mathcal{S} : \mathbb{S}^m \to \mathbb{S}^m_\circ$, defined in (C.1), is a surjective, non-injective linear transformation. Its kernel is given by*

$$\ker(\mathcal{S}) = \{\boldsymbol{v}\mathbf{1}^\intercal + \mathbf{1}\boldsymbol{v}^\intercal \mid \boldsymbol{v} \in \mathbb{R}^m\}. \tag{C.3}$$

Consequently, the quotient space $\mathbb{S}^m / \ker(\mathcal{S})$ represents the space of equivalence classes of covariance matrices of the form

$$[\boldsymbol{\Sigma}] = \{\boldsymbol{\Sigma} + \boldsymbol{v}\mathbf{1}^\intercal + \mathbf{1}\boldsymbol{v}^\intercal \mid \boldsymbol{v} \in \mathbb{R}^m\}, \tag{C.4}$$

where all elements of $[\boldsymbol{\Sigma}]$ map to the same matrix $\mathbf{S}$ under $\mathcal{S}$.

Proof. To determine $\ker(\mathcal{S})$, consider $\boldsymbol{\Sigma}' \in \mathbb{S}^m$ such that $\mathcal{S}(\boldsymbol{\Sigma}') = \mathbf{0}$. From (C.1), it follows that

$$\operatorname{diag}(\boldsymbol{\Sigma}')\mathbf{1}^\intercal + \mathbf{1}\operatorname{diag}(\boldsymbol{\Sigma}')^\intercal - 2\boldsymbol{\Sigma}' = \mathbf{0}.$$

Rearranging, we find

$$\boldsymbol{\Sigma}' = \frac{1}{2}\left(\operatorname{diag}(\boldsymbol{\Sigma}')\mathbf{1}^\intercal + \mathbf{1}\operatorname{diag}(\boldsymbol{\Sigma}')^\intercal\right),$$

implying that any $\boldsymbol{\Sigma}' \in \ker(\mathcal{S})$ must be of the form given in (C.3). The equivalence class $[\boldsymbol{\Sigma}] = \mathcal{S}^{-1}(\mathbf{S})$ is obtained by adding elements of $\ker(\mathcal{S})$ to a representative $\boldsymbol{\Sigma}$, yielding (C.4).

To show $\mathcal{S}$ is surjective, observe that for any $\mathbf{S} \in \mathbb{S}^m_\circ$, the matrix $-\frac{1}{2}\mathbf{S} \in \mathbb{S}^m$ satisfies $\mathcal{S}(-\frac{1}{2}\mathbf{S}) = \mathbf{S}$. Hence, every $\mathbf{S} \in \mathbb{S}^m_\circ$ has at least one preimage, proving surjectivity. □

As demonstrated in Proposition C.1, the covariance matrix $\boldsymbol{\Sigma}$ is not unique, as adding any symmetric rank-one matrix to it leaves the likelihood invariant. Consequently, $\boldsymbol{\Sigma}$ is not identifiable by Definition B.1.

C.2 Interpretation and Visualization of Covariance

While the covariance matrix $\boldsymbol{\Sigma}$ is not unique, the matrix $\mathbf{S}$ is unique and identifiable. This makes $\mathbf{S}$ the preferred object for interpreting relationships between competitors. Unlike $\boldsymbol{\Sigma}$, which represents similarity, $\mathbf{S}$ plays the role of a *dissimilarity matrix*. In fact, under suitable conditions, $\mathbf{S}$ can be interpreted as a distance matrix.

For $\mathbf{S}$ to qualify as a distance matrix, it must be non-negative. This holds if and only if the doubly-centered covariance matrix, defined as

$$\tilde{\boldsymbol{\Sigma}} = \mathbf{P}\boldsymbol{\Sigma}\mathbf{P}, \tag{C.5}$$

is positive semi-definite. Here, $\mathbf{P}$ is the centering matrix from (B.7). The matrix $\tilde{\boldsymbol{\Sigma}}$ represents the covariance of mean-centered scores, $\tilde{\boldsymbol{x}} = \mathbf{P}\boldsymbol{x}$. Importantly, $\tilde{\boldsymbol{\Sigma}} \in [\boldsymbol{\Sigma}]$ and is unique within the equivalence class $[\boldsymbol{\Sigma}]$. Specifically, for any $\boldsymbol{\Sigma}', \boldsymbol{\Sigma}'' \in [\boldsymbol{\Sigma}]$, it holds that $\mathbf{P}\boldsymbol{\Sigma}'\mathbf{P} = \mathbf{P}\boldsymbol{\Sigma}''\mathbf{P} = \tilde{\boldsymbol{\Sigma}}$. This ensures that $\tilde{\boldsymbol{\Sigma}}$ is well-defined and serves as a canonical representation of the covariance structure.

A matrix $\mathbf{S}$ is called a Euclidean distance matrix if there exist spatial points $\boldsymbol{p}_1, \ldots, \boldsymbol{p}_m$ such that $s_{ij} = \|\boldsymbol{p}_i - \boldsymbol{p}_j\|^2$. (Mardia et al., 1979, Theorem 14.2.1) guarantees that $\mathbf{S}$ is a Euclidean distance matrix if and only if $\tilde{\boldsymbol{\Sigma}}$ is positive semi-definite. In our setting, this condition is always satisfied, as we enforce the factor covariance model in (12) as

$$\boldsymbol{\Sigma} = \mathbf{D} + \boldsymbol{\Lambda}\boldsymbol{\Lambda}^\intercal, \tag{C.6}$$

where $\mathbf{D}$ is a positive diagonal matrix ($\mathbf{D} > \mathbf{0}$), ensuring that $\boldsymbol{\Sigma}$ is positive definite. Consequently, $\tilde{\boldsymbol{\Sigma}}$ is positive semi-definite, making $\mathbf{S}$ a Euclidean distance matrix.

This property enables meaningful visualization of $\mathbf{S}$ using multi-dimensional scaling (MDS). MDS (see e.g., (Mardia et al., 1979, Chapter 14) or (Seber & Wild, 2005, Section 5.5)) constructs a set of points in two- or three-dimensional space such that their pairwise distances approximate the distances in $\mathbf{S}$. This approach is particularly suitable for visualizing $\mathbf{S}$ or similar distance matrices derived from covariance models.

Here, we use the matrix $\mathbf{Z}$ as the distance matrix, as defined in (8), where $z_{ij} = (x_i - x_j)/\sqrt{s_{ij}}$, capturing both score differences and dissimilarities derived from covariance. The dissimilarity represented by $\mathbf{Z}$ is visualized in Figure C.1 using MDS, showing the first two principal coordinates in a

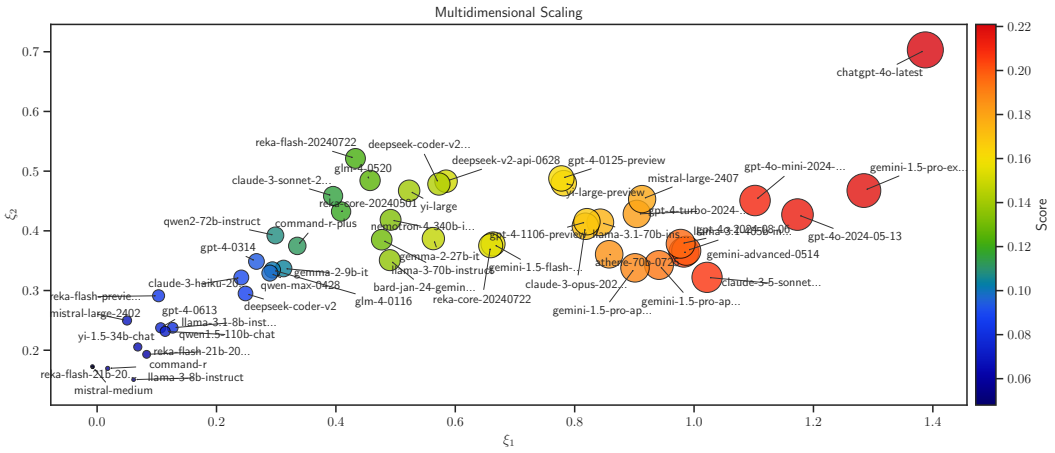

Figure C.1: MDS projection of the distance matrix $\mathbf{Z}$ onto two dimensions, focusing on the top 50 competitors ranked by Model 18 of Table D.1. In the scatter plot, circle size and color are proportional to the competitors' scores.

two-dimensional space. Interestingly, the spatial arrangement of points in the plot not only reflects the pairwise dissimilarities but also aligns well with the overall ranking of competitors, effectively capturing their relative scores. This highlights the ability of MDS to extract meaningful patterns from the dissimilarity matrix alone, without access to the absolute values of the scores.

