ICLR 2025

Siavash Ameli[1,2]
Siyuan Zhuang[1]
Ion Stoica[1]
Michael W. Mahoney[1,2,3]
@[1]UC Berkeley, [2]ICSI, [3]LBL

A Statistical FRAMEWORK *for* RANKING LLM-BASED CHATBOTS

Overview

Objective: Addressing key challenges in paired-comparison for ranking.

- Incorporating **ties** in user judgements effectively.
- Highlighting the use of **interpretable** metrics to analyze ranking systems.
- Exploring **relationship patterns** beyond ranking.

I: Statistical Models

- Overview of paired comparison
- A generalized model
- Experimental results

II: Statistical Interpretability

- Identifiability of parameters
- Dissimilarity metrics
- Visualizing dissimilarity

III: Ranking

- Comparing Rankings
- Analyze ranking similarities
- LLM Characteristics

I: Statistical Models

Setup

- Connected graph: $\mathcal{G}(V, E)$
- **Competitors**: $V := \{1, \ldots, m\}$
- **Matches**: $E \subseteq \{\{i,j\} \mid i,j \in V\}$
- **Win** frequency: $m \times m$ matrix $\mathbf{W} = [w_{ij}]$
- **Tie** frequency: $m \times m$ matrix $\mathbf{T} = [t_{ij}]$
- Data: $(\mathcal{G}, \mathbf{W}, \mathbf{T})$

Problem Statement

- Probability of i wining j: $P_{i \succ j}$
- Probability of i and j tying: $P_{i \sim j}$
- Parameters: $\boldsymbol{\theta}$, including **scores** $\boldsymbol{x} \in \mathbb{R}^m$
- **Likelihood function**: multinomial distribution

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

Probabilistic Models:

- $P_{i \succ j} + P_{i \prec j} + P_{i \sim j} = 1$
- Stochastic transitivity: $P_{i \succ j}, P_{j \succ k} \geq \frac{1}{2}$, then $P_{i \succ k} \geq \max\{P_{i \succ j}, P_{j \succ k}\}$
- Linear stochastic transitivity: $P = f(x_{ij})$, $x_{ij} := x_i - x_j$

Bradley-Terry (BT) Model:

$$P_{i \succ j} = P(x_{ij} > 0) := \frac{1}{1 + e^{-x_{ij}}}$$

- **Elo rating** is based on BT.
- Does not predict tie, i.e., $P_{i \sim j}$
- Over **a third** of Chatbot Area's data are cases of tie.

Rao-Kupper (RK) Model:

$$P_{i \succ j} = P(x_{ij} > \eta) := \frac{1}{1 + e^{-(x_{ij} - \eta)}}$$

$$P_{i \sim j} = P(|x_{ij}| \leq \eta) := (e^{2\eta} - 1) P_{i \succ j} P_{i \prec j}$$

- η: threshold parameter. $\eta = 0$ falls back to BT.
- Other models also exist that incorporate tie.
- $10 \sim 20\%$ prediction error in Chatbot Arena's data.

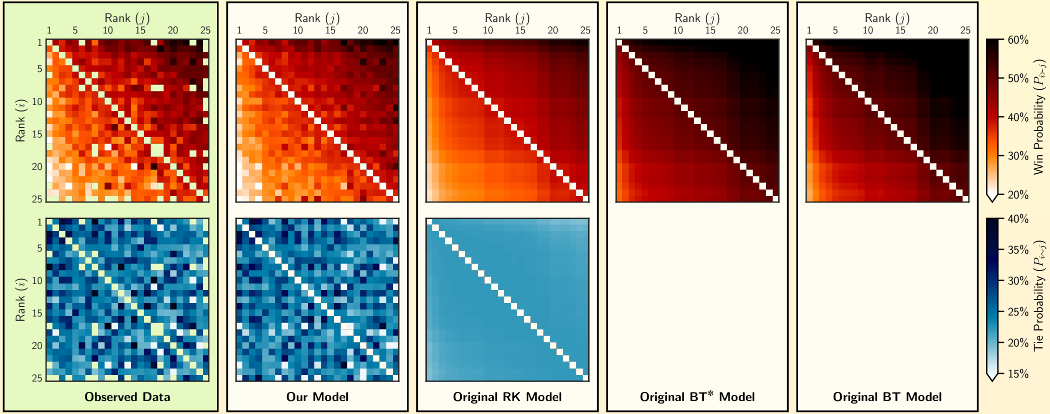
A Generalization for Models with Tie

- Pair-specific threshold: $\mathbf{H} = [\eta_{ij}]$.
- **Low**-rank **additive** factor model:

$$\boxed{\mathbf{H} := \mathbf{G}\boldsymbol{\Phi}^\intercal + \boldsymbol{\Phi}\mathbf{G}^\intercal}$$

- $\mathbf{G}$: parameters. $\boldsymbol{\Phi}$: **orthogonal basis** functions.

