Elo Uncovered:
Robustness and Best Practices in Language Model Evaluation

Meriem Boubdir*, Edward Kim†, Beyza Ermis*, Sara Hooker*, Marzieh Fadaee*

*Cohere For AI, †Cohere

NEURAL INFORMATION PROCESSING SYSTEMS

Cohere For AI

Motivation

- The evaluation of language models constitutes a critical yet **resource-intensive** undertaking, requiring substantial time, cost, and specialized human expertise.
- Despite advancements in automated metrics, **human feedback** remains key to assessing language models performance.
- The **Elo rating system**, designed for **dynamic** games like chess, provide a robust, dynamic, and interpretable framework for comparing models' capabilities.
- LLMs, unlike dynamic competitors that evolve with time, have **static capabilities** and operate in a **time-agnostic** context. This static nature prompts a critical investigation of Elo's suitability and reliability when applied to LLM evaluations.

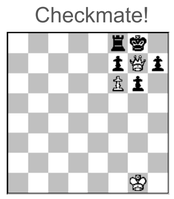
The Elo Rating System

- The rating system assigns scalar values to players' skill level, adjusting dynamically based on match outcomes in a zero-sum manner.

- Each match outcome affect future rankings by updating the ratings according to:

$$E_A = \frac{1}{1 + 10^{(R_B - R_A)/400}}$$

Initial Elo scores
Player A: 1950
Player B: 2000



$$R'_A = R_A + K(S_A - E_A)$$

Updated Elo scores
Player A: 1943 (-7pts)
Player B: 2007 (+7pts)

Desirable Properties of Elo

Our research focuses on two desirable properties of the Elo rating system:

1. Reliability:
sensitivity to hyperparameters (K-Factor) and match ordering.

2. Transitivity:

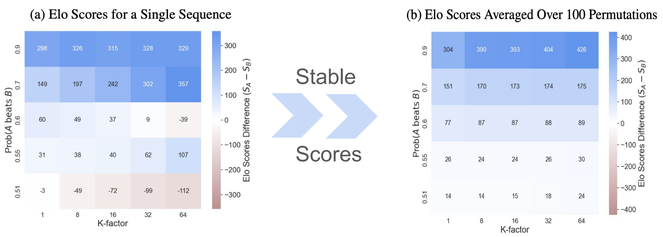
$$A > B \quad \text{and} \quad B > C \implies A > C$$

Rank	Model	Elo Rating
1	GPT-4	1225
2	Claude-v1	1195
3	Claude-instant-v1	1153

References:
[Elo] Arpad E. Elo. The Rating of Chessplayers, Past and Present. Arco Pub., New York, 1978.
[LMSYS-ChatBot] Wei lin Chiang, Lianmin Zheng, Lisa Dunlap, Joseph E. Gonzalez, Ion Stoica, Paul Mooney, Sohier Dane, Addison Howard, and Nate Keating. Lmsys - chatbot arena human preference predictions. https://kaggle.com/competitions/lmsys-chatbot-arena, 2024. Kaggle.

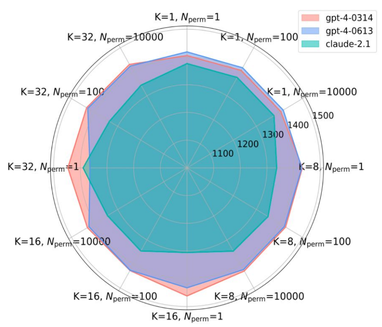
How Robust Are Elo Scores?

- **Methodology:** We model LLM evaluations as **Bernoulli** processes to simulate binary win/loss outcomes between two models **A** and **B**.
  → Allow us to rigorously test the Elo system behavior under controlled settings, including win rates, match ordering and the K-factor.
- **Findings:**
  a. Elo ratings display sensitivity to the ordering of matches and hyperparameters, such as the K-factor.
  b. Transitivity fails under some conditions, undermining the reliability of rankings derived from Elo scores.
  c. Volatility in Elo ratings becomes more prominent when the win rates are similar ($P_{win}(A) \approx P_{win}(B)$).

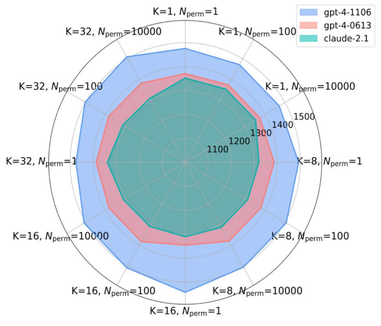


Validation on Real-World Data

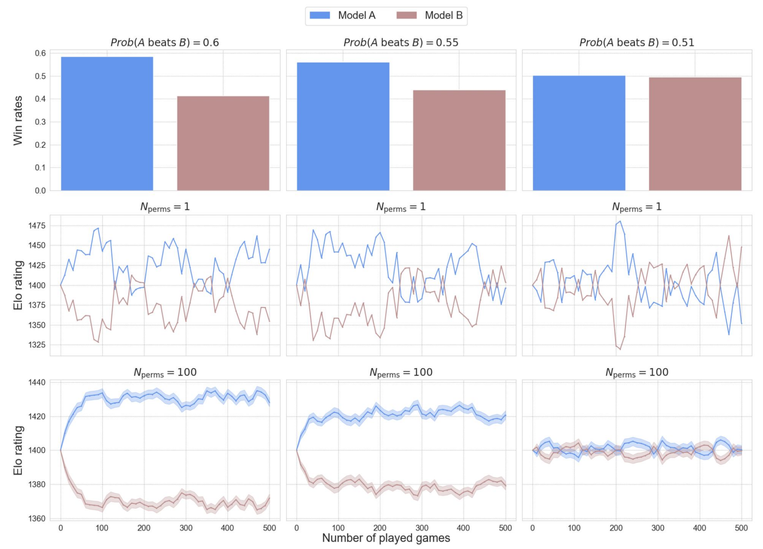
- We extend our synthetic analyses to real-world application by using human feedback from the open-source **LMSYS-ChatBot** dataset.
- We show the Elo scores for 3 models under various permutations N_{perms} and K-factor settings, and *how their final ranking is affected*.



(a) GPT-4-0314 vs. GPT-4-0613 and GPT-4-0613 vs. Claude-2.1
Recorded Win rates: 0.51 vs 0.49 and 0.61 vs 0.39

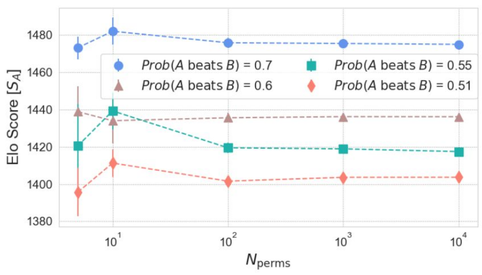


(b) Gpt-4-1106-preview vs. Gpt-4-0613 and Gpt-4-0613 vs. Claude-2.1
Recorded Win rates: 0.67 vs 0.33 and 0.61 vs 0.39



Guidelines for **Robust** Elo Ratings

1) Elo Score Stability



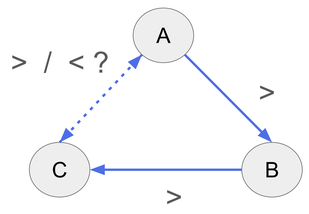
2) K-Factor Tuning

If Wins(A) ≈ Wins(B): *low K values*

Else: *large K values for a faster convergence of Elo*

3) Transitivity

Can be vulnerable for Elo ratings. Final ranking depends on recorded win rates and choice of hyperparameters.



meri.boubdir@gmail.com