A companion approach to MDS is kernel PCA (Schölkopf et al., 1999; Williams, 2000), applied earlier in Section 4.2 and visualized in Figure 2. Both techniques aim to uncover relationships in lower-dimensional spaces, with kernel PCA operating on points in a feature space and MDS working directly with pairwise distances. These methods are dual representations: PCA identifies principal components of the data, while MDS identifies principal coordinates of a distance matrix (Mardia et al., 1979, Section 14.3). Notably, both the kernel PCA in Figure 2 (projected into three dimensions) and the MDS in Figure C.1 (projected into two dimensions) rely on the same dissimilarity matrix $\mathbf{Z}$, offering consistent visual representations of the relationships between competitors.

C.3 Hierarchical Clustering of Competitor Performance

To complement the analysis of dissimilarity using PCA and MDS, we applied hierarchical agglomerative clustering (Hastie et al., 2009, Section 14.3.12) with optimal leaf ordering (Bar-Joseph et al., 2001) to the dissimilarity matrix $\mathbf{Z}$. We recall that the matrix $\mathbf{Z}$ incorporates both score differences and dissimilarities derived from Thurstonian covariance in our generalized models. This integration of covariance within the model enhances the interpretability of the clustering results by capturing both the performance levels of competitors and their correlation structure.

The clustering analysis focuses on the top 100 competitors ranked using Model 18 from Table D.1. The resulting dendrogram, shown in Figure C.2, reveals a hierarchical structure that organizes competitors into distinct tiers. Interestingly, despite the clustering algorithm having access only to the dissimilarity matrix $\mathbf{Z}$, which encodes score differences (but not the absolute values of the scores) and covariance-derived uncertainties, the clustering order closely aligns with the competitors' rankings. Each cluster roughly corresponds to a contiguous range of ranks, albeit with minor variations, demonstrating the ability of the clustering method to infer meaningful performance groupings solely from relative differences.

This hierarchical structure provides additional insights into the relationships between competitors, suggesting a natural stratification into performance-based tiers. Tier I includes the two leading models, ChatGPT-4 Latest (as of September 2024) and Gemini 1.5 Pro Experimental, reflecting their dominance. The remaining 98 models form Tier II, which is further divided into two subgroups: II_A (green theme) and II_B (red theme). These subgroups are further refined into smaller clusters: II_A

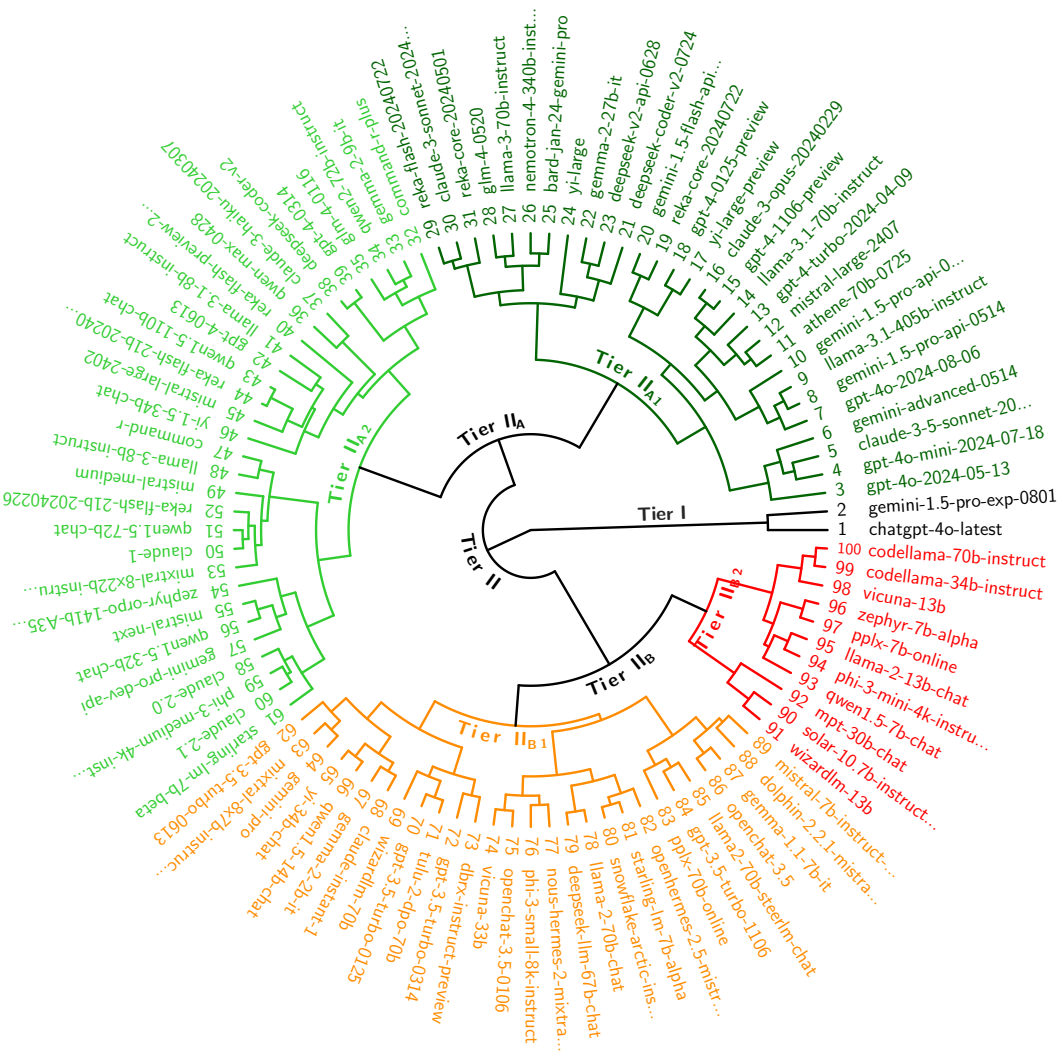

Figure C.2: Hierarchical clustering of the top 100 competitors based on the distance matrix $\mathbf{Z}$ derived from model 18 of Table D.1. The clustering reveals performance tiers, with Tier I consisting of the top two competitors (ChatGPT-4 Latest and Gemini 1.5 Experimental) and Tier II further subdivided into groups representing decreasing performance levels. This structure highlights the relationships and relative strengths among competitors.

is split into II_{A_1} (dark green) and II_{A_2} (light green), while II_{B} is split into II_{B_1} (light red) and II_{B_2} (dark red).

The meaningfulness of these groupings arises from the use of $\mathbf{Z}$, which integrates performance scores with covariance-based dissimilarities—a capability enabled by incorporating Thurstonian models in our framework. This analysis complements rankings by uncovering hierarchical structures and relational patterns among competitors.

While our analysis focuses on hierarchical clustering derived from the statistical properties of the dissimilarity matrix, alternative clustering approaches, such as those leveraging semantic relationships (e.g., embeddings or linguistic features), could provide complementary insights into model relationships. However, incorporating semantic clustering would necessitate additional datasets or features, which fall outside the scope of this work. Future research may explore such directions, integrating semantic and statistical perspectives to uncover deeper insights into competitor relationships.

C.4 Constraint on Thurstonian Covariance Model

In Section 2.5 we presented constraints to resolve the symmetry of the likelihood functions with respect to transformations of the parameters. Here, we provide further detail on the second symmetry presented therein. Specifically, we recall that the models are invariant under the transformation $(x_i, \sigma_{ij}) \mapsto (tx_i, t^2\sigma_{ij})$ for an arbitrary $t > 0$.

One approach to resolve the arbitrariness of the parameters introduced by such translation is to impose the constraint

$$\mathcal{C} \coloneqq \frac{1}{2m} \sum_{i,j=1}^{m} s_{ij} = 1, \tag{C.7}$$

where the constant $\frac{1}{2m}$ is arbitrary, chosen for convenience as we will explain momentarily. Since s_{ij} represents the variance of the difference $x_{ij} = x_i - x_j$, this constraint ensures that the total variance of all random processes x_{ij} is fixed. We will further show that this constraint can be directly expressed in terms of the covariance matrix $\mathbf{\Sigma}$.

Proposition C.2. *The constraint in (C.7) is equivalent to*

$$\operatorname{trace}(\tilde{\mathbf{\Sigma}}) = 1, \tag{C.8}$$

where $\tilde{\mathbf{\Sigma}} \coloneqq \mathbf{P}\mathbf{\Sigma}\mathbf{P}$ is the doubly-centered covariance matrix given in (C.5), and $\mathbf{P} \coloneqq \mathbf{I} - \frac{1}{m}\mathbf{1}\mathbf{1}^\intercal$ is the projection matrix defined in (B.7).