Comparison of Models' Predictions



- **Win** (top) and **tie** (bottom) probabilities for 25 LLMs
- **Observed** data (first column), model **predictions** (other columns)
- Our model best matches the observed data.

II: Statistical Interpretability

Thurstonian Model

- Scores as stochastic process: $\boldsymbol{x} = \boldsymbol{\mu} + \boldsymbol{\epsilon}$
- $\boldsymbol{\mu}$: deterministic component
- $\boldsymbol{\epsilon}$ stochastic component with covariance $\boldsymbol{\Sigma}$
- $\boldsymbol{\Sigma}$: **comparative dispersion**
- Diagonal plus low-rank **factor model** for covariance:

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

Dissimilarity Metric

Define the map $\mathcal{S} : \mathbb{S}^m \to \mathbb{S}^m_\circ$

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

- $\mathbf{S}$: **discriminal dispersion**.
- $\mathbf{S} = [s_{ij}]$: variances of score differences $x_i - x_j$

Non-uniqueness of Covariance

- Multiple $\boldsymbol{\Sigma}$ correspond to the same $\mathbf{S}$.
- Equivalence class of covariances:

$$[\boldsymbol{\Sigma}] = \mathcal{S}^{-1}(\mathbf{S}) = \{\boldsymbol{\Sigma}' \in \mathbb{S}^m \mid \mathcal{S}(\boldsymbol{\Sigma}') = \mathbf{S}\}$$

- $[\boldsymbol{\Sigma}]$ defines **invariance of likelihood** function.

Proposition (Equivalence Class of Covariance)

*The map $\mathcal{S} : \mathbb{S}^m \to \mathbb{S}^m_\circ$ is a surjective, **non-injective** linear transformation. Its kernel is given by*

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

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

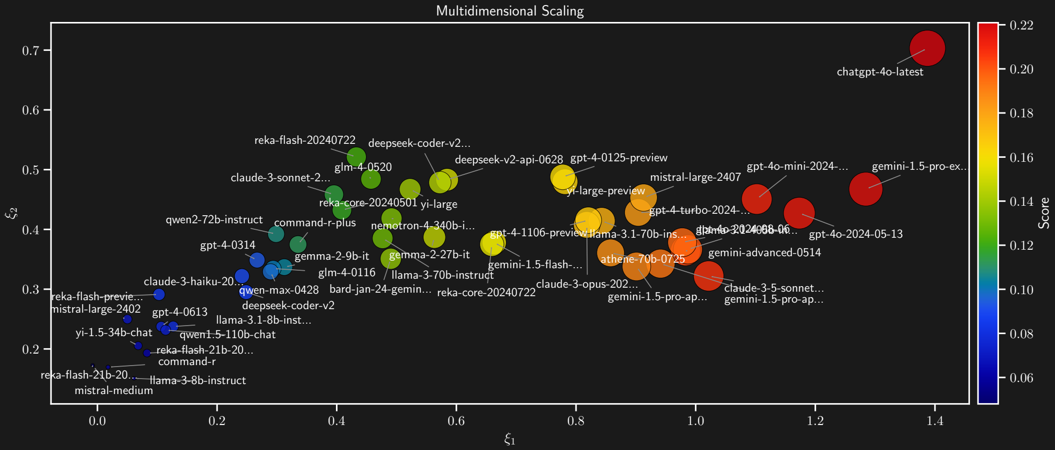
$$[\boldsymbol{\Sigma}] = \{\boldsymbol{\Sigma} + \boldsymbol{v}\mathbf{1}^\intercal + \mathbf{1}\boldsymbol{v}^\intercal \mid \boldsymbol{v} \in \mathbb{R}^m\},$$

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

Takeaway

- $\boldsymbol{\Sigma}$ is unique modulo **symmetric rank one** matrix.
- Covariance is **non-identifiable**.
- Same applies to: confident intervals, errors, etc.
- Dissimilarity metric is **unique**, and **identifiable**.

