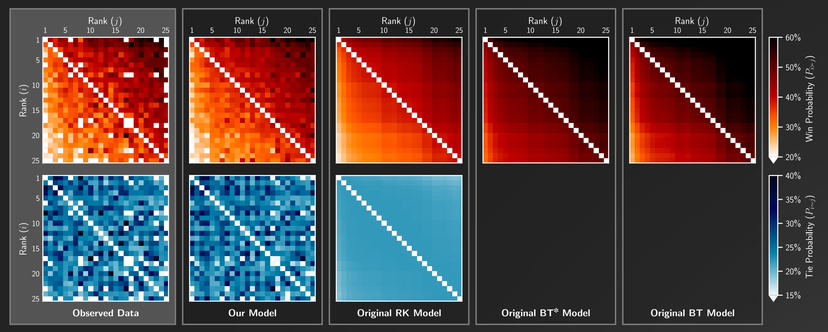
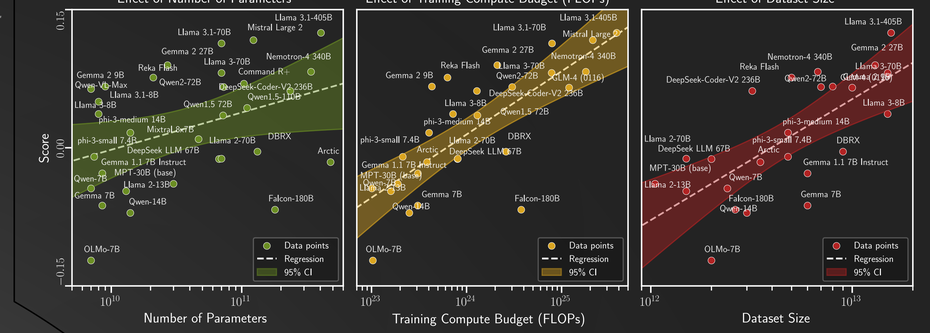
A Statistical FRAMEWORK *for* RANKING LLM-BASED CHATBOTS

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

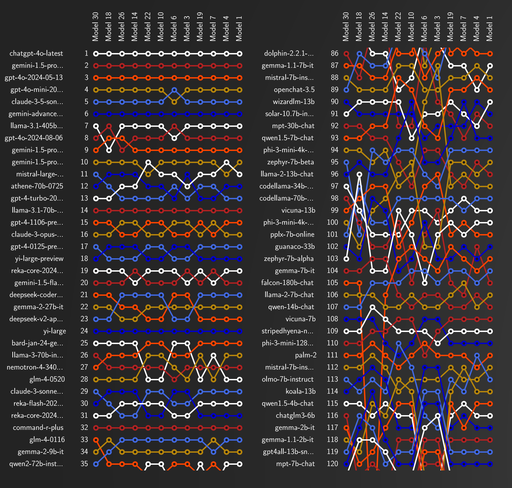
https://leaderbot.org



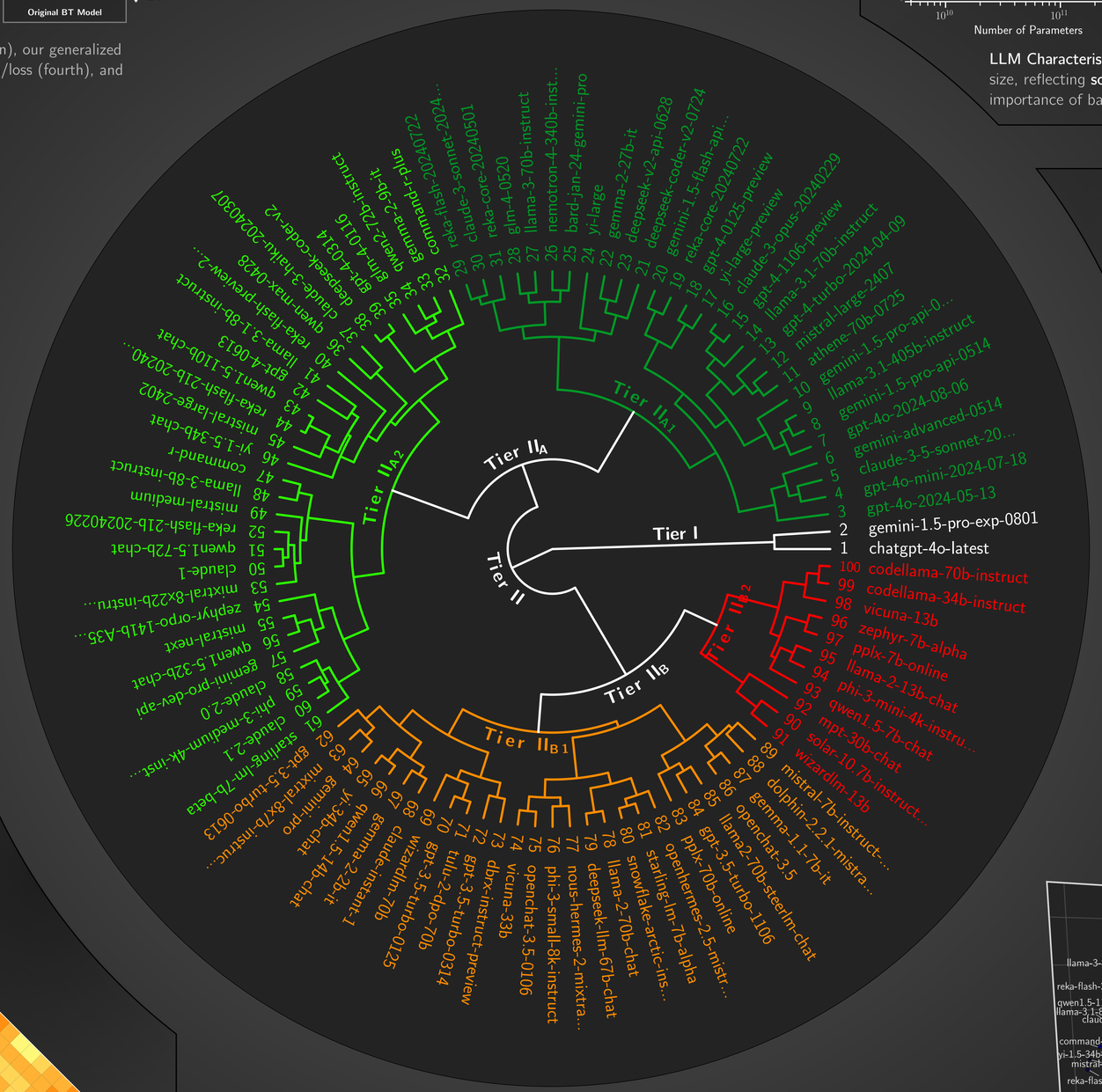
Win (top) and tie (bottom) probabilities for 25 LLMs: observed data (first column), our generalized model (second), Rao-Kupper with ties (third), Bradley-Terry with ties as half win/loss (fourth), and without ties (fifth). Our model best matches the observed data.