Proof. Recall that the elements s_{ij} of the matrix $\mathbf{S}$ are related to $\mathbf{\Sigma}$ by

$$\mathbf{S} = \operatorname{diag}(\mathbf{\Sigma})\mathbf{1}^\intercal + \mathbf{1}\operatorname{diag}(\mathbf{\Sigma})^\intercal - 2\mathbf{\Sigma}. \tag{C.9}$$

The constraint in (C.7) can be written equivalently as

$$\mathcal{C} = \frac{1}{2m}\mathbf{1}^\intercal\mathbf{S}\mathbf{1}. \tag{C.10}$$

Substituting (C.9) into (C.10) and noting that $\mathbf{1}^\intercal \operatorname{diag}(\mathbf{\Sigma}) = \operatorname{trace}(\mathbf{\Sigma})$, we obtain

$$\mathcal{C} = \operatorname{trace}(\mathbf{\Sigma}) - \frac{1}{m}\mathbf{1}^\intercal\mathbf{\Sigma}\mathbf{1}. \tag{C.11}$$

Next, let $\mathbf{J}_\circ \coloneqq \frac{1}{m}\mathbf{1}\mathbf{1}^\intercal$, so that $\mathbf{P} = \mathbf{I} - \mathbf{J}_\circ$. Expanding $\tilde{\mathbf{\Sigma}} \coloneqq \mathbf{P}\mathbf{\Sigma}\mathbf{P}$, we use the cyclic property of the trace and the idempotence of $\mathbf{J}_\circ$ ($\mathbf{J}_\circ^2 = \mathbf{J}_\circ$) to write

$$\operatorname{trace}(\tilde{\mathbf{\Sigma}}) = \operatorname{trace}(\mathbf{\Sigma}) - \operatorname{trace}(\mathbf{\Sigma}\mathbf{J}_\circ). \tag{C.12}$$

Moreover, by the cyclic property of the trace,

$$\operatorname{trace}(\mathbf{\Sigma}\mathbf{J}_\circ) = \frac{1}{m}\mathbf{1}^\intercal\mathbf{\Sigma}\mathbf{1}. \tag{C.13}$$

Substituting this into (C.12), we find that $\operatorname{trace}(\tilde{\mathbf{\Sigma}})$ equals $\mathcal{C}$ as expressed in (C.11). Thus, the constraint (C.7) is equivalent to $\operatorname{trace}(\tilde{\mathbf{\Sigma}}) = 1$, completing the proof. □

By using the factor model (12), the constraint can be written in terms of $\mathbf{D}$ and $\mathbf{\Lambda}$ as

$$\mathcal{C} = \left(1 - \frac{1}{m}\right)\operatorname{trace}(\mathbf{D}) + \|\mathbf{\Lambda}\|_F^2 - \|\mathbf{1}^\intercal\mathbf{\Lambda}\|_2^2, \tag{C.14}$$

where $\|\cdot\|_F$ is the Frobenius norm. Also, the derivative of the constraint with respect to these matrices (which is needed for the optimization methods utilizing Jacobin of the loss function) is can be derived as

$$\frac{\partial\mathcal{C}}{\partial\mathbf{D}} = \left(1 - \frac{1}{m}\right)\mathbf{I}, \tag{C.15a}$$

$$\frac{\partial\mathcal{C}}{\partial\mathbf{\Lambda}} = (2\mathbf{I} - \mathbf{1}\mathbf{1}^\intercal)\mathbf{\Lambda}. \tag{C.15b}$$

Id	Model	Model Features		Num. Param.	$-\ell(\boldsymbol{\theta})$	Cross Entropy			Training Time
		Cov. (k)	Tie (k)			Win	Loss	Tie	
1	Bradley-Terry	✗	✗	129	0.6554	0.3177	0.3376	—	2.3
2	(*with tie data*)	0	✗	258	0.6552	0.3180	0.3371	—	3.8
3		3	✗	645	0.6549	0.3178	0.3370	—	34.1
4	Bradley-Terry	✗	✗	129	0.6351	0.3056	0.3295	—	0.0
5	(*without tie data*)	0	✗	258	0.6346	0.3059	0.3287	—	1.7
6		3	✗	645	0.6342	0.3057	0.3285	—	27.5
7	Rao-Kupper	✗	0	130	1.0095	0.3405	0.3462	0.3227	5.8
8		✗	1	258	1.0106	0.3401	0.3459	0.3245	6.9
9		✗	10	1419	1.0055	0.3404	0.3455	0.3196	208.1
10		✗	20	2709	1.0050	0.3403	0.3455	0.3192	396.9
11		0	0	259	1.0092	0.3408	0.3457	0.3228	8.4
12		0	1	387	1.0103	0.3404	0.3454	0.3245	7.5
13		0	10	1548	1.0052	0.3407	0.3449	0.3196	293.7
14		0	20	2838	1.0048	0.3406	0.3449	0.3193	664.9
15		3	0	646	1.0089	0.3405	0.3457	0.3227	36.0
16		3	1	774	1.0100	0.3400	0.3454	0.3245	36.9
17		3	10	1935	1.0049	0.3403	0.3449	0.3196	363.5
18		3	20	3225	1.0044	0.3403	0.3449	0.3193	817.3
19	Davidson	✗	0	130	1.0100	0.3409	0.3461	0.3231	6.0
20		✗	1	258	1.0077	0.3413	0.3466	0.3198	10.5
21		✗	10	1419	1.0057	0.3404	0.3456	0.3197	253.2
22		✗	20	2709	1.0052	0.3404	0.3455	0.3193	602.8
23		0	0	259	1.0098	0.3411	0.3455	0.3231	8.7
24		0	1	387	1.0074	0.3415	0.3460	0.3200	8.3
25		2	10	1548	1.0055	0.3407	0.3451	0.3197	286.9
26		0	20	2838	1.0050	0.3407	0.3450	0.3194	665.1
27		3	0	646	1.0094	0.3410	0.3453	0.3231	34.6
28		3	1	774	1.0070	0.3412	0.3460	0.3199	35.8
29		3	10	1935	1.0051	0.3407	0.3448	0.3197	366.4
30		3	20	3225	1.0047	0.3405	0.3448	0.3194	804.9

Table D.1: Configurations and training details of the 30 statistical models used throughout the analysis. These models are referenced by their ID in various sections of the paper.

APPENDIX D EVALUATION OF STATISTICAL MODELS

D.1 MODEL SELECTION

The models used in our analysis are listed in Table D.1. Rows 1 to 6 include the Bradley-Terry (BT) model and its variants, rows 7 to 18 cover the Rao-Kupper model and its extensions, and rows 19 to 30 represent the Davidson model and its variants. For the BT model, we analyze two forms of the dataset. In rows 1 to 3, we modify the input matrices to incorporate ties by treating each tie as half a win and half a loss, i.e., $\mathbf{W} \leftarrow \mathbf{W} + \frac{1}{2}\mathbf{T}$, as done by Chiang et al. (2024). In rows 4 to 6, we did not modify $\mathbf{W}$. We recall that in both cases, the BT model does not account for ties, meaning $\mathbf{T}$ is effectively ignored.

Rows 1, 4, 7, and 19 represent the standard versions of the BT, Rao-Kupper, and Davidson models as found in the literature. All other rows reflect our extensions, detailed in the third and fourth columns of the table. In the third column, k refers to the rank of $\boldsymbol{\Lambda}$ in the factor model for covariance, as given in (12). The symbol "✗" means the model does not include a covariance structure, excluding the Thurstonian representation. $k = 0$ implies a diagonal covariance matrix, i.e., $\boldsymbol{\Sigma} = \mathbf{D}$.

In the fourth column, k represents the number of columns of the matrix $\mathbf{G}$ in the factor model for ties, as defined in (6). The symbol "✗" indicates that the model does not account for ties, while $k = 0$ corresponds to the original Rao-Kupper and Davidson models with a single tie parameter.

We trained these models (except for models 1 and 4) by maximizing the likelihood function (1) using the BFGS optimization method, while satisfying the constraints in Section 2.5. This method requires both the loss function $-\ell(\boldsymbol{\theta})$ and its Jacobian $-\partial\ell(\boldsymbol{\theta})/\partial\boldsymbol{\theta}$, which we analytically derived with respect to all parameters for each model and provided during training. The negative log-likelihood (NLL) is shown in the fifth column, and training time (in seconds), using an AMD EPYC 7543 processor with 32 cores, is given in the last column.

Models in rows 1 and 4 were trained using the iterative minorization–maximization (MM) algorithm of Newman (2023), which offers notable speed advantages over conventional maximum likelihood estimation. MM algorithms have also been extended to certain generalizations of the Bradley-Terry model, as shown by Hunter (2004). Whether MM methods are directly applicable to the more complex generalized models proposed in this work remains an open question and warrants further investigation.

Since the BT models do not account for ties, their NLL values are generally lower compared to other models, making direct comparison between the NLLs of BT and tie-inclusive models not applicable. However, within each model category, we observe that incorporating the Thurstonian covariance structure and the tie factor model (with increasing k) improves the NLL, indicating a better fit. Further evaluation metrics, including goodness-of-fit and generalization performance, are provided in Appendix D.2 and Appendix D.3, respectively.

