Elo Uncovered:
Robustness and Best Practices in Language Model Evaluation

Meriem Boubdir[1], Edward Kim[2], Beyza Ermis[1],
Sara Hooker[1], Marzieh Fadaee[1]

[1]Cohere For AI, [2]Cohere

NeurIPS 2024

Dec 10-15 Vancouver

Cohere For AI

NEURAL INFORMATION PROCESSING SYSTEMS

Challenges in Evaluating LLMs

- Rapid advancements of Large Language Models (LLMs) make the evaluation increasingly complex.
- Automated evaluations fail to capture the subtle nuances and contextual understanding inherent in human language.
- Human feedback via "A vs. B" comparisons has emerged as valuable evaluation but is resource-intensive.
- Elo Rating System in NLP:
  - Originally designed by Arpad Elo for ranking chess players.
  - Adopted to efficiently aggregate and interpret pairwise human evaluations of LLMs.

$$R'_A = R_A + K(S_A - E_A) \quad \text{with} \quad E_A = \frac{1}{1 + 10^{(R_B - R_A)/400}}$$

Is the Elo Rating System suitable for LLMs?

- Elo assumes players skills evolve over time.
- The Sequence of comparisons can influence final ratings.

→ LLMs are static entities; they don't "learn" between matches.

→ Elo ratings for LLMs should be order-agnostic

- Number of comparisons grows quadratically.
  - In chess, all players play against each other in a tournament!
- Choices like the K-factor affect rating updates.
  - In chess, 16 for masters and 32 for novice players.

Desirable Properties for Robust Evals

Transitivity

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

Ensures consistent rankings.

Failure can lead to unreliable model ranking.

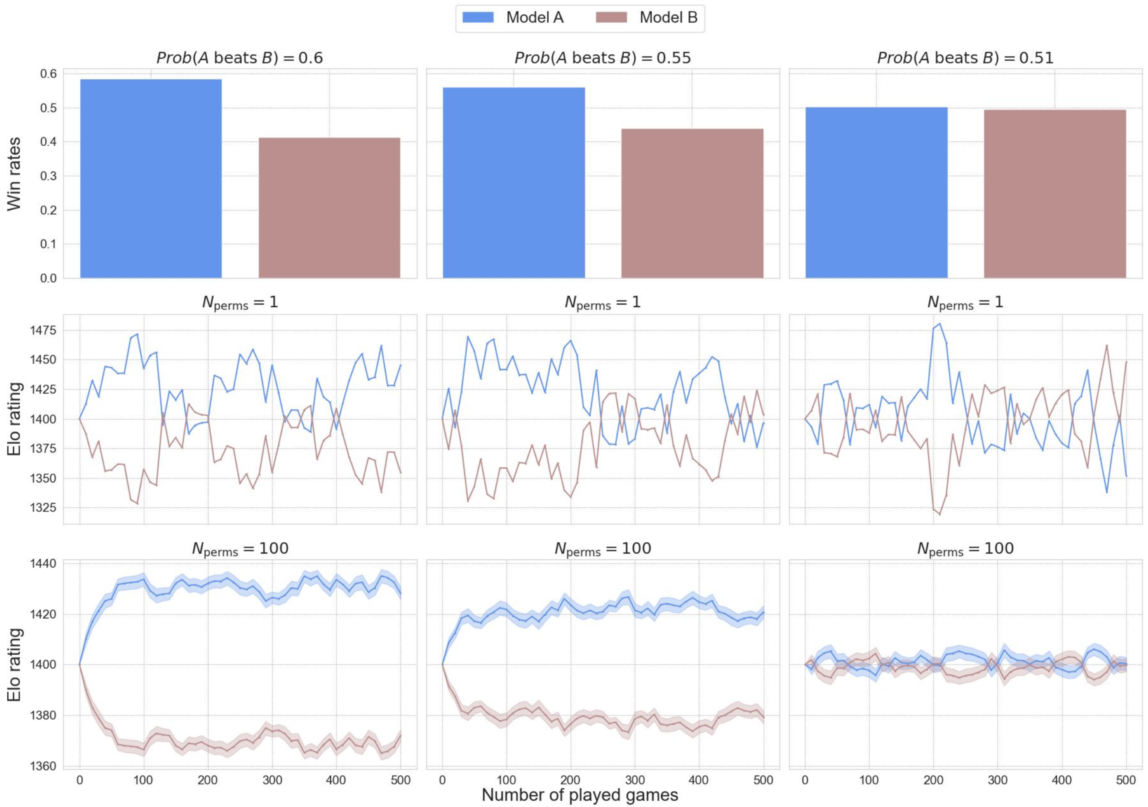
Reliability

Ratings should be robust against:

Ordering of matches.

