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]
Cohere for AI[1] Cohere[2]
{meriem,edward,beyza,sarahooker,marzieh}@cohere.com

Abstract

In Natural Language Processing (NLP), the Elo rating system, well-established for ranking dynamic competitors in games like chess, has seen increasing adoption for evaluating Large Language Models (LLMs) through "A vs B" paired comparisons. However, while popular, the system's suitability for assessing entities with constant skill levels, such as LLMs, remains relatively unexplored. Our study investigates the sensitivity and reproducibility of Elo scores for LLMs, integrating both synthetic and human feedback. We show that Elo ratings for LLMs stabilize with 100 or more comparison permutations ($N_{\text{perms}} \geq 100$). A lower K-factor is preferable for closely matched models, whereas a higher K-factor better distinguishes models with clear performance differences. We also report that transitivity ($A > B$ and $B > C$ implies $A > C$) does not consistently hold, particularly when models demonstrate similar performance. Our empirical findings provide guidelines for more reliable LLM evaluation.

1 Introduction

In the rapidly evolving field of Natural Language Processing (NLP), the task of accurately and reliably evaluating LLMs has become increasingly challenging (Liang et al., 2022; Chang et al., 2023; Srivastava et al., 2023; Kaddour et al., 2023; Pozzobon et al., 2023). Human feedback has emerged as an indispensable tool in this performance assessment process, serving as a qualitative metric that captures nuances that automated scoring mechanisms often fail to address (Askell et al., 2021; Bai et al., 2022a,b; Srivastava et al., 2023; Ding et al., 2023; Dettmers et al., 2023).

These human-centered evaluations, highly valuable to the overall progress of the NLP field, typically adopt an *"A vs B"* comparative setup, turning evaluations into a zero-sum game between language models.

This paired feedback structure (Zhao et al., 2023) naturally lends itself to the Elo rating system (Elo, 1978), originally designed for ranking chess players for better matchmaking.

Variants such as Glicko (Glickman, 1995, 1999, 2012) and TrueSkill™ (Herbrich et al., 2006; Minka et al., 2018) have incorporated more complex statistical methods into the original Elo framework, to address some of the limitations of the Elo system, particularly in the context of games with more than two players or teams, or games with more complex outcomes than just win or loss. There is ongoing research into the efficacy of these systems in diverse and dynamic environments, and new methods continue to be developed (Dehpanah et al., 2021; Bertrand et al., 2023).

Despite these limitations, the core principles of Elo have proven to be incredibly resilient and adaptable. As a result, the Elo system has found diverse applications, from predicting sports events outcomes (Binder and Findlay, 2009; Hvattum and Arntzen, 2010; Leitner et al., 2010; Wise, 2021), and facilitating matchmaking in massively multiplayer online games like StarCraft II and Dota (Ebtekar and Liu, 2021; Reid; Liquipedia; ESL), to its recent use in the evaluation of LLMs (Askell et al., 2021; Bai et al., 2022a,b; Srivastava et al., 2023; Ding et al., 2023; Dettmers et al., 2023; Wu et al., 2023; Lin and Chen, 2023).

However, its application to LLM evaluations landscape has been insufficiently studied. Unlike dynamic competitors that evolve, LLMs have static capabilities and operate in a time-agnostic context. In this setting, not only are LLM evaluations unconstrained by tournament timelines or predefined match sequences, but the ordering of matches can also significantly influence the final Elo scores and, consequently, models rankings. This oversight is especially concerning, given the direct impact of Elo system rankings on both research directions and real-world applications in NLP.

Proceedings of the Third Workshop on Natural Language Generation, Evaluation, and Metrics (GEM), pages 339–352
December 6, 2023 ©2023 Association for Computational Linguistics

This study aims to close this research gap by scrutinizing both the reliability and limitations of the Elo rating system when applied to LLMs. Through theoretical and empirical analyses grounded in collected human feedback data, our contributions provide a comprehensive understanding of when and how to reliably employ the Elo system for LLM evaluation, thus offering valuable guidelines for researchers and practitioners in the NLP field.