D.2 Model Fit

One method of assessing model fit, as presented in Table D.2, is to compare the cross-entropy between the predicted probabilities of win, loss, and tie with the observed probabilities. The sum of these cross-entropies matches the NLL, as shown in the fifth column. As with the NLL, the BT models show lower cross-entropy, though this is due to their differing dimensionality. Specifically, the BT model predicts two outcomes (win and loss), yielding only one independent output, since the probability of loss complements the probability of a win. In contrast, the Rao-Kupper and Davidson models predict three outcomes (win, loss, tie), resulting in two independent output variables. Thus, the BT model fits a one-dimensional output space, while the other models fit a two-dimensional space. Although the BT models achieve lower error rates, the complexity of the Rao-Kupper and Davidson models offers richer predictions.

Within each model category, increasing the rank k for covariance or tie models consistently improves fit, as indicated by decreasing cross-entropies. Further improvements are also reflected in other metrics, such as RMSE and divergence values, which will be discussed in the following paragraphs.

Another metric for comparison, provided in the fifth to eighth columns of Table D.2, is training accuracy via the root-mean-square error (RMSE) between predicted and observed data. Given the non-uniform number of comparisons per pair, we use weighted RMSE, with weights proportional to the number of comparisons. Results for win, loss, and tie are presented in the fifth to seventh columns, while overall RMSE is in the eighth. Similar to earlier trends, BT models show lower errors, but models with higher k values for covariance and ties show significant improvements in model accuracy.

We also compare models using the divergence between predicted probabilities and observed data. For each pair, we compute the Kullback-Leibler (KL) divergence between the predicted and observed probability mass functions. The KL divergence $D_{\mathrm{KL}}(P\|Q)$, averaged over all pairs, is shown in the ninth column of Table D.2. Additionally, the Jensen-Shannon (JS) divergence $D_{\mathrm{JS}}(P\|Q)$, which is symmetric and ranges between 0 and 1, is provided in the tenth column. Lower KL and JS values indicate better model fit. Notably, models incorporating covariance and tie factor models yield better results in terms of divergence, reaffirming the effectiveness of these extensions.

D.3 Generalization Performance

To evaluate the models' generalization performance, we trained each model on 90% of the data and tested predictions on the remaining 10%. The weighted RMSE of the predictions is presented in the fifth to eighth columns of Table D.3, and the KL and JS divergences are shown in the ninth and tenth columns, respectively.

		Model Features		RMSE				Divergence ($\times 10^2$)	
Id	Model	Cov. (k)	Tie (k)	Win	Loss	Tie	All	KLD	JSD
1	Bradley-Terry	✗	✗	29.7	29.7	—	29.7	1.49	0.44
2	(*with tie data*)	0	✗	26.2	26.2	—	26.2	1.42	0.42
3		3	✗	17.4	17.4	—	17.4	1.30	0.39
4	Bradley-Terry	✗	✗	35.1	35.1	—	35.1	1.82	0.52
5	(*without tie data*)	0	✗	31.5	31.5	—	31.5	1.71	0.49
6		3	✗	17.3	17.3	—	17.3	1.58	0.46
7	Rao-Kupper	✗	0	48.2	69.9	103.5	77.3	3.32	0.92
8		✗	1	46.4	67.8	99.2	74.3	3.45	0.91
9		✗	10	34.1	34.2	23.1	30.9	2.63	0.73
10		✗	20	34.3	32.2	16.8	28.8	2.35	0.65
11		0	0	46.5	67.9	103.6	76.4	3.23	0.90
12		0	1	43.5	66.8	99.4	73.5	3.36	0.89
13		0	10	29.8	31.6	22.7	28.3	2.55	0.70
14		0	20	30.4	29.1	16.7	26.1	2.26	0.63
15		3	0	49.0	61.7	104.7	75.6	3.09	0.86
16		3	1	48.6	58.7	100.9	73.0	3.18	0.84
17		3	10	20.0	21.2	22.1	21.1	2.42	0.67
18		3	20	18.7	18.9	15.8	17.9	2.12	0.59
19	Davidson	✗	0	51.0	71.8	109.8	81.3	3.41	0.94
20		✗	1	44.4	63.3	90.1	68.6	2.99	0.82
21		✗	10	37.1	39.6	25.7	34.7	2.69	0.75
22		✗	20	37.7	37.1	17.2	32.1	2.50	0.70
23		0	0	49.4	70.5	109.9	80.6	3.32	0.92
24		0	1	41.1	62.4	91.4	68.1	2.94	0.81
25		0	10	32.8	37.7	27.0	32.8	2.73	0.76
26		0	20	35.7	32.6	18.8	30.0	2.56	0.72
27		3	0	55.1	61.1	111.0	79.8	3.18	0.89
28		3	1	46.5	50.0	90.6	65.5	2.80	0.78
29		3	10	20.8	22.0	25.0	22.7	2.57	0.72
30		3	20	19.1	19.0	17.1	18.4	2.43	0.68

Table D.2: Goodness-of-fit metrics, including root-mean-square error (RMSE), Kullback-Leibler divergence (KLD), and Jensen-Shannon divergence (JSD), for the 30 statistical models introduced in Table D.1.

An important observation is that increasing the number of parameters, such as k in the covariance or tie factor models, improves the fit to training data (Table D.2), but can reduce generalization performance, as seen in Table D.3. While adding parameters helps mitigate underfitting in simpler models like the original Rao-Kupper and Davidson, too many parameters lead to overfitting. Models with k in the range of 1 to 10 strike a balance between fit and generalization, whereas higher k values, such as $k = 20$, tend to overfit the data, reducing generalization performance.

D.4 Evaluating Marginal Probabilities of Win, Loss, and Tie

The errors in previous sections were calculated based on pairwise probabilities, such as $P_{i \succ j}$, with errors averaged over all pairwise comparisons in E. Here, we assess the *marginal* probabilities for each competitor, which represent the overall likelihood of winning, losing, or tying against all other competitors. Specifically, we denote these probabilities respectively by $P(i \succ V \setminus \{i\} \mid E)$,

Id	Model	Model Features		RMSE				Divergence ($\times 10^2$)	
		Cov. (k)	Tie (k)	Win	Loss	Tie	All	KLD	JSD
1	Bradley-Terry	✗	✗	27.4	27.4	—	27.4	1.46	0.41
2	(*with tie data*)	0	✗	27.6	27.6	—	27.6	1.48	0.41
3		3	✗	27.1	27.1	—	27.1	1.55	0.43
4	Bradley-Terry	✗	✗	30.0	30.0	—	30.0	1.74	0.48
5	(*without tie data*)	0	✗	30.3	30.3	—	30.3	1.77	0.48
6		3	✗	30.4	30.4	—	30.4	2.06	0.54
7	Rao-Kupper	✗	0	54.5	29.1	67.4	52.8	3.16	0.88
8		✗	1	49.9	41.9	74.3	57.0	3.31	0.87
9		✗	10	26.3	38.6	32.1	32.7	3.17	0.85
10		✗	20	29.0	56.3	66.4	53.0	4.14	1.00
11		0	0	52.9	31.5	67.8	52.9	3.19	0.88
12		0	1	46.8	45.5	75.0	57.4	3.31	0.87
13		0	10	25.9	38.5	31.6	32.4	3.10	0.84
14		0	20	30.2	54.3	64.6	51.7	4.66	1.06
15		3	0	50.3	33.8	68.1	52.6	3.31	0.90
16		3	1	51.6	42.1	75.0	57.9	3.59	0.93
17		3	10	28.5	34.9	32.3	32.0	3.25	0.87
18		3	20	35.9	59.6	67.1	55.8	4.71	1.07
19	Davidson	✗	0	54.7	30.8	70.0	54.3	3.28	0.91
20		✗	1	43.1	24.6	42.7	37.8	2.76	0.77
21		✗	10	27.6	41.8	31.4	34.2	2.95	0.81
22		✗	20	28.6	65.5	73.4	59.1	3.42	0.93
23		0	0	54.0	32.8	70.2	54.5	3.31	0.92
24		0	1	44.2	26.2	45.0	39.4	2.91	0.81
25		0	10	26.9	40.0	28.6	32.3	3.06	0.84
26		0	20	31.0	68.3	80.4	63.5	3.51	0.96
27		3	0	52.5	34.2	70.2	54.3	3.40	0.93
28		3	1	40.7	28.2	44.0	38.2	2.87	0.79
29		3	10	33.6	32.7	28.5	31.6	3.31	0.89
30		3	20	32.8	71.6	83.3	66.2	3.70	1.00

Table D.3: Generalization performance of the 30 statistical models introduced in Table D.1, evaluated on test data using a 90/10 train-test split, including root-mean-square error (RMSE), Kullback-Leibler divergence (KLD), and Jensen-Shannon divergence (JSD).