Visualizing Dissimilarity I

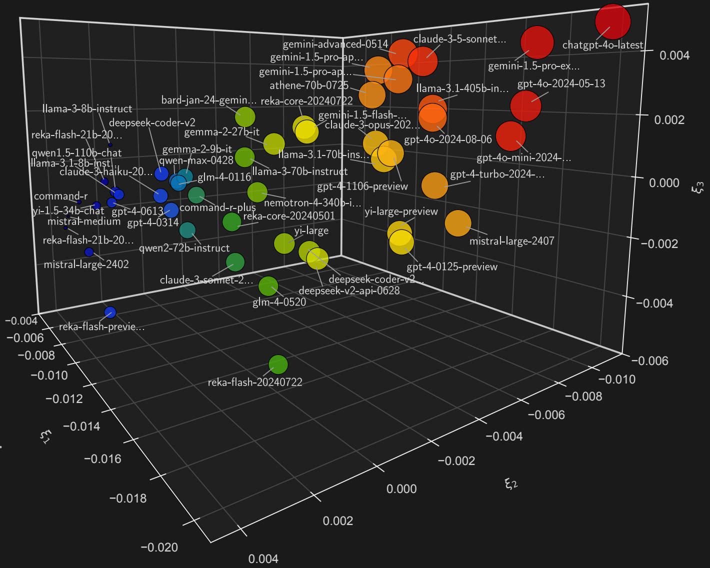


Multi-Dimensional Scaling (MDS)

- 2D projection of LLMs' dissimilarities.

- **Distances** reflecting **dissimilarities**.
- Circle size/color indicate **scores**.

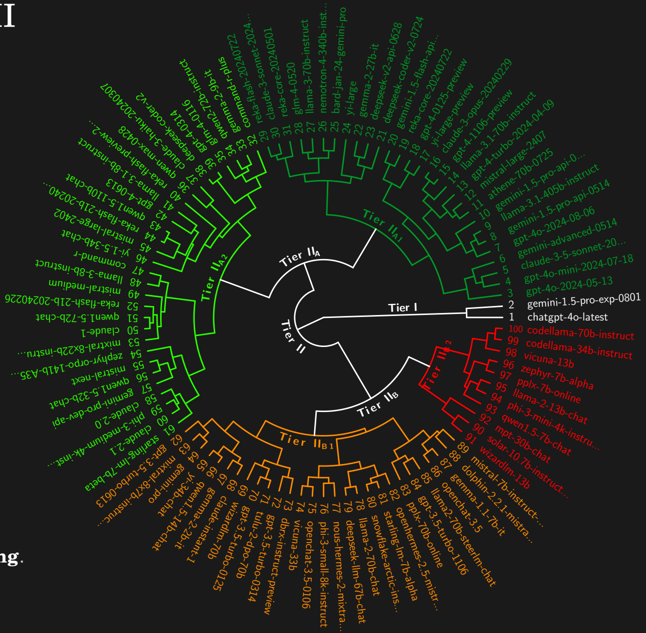
Visualizing Dissimilarity II



Kernel PCA

- 3D Projection of LLMs' dissimilarities.
- **Distances** reflecting **dissimilarities**.
- Circle size/color indicate **scores**.

Visualizing Dissimilarity III

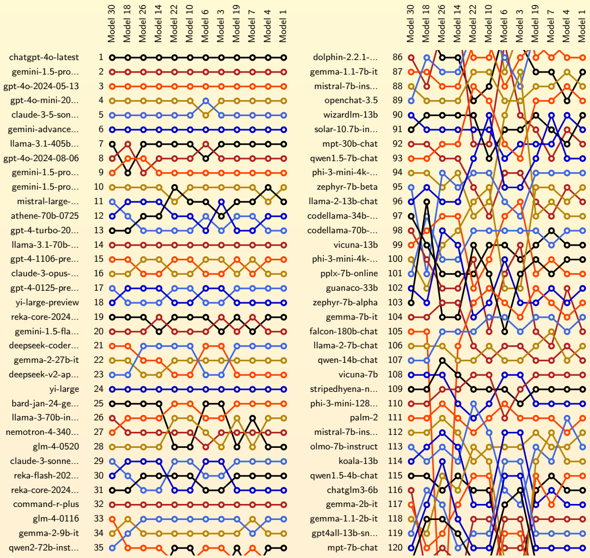


Hierarchical Clustering

- Derived from **dissimilarity** matrix.
- Optimal leaf ordering reveals **tiers**.
- Performance tiers **aligns with ranking**.