We find that Elo ratings are far from stable, and are highly sensitive to permutation of ordering and hyperparameter choice. Desirable properties such as transitivity are not always guaranteed, and can be unreliable unless there is comprehensive human feedback data for all unique pairwise comparisons among models in the feedback pool. The sensitivity of Elo ratings becomes more pronounced when dealing with models that exhibit similar performance levels. We illustrate the best practices for mitigating these sensitivities by offering guidelines for hyperparameter selection and matchmaking scenarios.

2 Elo Algorithm Explained

We provide the mathematical formulation of the Elo algorithm, contextualized to the setting of LLM evaluation. In this formulation, let $\mathcal{M}$ be a set of models and each model $i \in \mathcal{M}$ is assigned an initial numerical rating R_i.

Expected Score Computation. For a given paired match-up between two models A and B ($A, B \in \mathcal{M}$), each with respective ratings R_A and R_B, the expected scores E_A and E_B are computed as:

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

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

In this context, the factor of 400 (Elo, 1978) precisely adjusts the sensitivity of the expected score to differences in ratings. A 400-point advantage in ratings translates to a $10 : 1$ odds in favor of the higher-rated model, providing an interpretable metric for performance comparison. For evenly matched models ($R_A = R_B$), both E_A and E_B equate to 0.5, reflecting a $50 : 50$ win probability for both models.

Rating Update Mechanism. Following each match, the Elo ratings are updated based on the observed outcome. The rating adjustment is dictated by the equation:

$$R'_A = R_A + K(S_A - E_A) \tag{2}$$

Here, S_A represents the actual score achieved by model A, which can take on either the value 0 or 1. The K-factor serves as a variable hyperparameter to adapt the rate of change in rating to different scenarios.

Given the costly and time-consuming nature of human evaluations, studying the Elo system's behavior under various scenarios becomes challenging. To circumvent these limitations, we turn to synthetic data generation through Bernoulli processes to simulate various scenarios of human feedback. In the following section, we rigorously evaluate the Elo rating system's robustness and reliability using synthetic data, ensuring it upholds desirable properties like transitivity when rating LLMs.

3 Synthetic Human Feedback

This time-agnostic and independent setup of LLM evaluations resembles a Bernoulli process(Bernoulli, 1713), a sequence of independent experiments, each with two possible outcomes; one model outperforming the other. We use this synthetic setting where we can control characteristics of the distribution to evaluate different desirable properties of a rating system. In this controlled setting where we can precisely control the data distribution, we ask whether the Elo score respects **transitivity** and quantify the degree of sensitivity to **ordering of models** and **hyperparameter choices** like the K-factor.

3.1 The Bernoulli Analogy

Pairwise comparisons in LLM evaluation draw parallels with the foundational principles of the Bernoulli experiment in probability theory. This section delves into the similarity between human feedback-based evaluations and the Bernoulli experiment's principles.

Preliminaries. A Bernoulli trial is a random experiment with exactly two possible outcomes, "success" or "failure". These outcomes adhere to the condition:

$$P(A) + P(A^c) = 1 \tag{3}$$

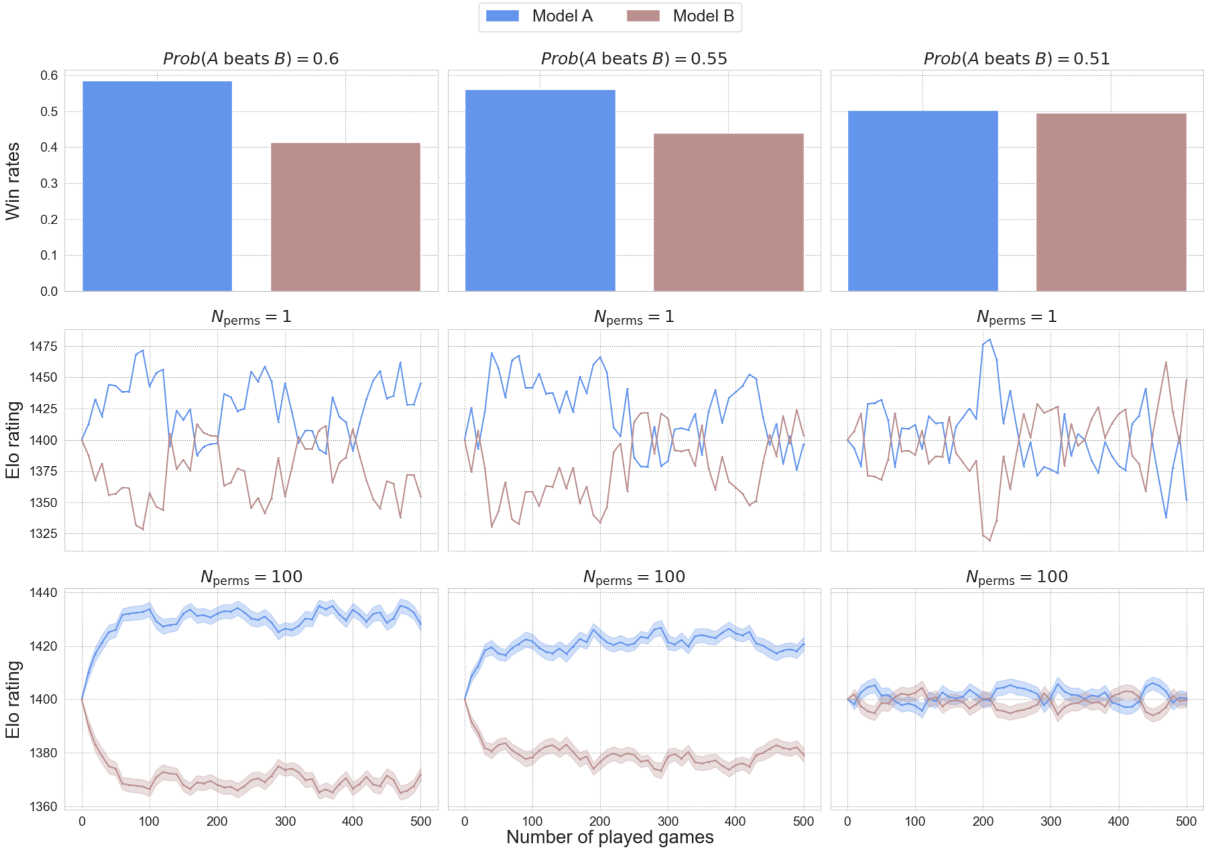


Figure 1: **Impact of win probabilities and permutation sampling on Elo ratings**: Comparing Model A and Model B across three different win probabilities ($Prob(A \text{ beats } B) = \{0.6, 0.55, 0.51\}$) with two levels of permutation sampling ($N_{\text{perms}} = 1$ and $N_{\text{perms}} = 100$). The top row displays the observed win rates, the middle row illustrates Elo ratings with a single permutation, and the bottom row shows the mean and standard error of the mean (SEM) of Elo ratings across 100 permutations.

Here, the random variable $\mathcal{X}$ denotes the outcome, where $\mathcal{X} = 1$ implies success, and $\mathcal{X} = 0$ signifies failure. The probabilities are:

$$P(\mathcal{X} = 1) \quad = p, \quad P(\mathcal{X} \quad = 0) = 1 - p \tag{4}$$

with $0 \leq p \leq 1$, the "success" probability.

Mapping to Human Feedback. When comparing two models, A and B, across N pairwise evaluations, the setup aligns with a Bernoulli process. This process comprises a sequence of independent and identically distributed (*i.i.d*) Bernoulli trials.