$P(i \prec V \setminus \{i\} \,|\, E)$, and $P(i \sim V \setminus \{i\} \,|\, E)$, which are defined by

$$P(i \succ V \setminus \{i\} \,|\, E) = \sum_{\{i,j\} \in E(i)} P(i \succ j \,|\, \{i,j\})\, P(\{i,j\} \,|\, E) \tag{D.1a}$$

$$P(i \prec V \setminus \{i\} \,|\, E) = \sum_{\{i,j\} \in E(i)} P(i \prec j \,|\, \{i,j\})\, P(\{i,j\} \,|\, E) \tag{D.1b}$$

$$P(i \sim V \setminus \{i\} \,|\, E) = \sum_{\{i,j\} \in E(i)} P(i \sim j \,|\, \{i,j\})\, P(\{i,j\} \,|\, E) \tag{D.1c}$$

where $E(i) \coloneqq \{e \in E \,|\, i \in e\}$ represents the set of edges incident to the vertex $i \in V$, and $P(\{i,j\} \,|\, E)$ is the probability of observing a match for the pair $\{i,j\}$, which can be empirically obtained as

$$P(\{i,j\} \,|\, E) = \frac{n_{ij}}{\sum_{\{k,l\} \in E} n_{kl}}. \tag{D.2}$$

For brevity, we denote the marginal probabilities in (D.1a) to (D.1c) by $P_{i \succ V_{-i}}$, $P_{i \prec V_{-i}}$, and $P_{i \sim V_{-i}}$, respectively, where $V_{-i} \coloneqq V \setminus \{i\}$.

Here, we evaluate the goodness of fit of the models by comparing the predicted marginal probabilities of winning, losing, and tying for each competitor against their corresponding empirical marginal probabilities. Figure D.1 illustrates the marginal probabilities for a selected set of models. Specifically, the first two rows of the figure show results from the BT models (Models 1 and 4 of Table D.1, with and without modified input data), while rows 3 to 5 correspond to the Rao-Kupper models (Model 7 as the standard model with one tie parameter corresponding to $k = 0$, Model 8 with factored tie model and $k = 1$, and Model 10 with factored tie model and $k = 20$). Results for the Davidson models are omitted as they closely resemble those of the Rao-Kupper models under similar conditions.

The left, middle, and right columns in the figure show the marginal probabilities of win ($P_{i \succ V_{-i}}$), loss ($P_{i \prec V_{-i}}$), and tie ($P_{i \sim V_{-i}}$), respectively. Each row shares the same legend, shown only in the rightmost column. The abscissa represent competitors ordered by their rank in Model 18. The colored curves represent predicted marginal probabilities for each model, with red-themed curves for BT models and green-themed curves for Rao-Kupper models. The black curve represents the empirical marginal probabilities from the observed data, though in some cases, it may overlap with the colored curves. The relative error between model predictions and empirical probabilities is presented in the sixth row, using the same color scheme for consistency. Key observations are as follows:

First, in the top two rows of the figure, the BT model predictions (red curves) noticeably deviate from the empirical probabilities (black curve). This is because the BT models do not account for ties, resulting in a different distribution of win and loss probabilities. To provide a fair comparison, we compare BT model predictions with adjusted empirical probabilities, represented by the dotted black curve, which excludes ties. Accordingly, the relative error for BT models in the sixth row is computed against these adjusted probabilities. By contrast, Rao-Kupper models are compared directly with empirical probabilities from the input data, which include ties.

Second, in the last row of the figure, we observe that the errors for BT models are generally lower than those of the Rao-Kupper models. This is due to the smaller output dimension of the BT models, which fit only win and loss outcomes. If BT models were compared to the actual empirical probabilities (solid black curves), their error would be higher than that of the Rao-Kupper models. However, such a comparison would be unfair, as BT models are trained on modified data that does not account for ties.

Finally, the original Rao-Kupper model with a single tie parameter (Model 7, shown in the third row) exhibits errors on the order of $\mathcal{O}(1)$ to $\mathcal{O}(10)$, making it impractical for real-world applications, particularly in predicting ties. However, our generalized Rao-Kupper models, which incorporate factored tie models (Models 8 and 10, shown in the fourth and fifth rows), demonstrate a substantial improvement in accuracy. This enhancement not only elevates the prediction of ties but also improves the prediction of win and loss outcomes by one to two orders of magnitude. This result is significant, as it brings the Rao-Kupper and Davidson models back into practical relevance, offering richer predictions for win, loss, and tie outcomes—unlike the BT models—without compromising on accuracy.

APPENDIX E COMPARATIVE ANALYSIS OF RANKING VARIABILITY

This section expands on the ranking comparisons of chatbots presented in Section 4, providing a detailed examination of how ranking outcomes vary across statistical models with different parameter configurations. We include visualizations such as the baseline model's score distribution (Appendix E.1), a comparison of ranking consistency across models (Appendix E.2), and Kendall's tau correlation matrix (Appendix E.3) to assess ranking similarities. Further exploration using hierarchical clustering (Appendix E.4) reveals underlying structure in model-based rankings, highlighting the distinct influences of covariance and tie parameters on ranking alignment.

E.1 BASELINE MODEL RANKINGS AND SCORE DISTRIBUTION

The score plot in Figure E.1 illustrates the scores for the top 50 chatbots, ranked according to Model 18 from Table D.1. This model, a generalized Rao-Kupper (RK) variant with a higher covariance factor $k = 3$ and tie factor $k = 20$, achieves improved goodness of fit, making it a suitable baseline

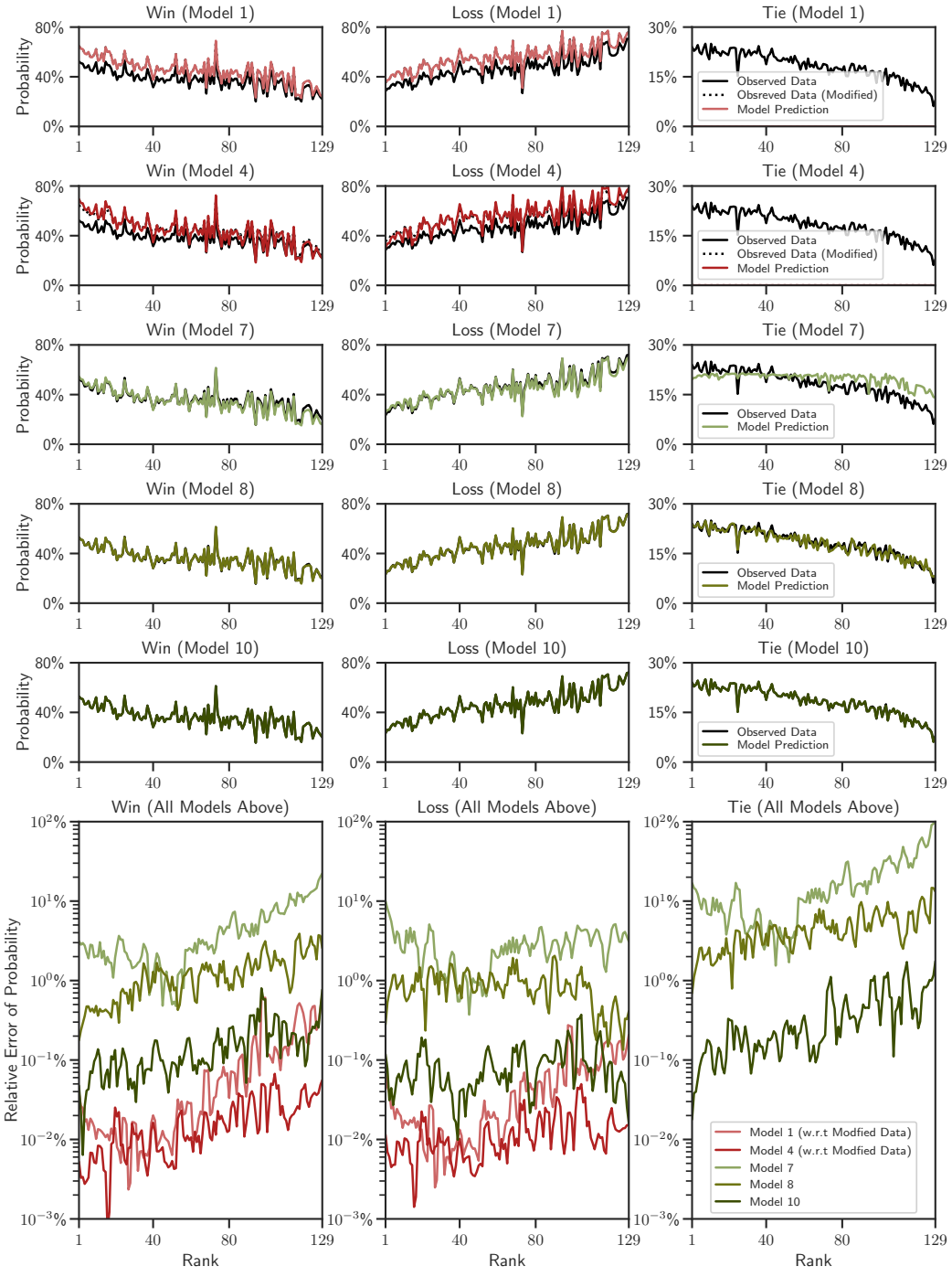

Figure D.1: Comparison of predicted (colored curves) and empirical (black curves) marginal probabilities of win (left), loss (middle), and tie (right) for selected models. First and second rows: BT models, third row: original Rao-Kupper with tie factor $k = 0$, fourth and fifth rows: generalized Rao-Kupper with tie factors $k = 1$ and $k = 20$. Sixth row: relative errors for rows one to five, calculated between predicted and empirical probabilities.