III: Ranking

Comparison of Ranking Systems



Bump chart of ranking

- Comparing rankings across 12 models.
- Each line tracks a chatbot's rank.
- Models ordered: simpler (right) to complex (left).
- Rankings at the top remain similar.
- Greater divergence observed at lower ranks.

Correlation Among Ranking Systems

1. Kendall τ Correlation

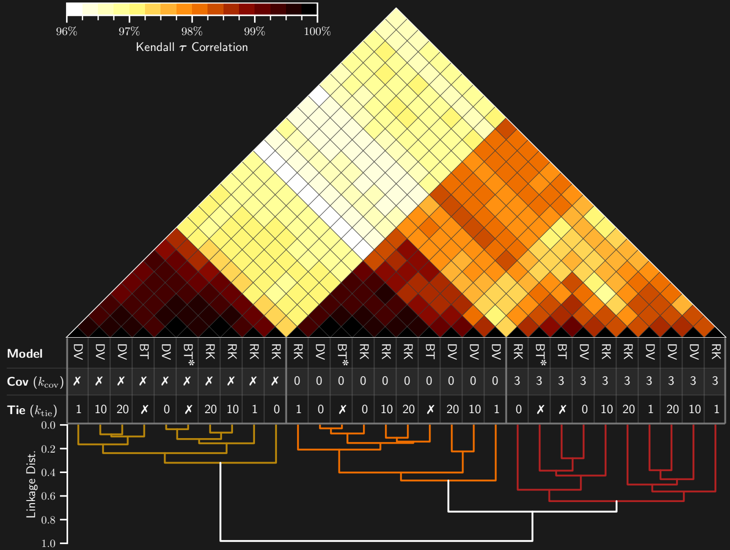
- Analyze ranking similarity among 30 models.
- Models ordered using **optimal leaf ordering** via hierarchical clustering.

2. Model Settings

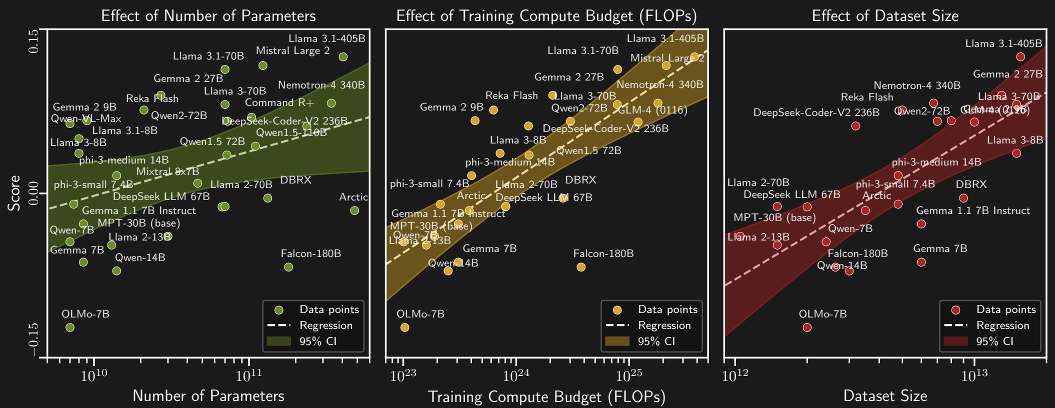
- First row: model types.
- Second/third rows: model features

3. Hierarchical Clustering

- Dendrogram reveals two primary clusters.
- Models with and without **covariance** form distinct groups.



LLM Characteristics and Ranking



- Strong correlation with **FLOPs** and **dataset size**.
- Weaker correlation with number of parameters.
- Reflecting **scaling laws**.
- May **predict** ranking of future LLMs.

Takeaways

- Including **ties** significantly improves model **predictions**.
- **Covariance** has a notable impact on **ranking**.
- Dissimilarity metrics offer a more **interpretable** alternative.

Explore

- Try our Python package: `https://`**`leaderbot`**`.org`
- Check out **`/poster`**, and **`/paper`**