To frame this analogy, we designate a win probability, $P(A_{\text{win}})$, to model A. Leveraging a Bernoulli random variable, $\mathcal{X}$, as a means to simulate synthetic human feedback, we proceed as follows:

1. A sample is drawn from $\mathcal{X}$ using $P(A_{\text{win}})$.
2. If $\mathcal{X} = 1$, feedback suggests a preference for model A.
3. Otherwise, model B is favored.

Extending to Multiple Players. Given a finite set of models, $\mathcal{M}$, with n distinct models, their pairwise comparisons can be formulated as:

$$\binom{n}{2} = \frac{n!}{2!(n-2)!} \tag{5}$$

This formula yields $\binom{n}{2}$ unique pairs (A, B) where $A, B \in \mathcal{M}$ and $A \neq B$. For each of these pairs, a Bernoulli process, comprising multiple Bernoulli experiments, is conducted to discern which model performs better over a sequence of trials.

3.2 Synthetic Data Generation

Building upon the Bernoulli process analogy, when conducting multiple independent evaluations between two models, the distribution of the number of times one model is preferred over the other naturally follows a binomial distribution. For N pairwise comparisons, the relation is:

$$P(k; N, p) = \binom{N}{k} p^k (1-p)^{N-k} \tag{6}$$

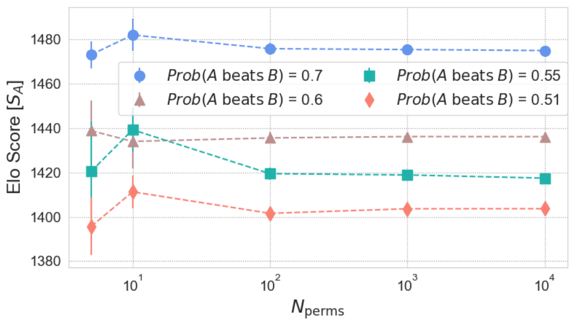


Figure 2: Variation of Model A's average Elo score with increasing number of permutations (N_{perms}) for different probabilities of Model A winning ($Prob(A \text{ beats } B)$). Error bars indicate standard errors of the mean.

where $P(k; N, p)$ is the probability of one model being preferred k times out of N evaluations. p is the success probability and $\binom{N}{k}$ is the binomial coefficient, representing the number of ways to choose k successes from N trials.

4 How Robust Are Elo Scores?

This section defines rigorous stress tests designed to investigate the robustness and overall reliability of the Elo rating system in evaluating LLMs. We focus on critical desirable properties of a ranking mechanism – that it should 1) be insensitive to match-up ordering, 2) not be overly sensitive to hyperparameters like K-factor 3) preserve properties of transitivity. Subsequently, we provide empirically-grounded guidelines for safe and interpretable application of Elo ratings.

4.1 Impact of Ordering on Elo Ratings

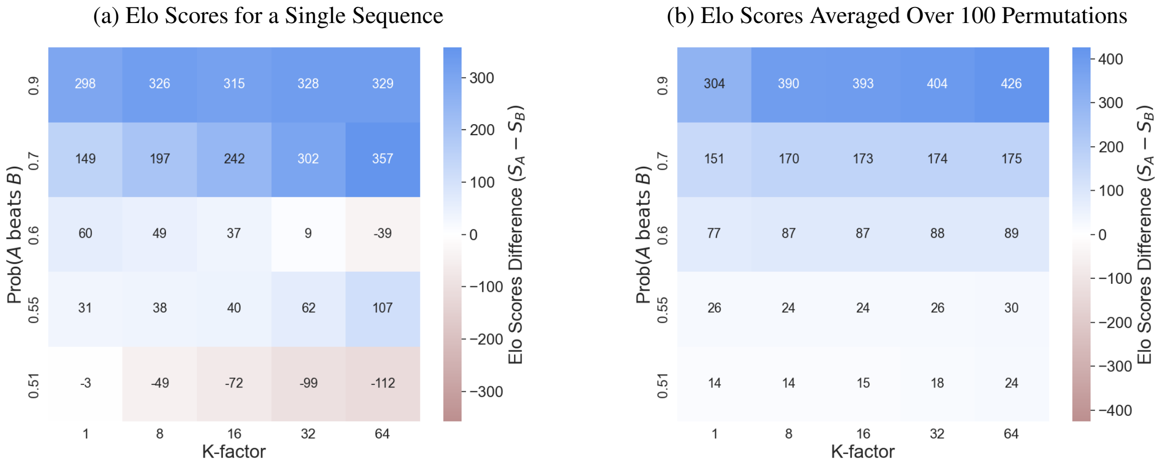
Problem Statement. Unlike chess or time-bound sports where match sequences are structured, in LLM evaluations all matches can occur independently and in parallel, amplifying the sequence's influence on final models ranking. This inherent variability prompts us to investigate the extent to which match-up ordering affects the robustness of Elo ratings.