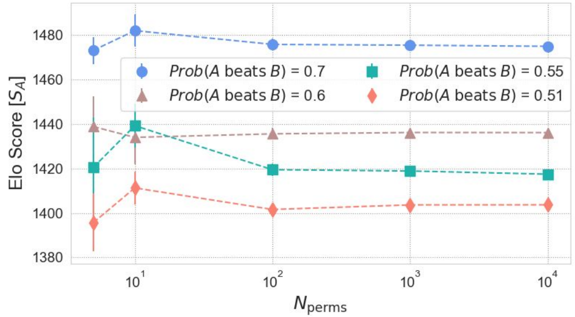
Choice of hyperparameters.

Impact of Match Ordering on Elo Ratings



Elo ratings sensitive to the ordering of matches:

- Early wins can bias subsequent ratings.
- Volatility in Elo scores for win probabilities around 0.5.
- Stability increases with averaging across permutations.



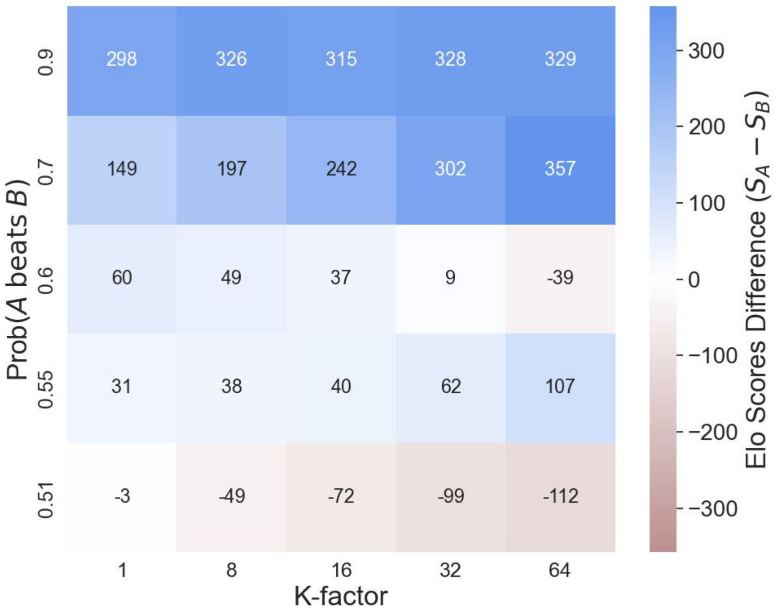
Cohere For AI

Hyperparameter Sensitivity in Elo Ratings

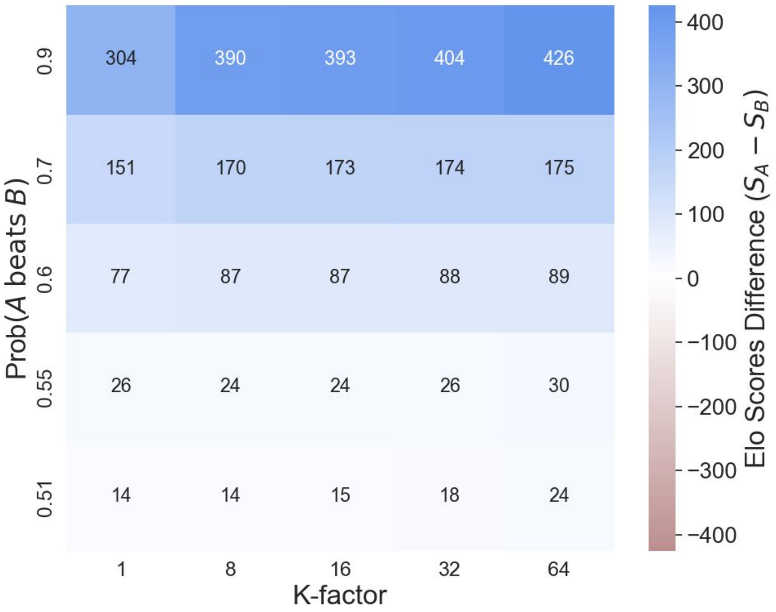
- ➔ K-factor adjusts the rate of update in Elo ratings post-match.
- ➔ High K-factors can lead to rapid but unstable rating updates.