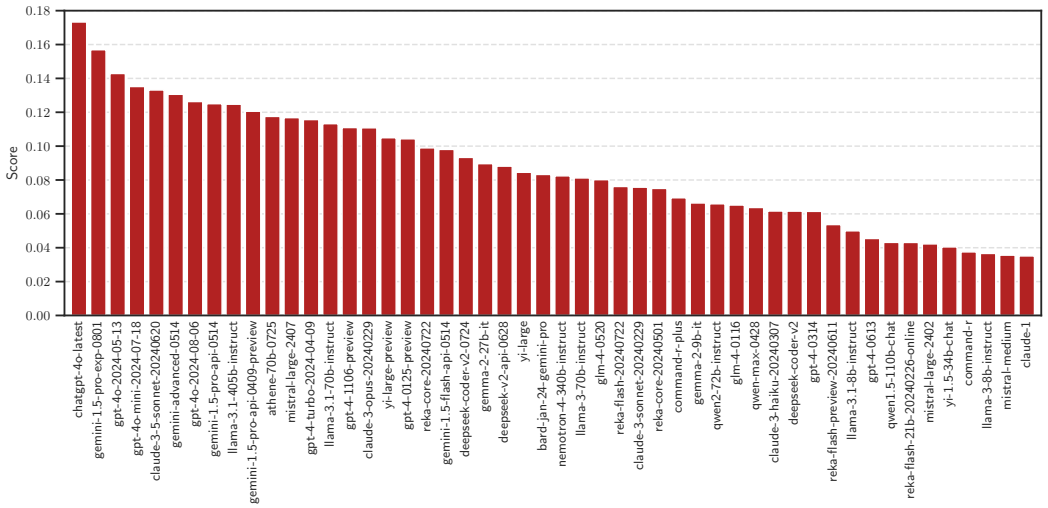

Figure E.1: Competitors ranked by their scores according to Model 18 from Table D.1.

for comparison. We selected this RK model over a similar generalized Davidson model due to their comparable performance. Each bar in the plot represents a chatbot, ordered by their relative score from highest to lowest. We use the ranking by this model as a basis for comparing consistency across top competitors, examined further in the bump chart analysis.

E.2 Exploring Ranking Consistency Across Models

The bump chart from Section 4 (now displayed here in Figure E.2) compares rankings across 12 selected models from Table D.1. Each row represents a competitor, ranked by the leftmost model, and columns represent models of increasing complexity from right to left. By following the colored lines across the chart, we observe how rankings shift as models incorporate additional parameters for ties and covariances. The rankings of top competitors remain consistent across all models, indicating robustness, while discrepancies grow more pronounced at lower ranks.

E.3 Quantifying Ranking Similarity Across Models

While the bump chart in Section 4.1 provides a visual overview of ranking shifts across models, quantifying the degree of similarity between these rankings requires statistical correlation measures. A variety of methods are available, including Pearson's correlation, Spearman's ρ, and Kendall's τ (Kendall, 1938). Pearson's correlation is most suitable for assessing linear relationships between continuous variables, while Spearman's rank correlation generalizes it for monotonic relationships in ordinal data. However, neither offers as direct an interpretation for pairwise ranking comparisons as Kendall's τ, which evaluates the ordering of pairs directly, making it particularly well-suited for ordinal data.

Kendall's ranking correlation quantifies the agreement between two ranking orders by comparing the relative ordering of pairs of objects. Given two score vectors, $\boldsymbol{x}^p = (x_1^p, \ldots, x_m^p)$ and $\boldsymbol{x}^q = (x_1^q, \ldots, x_m^q)$, from the p-th and q-th models, respectively, τ_{pq} reflects the extent to which the pairwise orderings are concordant. Two pairs, (x_i^p, x_j^p) and (x_i^q, x_j^q), are *concordant* if they maintain the same relative ordering—i.e., either $x_i^p < x_j^p$ and $x_i^q < x_j^q$, or $x_i^p > x_j^p$ and $x_i^q > x_j^q$. Conversely, they are *discordant* if the orderings are reversed. This concordance criterion can be expressed as $\mathsf{sgn}(x_i^p - x_j^p)\,\mathsf{sgn}(x_i^q - x_j^q) = 1$ for concordant pairs, and -1 for discordant pairs.

The Kendall correlation τ_{pq} is defined as the difference between the probabilities of concordant and discordant pairs and can be empirically obtained by computing the difference of counts for all

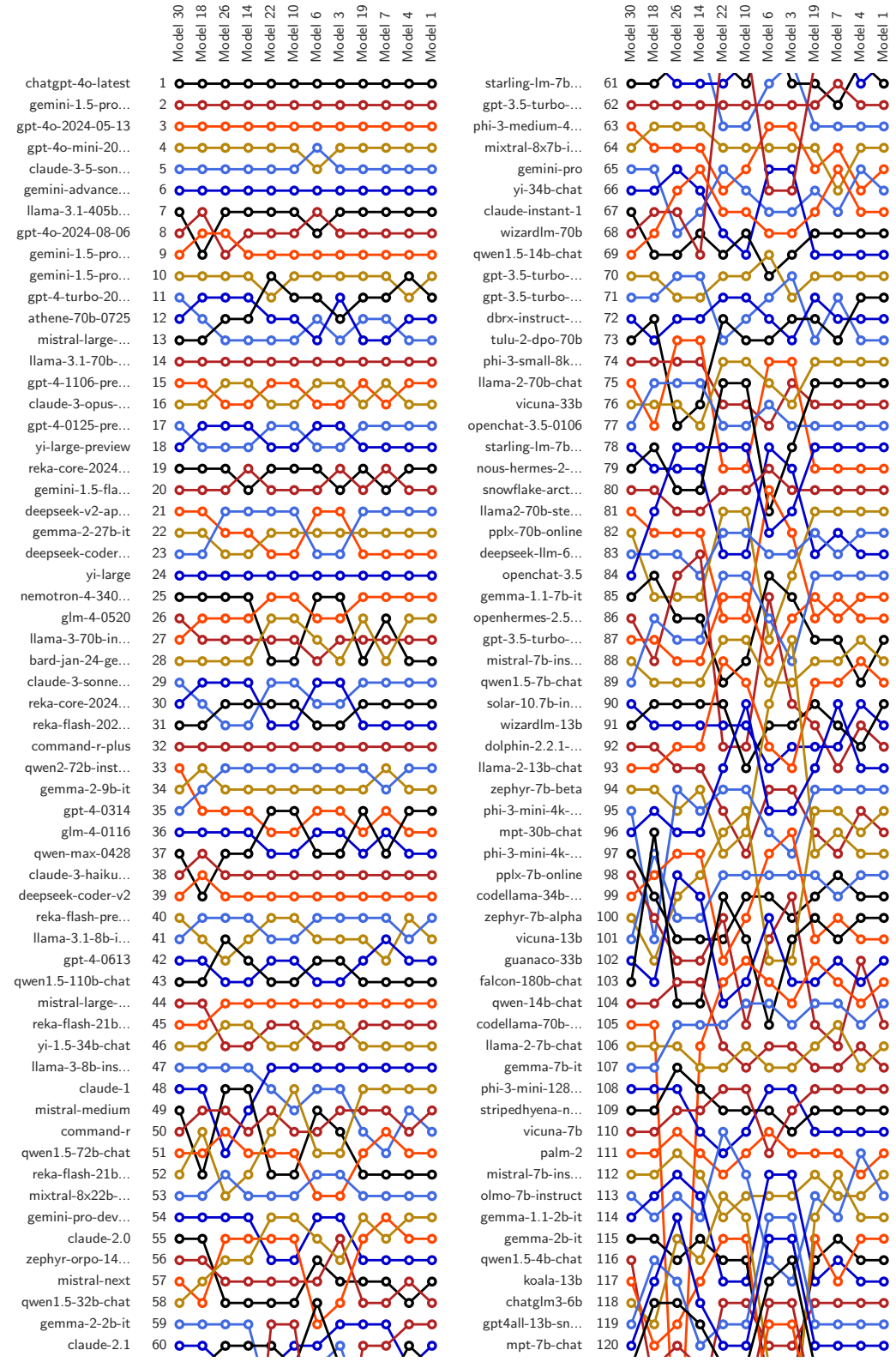

Figure E.2: Bump chart comparing chatbot rankings across 12 statistical models, with Model 1 representing the Elo-based ranking method used in Chiang et al. (2024). Models are arranged with increasing complexity from right to left. Lines track changes in ranking for each chatbot.