Experimental Setup. To quantify the effect of match-up ordering on Elo ratings, we generate a baseline sequence of $N_{\text{games}} = 1000$ match outcomes between models A and B, reflecting the scale typical of LLM evaluations via human feedback. We hold N_{games} constant for the entirety of our study to maintain consistency. From this baseline, we derive N_{perms} distinct permutations, each involving a complete reshuffling of the original match outcomes to simulate various chronological orders in which the games might unfold. Crucially, we are not generating new match outcomes for each permutation; rather, we are reordering the existing data to explore the potential impact of different match-up sequences. For each reordered sequence, we update the Elo ratings R_A and R_A according to equation 2, resetting both ratings to an initial value of 1400 at the start of each permutation. Following this, we compute the average Elo ratings per match across all N_{perms} permutations, ensuring a robust analysis that takes into account the full range of possible match-up orders.

We compare ratings' behavior for a set of selected winning probabilities $Prob(A \text{ beats } B)$ = $\{0.51, 0.55, 0.6\}$, inspecting a spectrum of real-world scenarios. N_{perm} is varied from a minimum of 1 to a maximum of 10k, providing a robust sample size for statistical analysis (see Figure 2). Subsequently, we compute the average Elo ratings per match across all permutations. These averages, $\bar{R}_A$ and $\bar{R}_B$. particularly for $N_{\text{perms}} = 1$ and $N_{\text{perm}} = 100$, are visualized to offer insights into the stability of the ratings, as shown in Figure 1.

Key Findings. Our analysis underscores the interplay between winning probability $P(A_{\text{win}})$ and the number of different orderings N_{perm} on the stability of Elo ratings after each update. For $P(A_{\text{win}}) \geq 0.6$, Elo ratings demonstrate high stability; additional results for $P(A_{\text{win}}) = 0.65$ and beyond are available in Appendix B. On the other hand, for $P(A_{\text{win}}) \approx 0.5$, ratings exhibit significant instability for a single sequence. As depicted in Figure 1, when both models have a win probabilities are around 0.5, Elo ratings frequently intertwine, making it challenging to discern a clear performance difference between the two. The instability plateaus as N_{perms} exceeds 100, resulting in stabilized Elo ratings that align closely with the preset winning probabilities. For instance, at $P(A_{\text{win}}) = 0.55$, the average Elo rating for Model A, $\bar{R}_A$, consistently exceeds that for Model B, $\bar{R}_B$, when averaged across multiple permutations, reflecting an accurate performance-based ranking of these models.

These observations validate our concerns highlighted earlier, emphasizing the critical role of N_{perms} for a reliable interpretation of Elo ratings in LLM evaluations. In Elo-based evaluations, the se-



(a) Elo Scores for a Single Sequence

(b) Elo Scores Averaged Over 100 Permutations

Figure 3: Final Elo scores difference ($S_A - S_B$) as a function of K-factor and N_{perms}. Positive values reflect the expected ranking where Model A is superior to Model B, while negative values indicate a discrepancy, falsely suggesting that Model B has a higher Elo score than Model A. We compare between a single sequence of outcomes and averages over $N_{\text{perms}} = 100$ unique permutations.

quence of which models are compared is not a mere procedural detail; it can significantly influence the final Elo scores.