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

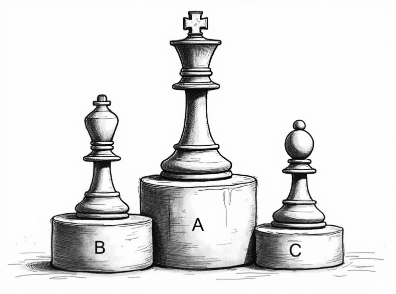
Elo Scores for a Single Sequence



Elo Scores Averaged Over **100** Perms



Can We Always Assume Transitivity?



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

Scenarios:

♔ A > B and B > C with high win probabilities (P_{win} = 0.75).

♖ A > B with high P_{win} = 0.75, B > C with P_{win} = 0.51.

♗ A > B with P_{win} = 0.51, B > C with high P_{win} = 0.75.

Observed inconsistent rankings

♘ A > B with P_{win} of 0.54, B > C with P_{win} of 0.51.

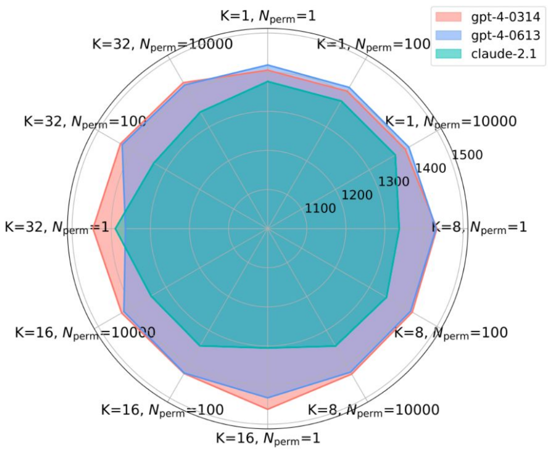
→ Transitivity of Elo ratings can be vulnerable around ~50% win rates.

Cohere For AI

Validation with Real-World Human Feedback

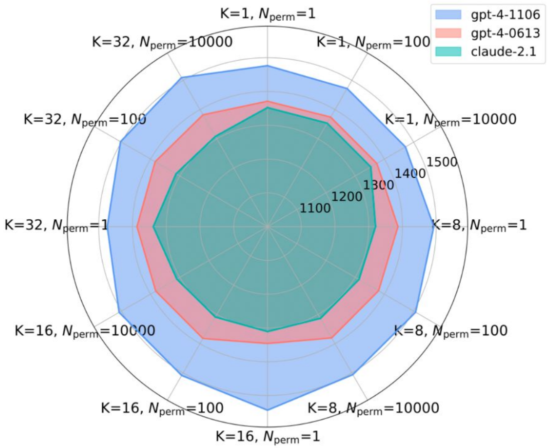
LMSYS Chatbot Arena dataset

Inconsistent Ranking of Models



(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

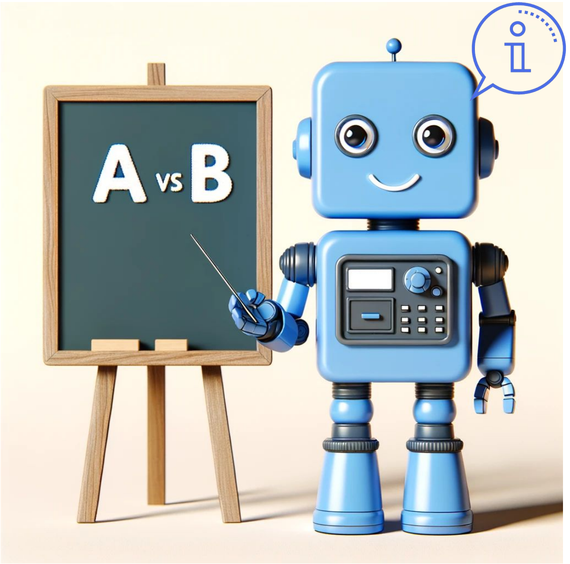
Recorded Win rates: 0.67 vs 0.33 and 0.61 vs 0.39

Consistent Ranking of Models

References: LMSYS Chatbot: Chiang, W.-L., Zheng, L., Dunlap, L., Gonzalez, J. E., Stoica, I., Mooney, P., Dane, S., Howard, A., & Keating, N. (2024). LMSYS - Chatbot Arena Human Preference Predictions. Kaggle.

Guidelines for Robust Elo Evaluation

- **Achieving Score Stability:** Average across a large number of permutations ($N_{perm} \geq 100$).
- **Tuning the K-factor given win rates:**
    - Opt for smaller K-factors when model performances are close.
    - A higher K-factor can quickly differentiate between models with clear performance gaps.
- **Transitivity** is not always guaranteed!



Thank you!

NEURAL INFORMATION
PROCESSING SYSTEMS

Cohere For AI