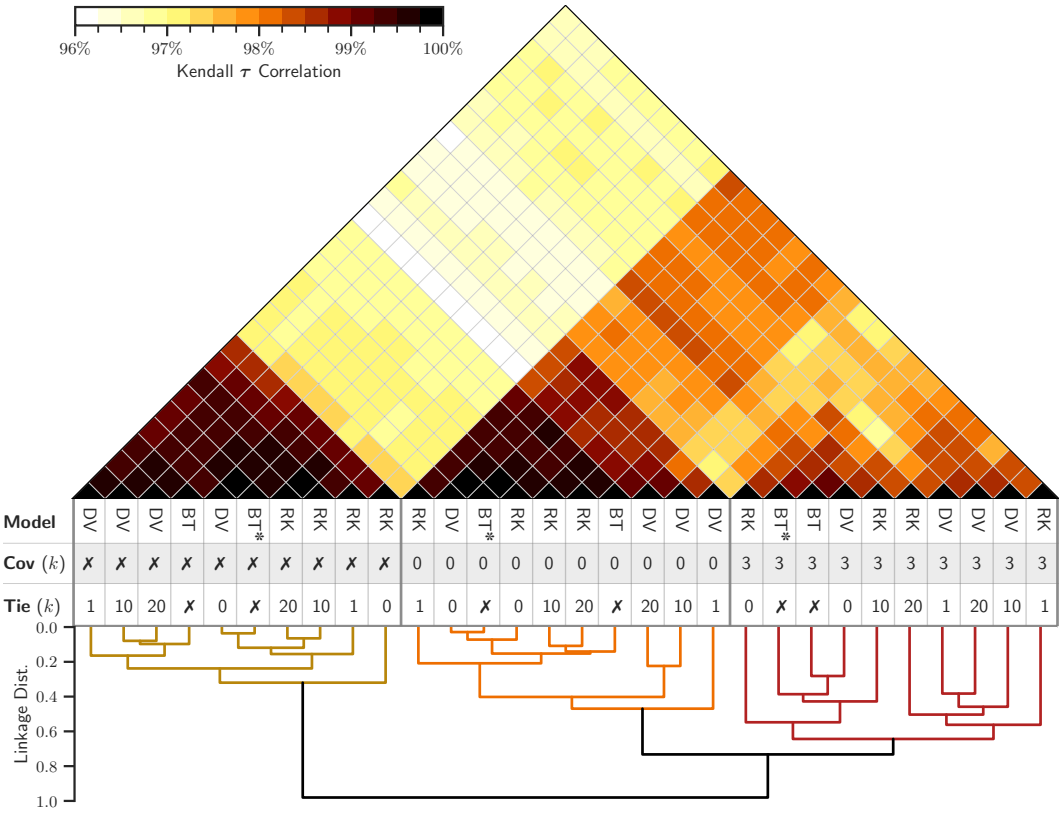

Figure E.3: Kendall $\boldsymbol{\tau}$ ranking correlation matrix for models in Table D.1. The table below categorizes models by configuration: model type (first row), covariance factor k (second row), and tie factor k (third row). In the first row, code names BT, RK, and DV denote Bradley-Terry, Rao-Kupper, and Davidson models, respectively, with BT* indicating the Bradley-Terry model treating ties as half wins and half losses. The model order is determined by hierarchical clustering on Kendall's correlation values, highlighting two main clusters based on the presence of covariance, with a further division within the covariance group based on factor k. A dendrogram below the table illustrates this clustering.

concordant and discordant pairs, normalized by the total number of pairs, $\binom{m}{2}$, as

$$\tau_{pq} = \frac{1}{\binom{m}{2}} \sum_{1 \le i < j \le m} \mathsf{sgn}(x_i^p - x_j^p)\,\mathsf{sgn}(x_i^q - x_j^q). \tag{E.1}$$

This correlation ranges from -1 to 1, where $\tau_{pq} = 1$ indicates identical rankings, and $\tau_{pq} = -1$ implies a complete reversal in ranking order (i.e., $x_i^p < x_j^p$ implies $x_i^q > x_j^q$ and vice versa). The probability that a pairwise order $x_i^p < x_j^p$ in one ranking aligns with $x_i^q < x_j^q$ in another is $\frac{1}{2}(\tau_{pq} + 1)$ (Gibbons & Chakraborti, 2003, p. 410).

In this analysis, we compute the Kendall correlation matrix $\boldsymbol{\tau} = [\tau_{pq}]$, $p, q = 1, \ldots, 30$, between each pair of models in Table D.1 using Kendall's τ-b method, which also accounts for ties in the scores (Kendall, 1945).

Figure E.3 shows the resulting $\boldsymbol{\tau}$ matrix, where each cell represents the Kendall correlation between two models. We present only the lower-triangular half of this symmetric matrix for clarity. An adjacent table below the matrix describes each model's type (first row), covariance factor k (second row), and tie factor k (third row). The codes BT, RK, and DV represent Bradley-Terry, Rao-Kupper, and Davidson models, respectively, while BT* denotes the Bradley-Terry model with ties treated as half win and half loss. Across our 30 models, the Kendall correlation ranged from 0.96 to 1, indicating overall similarity in rankings but with distinguishable differences driven by model

parameter variations. Further insights into these parameter-driven distinctions emerge by reordering the correlation matrix, as detailed in the next section.

E.4 Identifying Ranking Similarities via Hierarchical Clustering

To better interpret the distinctions between models, we performed hierarchical agglomerative clustering (Hastie et al., 2009, Section 14.3.12) on the distance matrix $\mathbf{J} - \boldsymbol{\tau}$, where $\mathbf{J}$ is a matrix of all ones, converting $\boldsymbol{\tau}$ into a dissimilarity measure. Using optimal leaf ordering (Bar-Joseph et al., 2001), this clustering reorders the rows and columns of $\boldsymbol{\tau}$ to reveal natural groupings based on ranking similarity across models.

The ordering of models shown in Figure E.3 is directly the arrangement produced by hierarchical clustering, visualized in the dendrogram below the figure, which reveals a distinct block structure in the $\boldsymbol{\tau}$ matrix. The clustering first divides the models into two main groups: those without covariance (indicated by ✗, columns 1 to 10, shown by the yellow branch) and those with covariance (columns 11 to 30). Within the group of models with covariance, further subdivision occurs, with models having $k = 0$ (columns 11 to 20, shown by the orange branch) forming a distinct sub-block from those with $k = 3$ (columns 21 to 30, shown by the red branch). This hierarchical structure highlights that the presence and type of covariance parameter k are primary factors influencing ranking similarity, more so than the tie factor k.

While covariance modeling prominently influences ranking consistency, earlier results (see Section 3 and Appendix D) demonstrated that our generalized tie modeling significantly enhances inference and predictive accuracy. This dual impact—covariance structure shaping ranking alignment and generalized tie modeling improving model fit and accuracy—illustrates the complementary strengths of these two generalizations in paired comparison models.

Appendix F Relationship Between LLM Characteristics and Scores

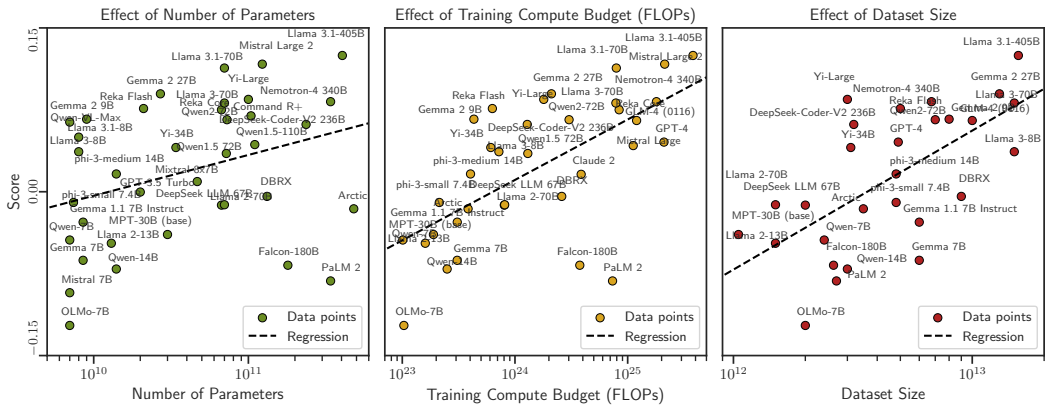

Figure F.1: Scatter plots showing the effect of LLM characteristics on scores. The abscissa represents the logarithmic scale of the characteristics: (left) number of parameters, (middle) computational budget (FLOPs), and (right) dataset size. The ordinate (y-axis) is shared across all panels and shown only for the left panel. Each point is labeled with the corresponding LLM name, and the regression lines are shown in dashed black.