4.2 The K-factor Dependency

Problem Statement. The K-factor in the Elo rating system serves as a crucial hyperparameter scaling constant for rating update and is a key determinant in the rate of convergence to a "true" rating of skill level. While conventional applications like chess use standard K-factor values (16 for experienced players and 32 for novices), these may not be directly applicable in the context of evaluating LLMs due to the unique characteristics and requirements of this domain.

Experimental Setup. We extend our previous approach by conducting tests across a range of winning probabilities and multiple K-factor values $(1, 8, 16, 32, 64)$. We compute and compare the average Elo scores $\bar{S}_A$ and $\bar{S}_B$ for $N_{\text{games}} = 1000$ and $N_{\text{perms}} = \{1, 100\}$. The differences between these final averages for Model A and Model B are summarized in Figure 3 to assess the stability and expected ranking between the two models.

Key Findings. As shown in Figure 3, notable instability is observed in model rankings based on the final Elo scores when we consider a single sequence of paired comparisons (i.e., $N_{\text{perms}} = 1$), especially for winning probabilities nearing 0.5. This instability is markedly exacerbated at higher K-factors. In contrast, the picture changes when coupling higher K-factors with raising the number of permutations to at least 100.

Higher K-factors, in this multi-permutation scenario, speed up the differentiation between models' Elo scores, enabling faster convergence to their true skill levels. This yields much more stable and reliable model rankings. It is noteworthy that this faster convergence is observed to be more reliable for higher winning probabilities, which corresponds to skewed win rates in a real-wold scenario.

4.3 Transitive Properties of Elo Scores

Problem Statement. A desirable property of any rating system is transitivity. The Elo rating system is often assumed to possess transitive properties – here we evaluate if that is actually the case. Transitivity in this context means that if player A beats player B, and player B beats player C, then player A is expected to beat player C. Prior work has already demonstrated limitations of Elo in maintaining transitivity, especially in non-transitive cyclic games such as rock-paper-scissors and StarCraft II (Bertrand et al., 2023; Vadori and Savani, 2023). While Elo's design inherently assumes transitivity, our synthetic data, which are derived from realistic scenarios, uncovers certain circumstances that violate this assumption. Such anomalies can subsequently affect the final ranking of language models and their relative performance assessments.

Experimental Setup. The transitivity property of the Elo scores is defined as:

$$A > B \quad \text{and} \quad B > C \implies A > C \qquad (7)$$

Table 1: Investigation of Elo score reliability in capturing true model hierarchies across varying configurations. Scenarios explore the transitive relationship $A > B$ and $B > C \implies A > C$. The star (*) indicates cases where the Elo score fails to accurately reflect the expected hierarchy of models. Symbols: $\approx$ represents models with similar performance; $\gg$ indicates that a model significantly outperforms the other one.

Scenario	Model	Models Ranking per Configuration			
		$N = 1, K = 1$	$N = 100, K = 1$	$N = 1, K = 16$	$N = 100, K = 16$
♔	A	1539.43	1528.50 ± 0.35	1650.93	1584.78 ± 3.09
$A \gg B$	B	1390.47	1410.33 ± 0.54	1381.17	1406.48 ± 3.23
$B \gg C$	C	1270.10	1261.17 ± 0.33	1167.90	1208.74 ± 2.71
♖	A	1502.09	1495.92 ± 0.36	1509.08	1526.04 ± 3.03
$A \gg B$	B	1337.48	**1342.70*** ± 0.53	1379.00	1340.83 ± 2.83
$B \approx C$	C	1360.42	**1361.38*** ± 0.38	1311.92	1333.13 ± 2.68
♗	A	1437.97	**1433.84*** ± 0.41	1440.31	1460.22 ± 2.90
$A \approx B$	B	1455.10	**1453.84*** ± 0.61	1481.04	1452.87 ± 3.25
$B \gg C$	C	1306.93	1312.32 ± 0.34	1278.65	1286.91 ± 2.72
♘	A	1426.33	1419.73 ± 0.36	1407.44	1432.26 ± 2.93
$A \approx B$	B	1390.47	1393.29 ± 0.59	1386.17	1392.75 ± 3.04
$B \approx C$	C	1383.20	1386.99 ± 0.41	1406.39	1374.99 ± 3.12

To test the transitivity property, we design four distinct scenarios:

♔ A beats B and B beats C both with high win probabilities ($P_{\text{win}} = 0.75$).