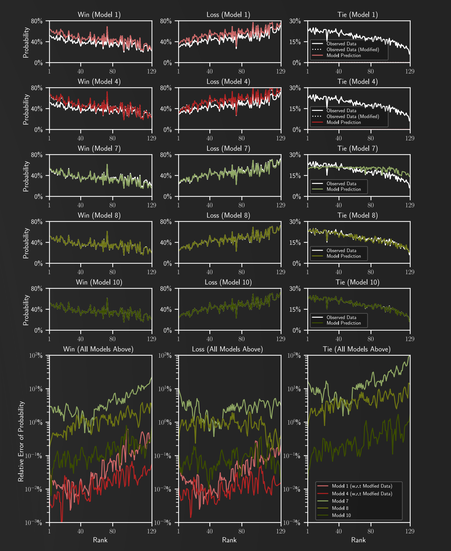


LLM Characteristics and Ranking: Scores **strongly correlate** with computational budget and dataset size, reflecting **scaling laws**. The number of parameters shows diminishing returns, emphasizing the importance of balanced scaling.LM Characteristics and Scores.



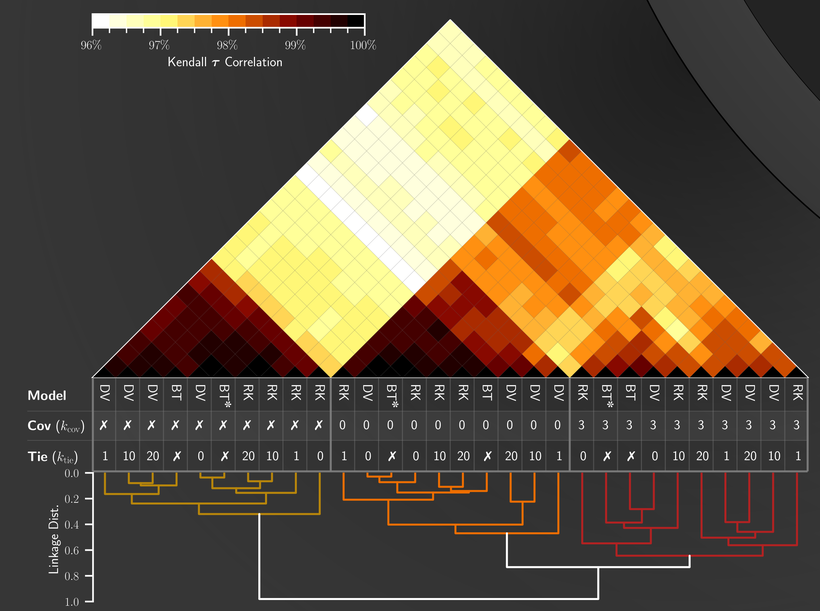
Bump chart visualizing chatbot rankings across 12 models, from simpler models on the right to complex ones on the left. Each line tracks a chatbot's rank, highlighting stable top rankings and diverging lower ranks with increasing model complexity.

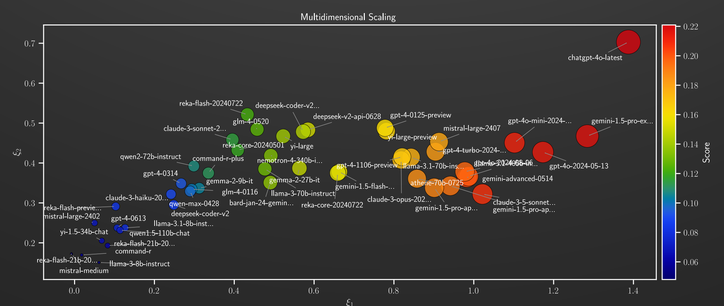




Predicted vs. empirical marginal probabilities of win, loss, and tie. Our models improve tie prediction accuracy while also enhancing win/loss predictionsarginal Probabilities Across Models

Kendall correlation matrix of rankings among 30 models, reordered using optimal leaf ordering via hierarchical clustering. The dendrogram reveals two primary clusters, indicating that the inclusion or exclusion of covariance is the key factor driving ranking similarity.





MDS projection of dissimilarities into 2D space. LLMs are spatially arranged based on pairwise differences, with size and color showing scores. The plot aligns relative scores with dissimilarities, uncovering meaningful patterns.

Kernel PCA projection of LLMs' dissimilarities into 3D space. Distances reflecting pairwise dissimilarities. Circle size and color indicate scores, highlighting clusters of similar performance.