This section explores the relationship between the scores derived from our ranking framework and three key characteristics of large language models: the number of parameters, computational budget (FLOPs), and dataset size. Using data from Epoch AI (2022), which provides these attributes for various LLMs, we matched these models with those evaluated in our analysis. Among the matched models, 37, 34, and 28 LLMs had available values for the number of parameters, FLOPs, and dataset size, respectively. Figure F.1 presents scatter plots illustrating these relationships, with the abscissa displayed on a logarithmic scale and dashed regression lines capturing the linear trends. The scores,

shown on the ordinate, are derived from our generalized Rao-Kupper model corresponding to model 18 of Table D.1.

Table F.1: Regression results examining the relationship between LLM characteristics and scores.

Variable	**Coeff** ($\times 100$)	p**-Value**	R^2	**Pearson** r
Number of Parameters	1.62 ($\pm$0.76)	4.028%	0.11	0.34
Training Compute (FLOPs)	2.40 ($\pm$0.51)	0.004%	0.41	0.64
Dataset Size	5.53 ($\pm$1.29)	0.022%	0.41	0.64

Table F.1 summarizes the results of independent regressions performed for each characteristic against the scores. The table reports the coefficients with their standard errors, p-values, coefficient of determination R^2, and Pearson correlations r. These results reveal consistent positive associations between all three characteristics and model scores. Computational budget (FLOPs) and dataset size exhibit the strongest associations, each with $R^2 = 0.41$ and $r = 0.64$. The number of parameters shows a weaker association, reflected by a lower $R^2 = 0.11$ and $r = 0.34$, though it remains statistically significant.

Our findings align with prior studies on the scaling laws of LLMs. Kaplan et al. (2020) emphasize that model performance depends strongly on scale, encompassing model size, dataset size, and computational budget. Hoffmann et al. (2022) further highlight the importance of balancing these factors, demonstrating that compute and dataset size play pivotal roles in maximizing performance within fixed budgets. Consistently, our analysis shows that scores improve across all three characteristics, with particularly strong trends observed for computational budget and dataset size.

The comparatively modest association observed for the number of parameters in our analysis reflects the diminishing returns noted in the literature when model size is increased without proportional scaling of compute and data resources. This observation underscores the importance of balanced scaling for optimal performance, as emphasized by prior studies.

Given the sparsity of available data for proprietary models in our analysis, these conclusions should be interpreted with caution. Further studies incorporating more comprehensive datasets could provide additional insights into the interplay between computational budget, dataset size, and the number of parameters in driving LLM performance.

Appendix G Implementation and Reproducibility Guide

We developed a Python package `leaderbot`[3] that implements the methods presented in this paper. The package allows users to reproduce the numerical results, evaluate model fit, and explore model generalization performance. Below, we provide examples of using `leaderbot` for common tasks such as model training, evaluation, and visualization. The full documentation, including further functionality and customization options, is available online.

G.1 Model Training and Visualization

Listing G.1 demonstrates the basic usage of `leaderbot` for training a statistical model and visualizing results. In this example, we replicate Model 23 from Table D.1 using the `DavidsonFactor` class, which includes both tie modeling and covariance with parameters $k = 0$ for covariance and $k = 0$ for the factor tie model. Once instantiated, the model is trained using the BFGS optimization method on the dataset. While `leaderbot` ships with the data used in this paper, users can download the latest dataset directly from Chatbot Arena through the function `load`, which provides additional options (see documentation for details).

After training, users can use the model for inference and prediction, retrieve the values of the loss function and Jacobian at either the trained optimal parameters or any specified parameter array,

[3] `leaderbot` is available for installation from PyPI at [URL removed for anonymization]. The source code of the package can be found at `https://anonymous.4open.science/r/leaderbot-CA90`.

generate leaderboard tables, and create visualizations—including replicating figures like the score plot in Figure E.1, the 3D kernel-PCA plot in Figure 2, and the match matrix in Figure 1.

Listing G.1: Basic usage of `leaderbot` for model training and visualization of results.

```
# Install leaderbot with "pip install leaderbot"
import leaderbot as lb

# Load default dataset shipped with the package
data = lb.data.load()

# Create Davidson model with covariance factor k = 0 (diagonal covariance)
# and tie factor k = 0. This corresponds to Model 23 in Table D.1
model = lb.models.DavidsonFactor(data, n_cov_factors=0, n_tie_factors=0)

# Train the model
model.train(method='BFGS', max_iter=1500, tol=1e-8)

# Make inference
probabilities = model.infer(data)

# Make prediction
preiction = model.predict(data)

# Compute loss function -ℓ(θ) and its Jacobian -∂ℓ(θ)/∂θ
loss, jac = model.loss(return_jac=True)

# Print leaderboard and plots overall probabilities
model.leaderboard(max_rank=None, plot=True)

# Generates Figure E.1
model.plot_scores(max_rank=50)

# Rank competitors based on their scores
rank = model.rank()

# Visualize correlation similar to Figure 2  using Kernel PCA method
# projected on 3-dimensional space for the top 40 ranks.
model.visualize(max_rank=40, method='kpca', dim='3d')

# Generate a plot similar to Figure 1  with the win/loss matrix W and
# tie matrix T for both observed and predicted probabilities.
model.match_matrix(max_rank=25, win_range=[0.2, 0.6], tie_range=[0.15, 0.4])
```

G.2 Model Evaluation: Fit and Consistency Metrics

Listing G.2 demonstrates the evaluation of model fit and consistency for five selected models. These models are chosen from the broader set of 30 models discussed in Table D.1 and include the original Bradley-Terry model, as well as original and generalized versions of the Rao-Kupper and Davidson models.

In this example, each model is trained on the full dataset to evaluate goodness of fit. The script produces a bump chart similar to Figure E.2, comparing the rankings generated by the five models. Additionally, it provides tables similar to Table D.1 and Table D.2, displaying model selection and goodness-of-fit metrics. These metrics enable users to analyze and compare model consistency in training performance.

G.3 Model Generalization: Performance on Test Data

Listing G.3 demonstrates the evaluation of model generalization using a 90/10 train-test split, where the same five models from the previous listing are trained on 90% of the data and tested on the re-

Listing G.2: Evaluating model fit and consistency metrics across models in `leaderbot`.

```python
import leaderbot as lb

# Load dataset
data = lb.data.load()

# List models 2, 11, 12, 23, and 24 from Table D.1
models = [
   lb.models.BradleyTerryFactor(data, n_cov_factors=0),
   lb.models.RaoKupperFactor(data, n_cov_factors=0, n_tie_factors=0),
   lb.models.RaoKupperFactor(data, n_cov_factors=0, n_tie_factors=1),
   lb.models.DavidsonFactor(data, n_cov_factors=0, n_tie_factors=0),
   lb.models.DavidsonFactor(data, n_cov_factors=0, n_tie_factors=1)
]

# Pre-train the models
for model in models: model.train()

# Compare model rankings, generating a bump chart like Figure E.2
lb.evaluate.compare_ranks(models, rank_range=[1, 60])

# Evaluate model-selection metrics, similar to Table D.1
mod_metrics = lb.evaluate.model_selection(models, report=True)

# Evaluate models for goodness of fit, similar to Table D.2
gof_metrics = lb.evaluate.goodness_of_fit(models, metric='RMSE', report=True)
```

maining 10%. The resulting RMSE, KLD, and JSD metrics, displayed in a table similar to Table D.3, offer insight into each model's predictive accuracy and robustness on unseen data.

Listing G.3: Evaluating model generalization using train-test split in `leaderbot`.

```python
import leaderbot as lb

# Load dataset
data = lb.data.load()

# Split data to training and test data
training_data, test_data = lb.data.split(data, test_ratio=0.1, seed=20)

# List models 2, 11, 12, 23, and 24 from Table D.1
models = [
   lb.models.BradleyTerryFactor(training_data, n_cov_factors=0),
   lb.models.RaoKupperFactor(training_data, n_cov_factors=0, n_tie_factors=0),
   lb.models.RaoKupperFactor(training_data, n_cov_factors=0, n_tie_factors=1),
   lb.models.DavidsonFactor(training_data, n_cov_factors=0, n_tie_factors=0),
   lb.models.DavidsonFactor(training_data, n_cov_factors=0, n_tie_factors=1)
]

# Evaluate models for generalization on test data, similar to Table D.3
gen_metrics = lb.evaluate.generalization(models, test_data=test_data,
                                         train=True, metric='RMSE', report=True)
```