♖ A beats B with a high win probability ($P_{\text{win}} = 0.75$), B beats C with a win probability close to 0.5 ($P_{\text{win}} = 0.51$).

♗ A beats B with a win probability close to 0.5 ($P_{\text{win}} = 0.51$), B beats C with a high win probability ($P_{\text{win}} = 0.75$).

♘ A beats B with a win probability of 0.54, B beats C with a win probability of 0.51.

In each of these scenarios, we simulate matches for paired comparisons A vs. B and B vs. C and then rearrange these matches in an arbitrary order to form our baseline sequence. This approach mimics how Elo ratings are computed for online leaderboards in the evaluation of large language models (Wu et al., 2023; Lin and Chen, 2023). We then analyze whether Elo scores maintain the expected model hierarchies.

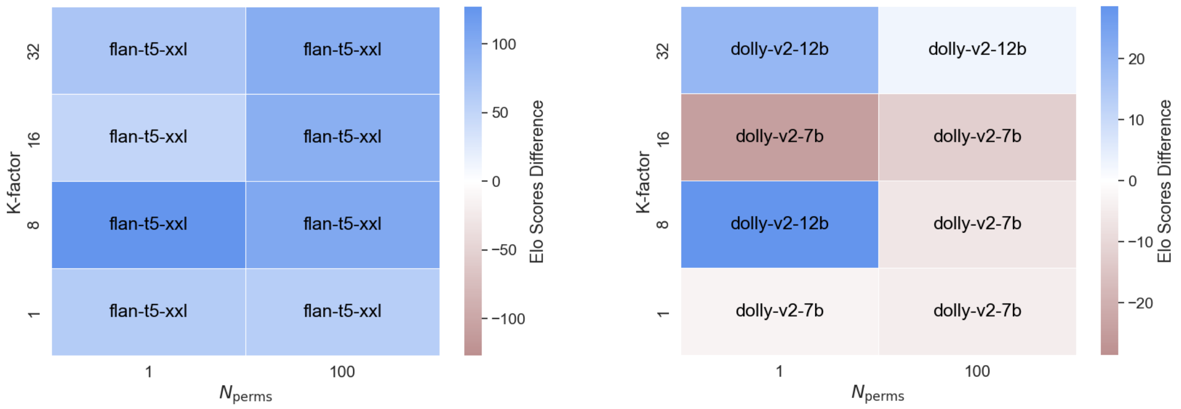
Key Findings. The results of all 4 scenarios are consolidated in table 1. These outcomes validate that the transitivity assumed by the Elo rating system can be vulnerable, especially when win rates hover around $\approx 50\%$. Once again, we observe that varying the number of permutations ($n = 1$ vs $N_{\text{perms}} = 100$) and the K-factor plays a critical role in stability. For $N_{\text{perms}} = 100$ and $K = 1$, we notice discrepancies in the models' rankings. This can be contrasted with $K = 16$, where rankings were much more consistent and reliable. The slower updates from $K = 1$ suggest that this setting is possibly too conservative to capture the transitive relations quickly, hence leading to inconsistencies.

5 Validation on Real-World Human Feedback

Building on the insights gained from our synthetic data experiments, we extend our validation efforts to include real-world human feedback. Our objective is two-fold: first, to ascertain how the demonstrated properties established using synthetic data generalize to real human annotations; and second, to evaluate the Elo rating system's utility for assessing large language models (LLMs) under practical conditions.

Table 2: Win rates per evaluated model across conducted paired comparison experiments.

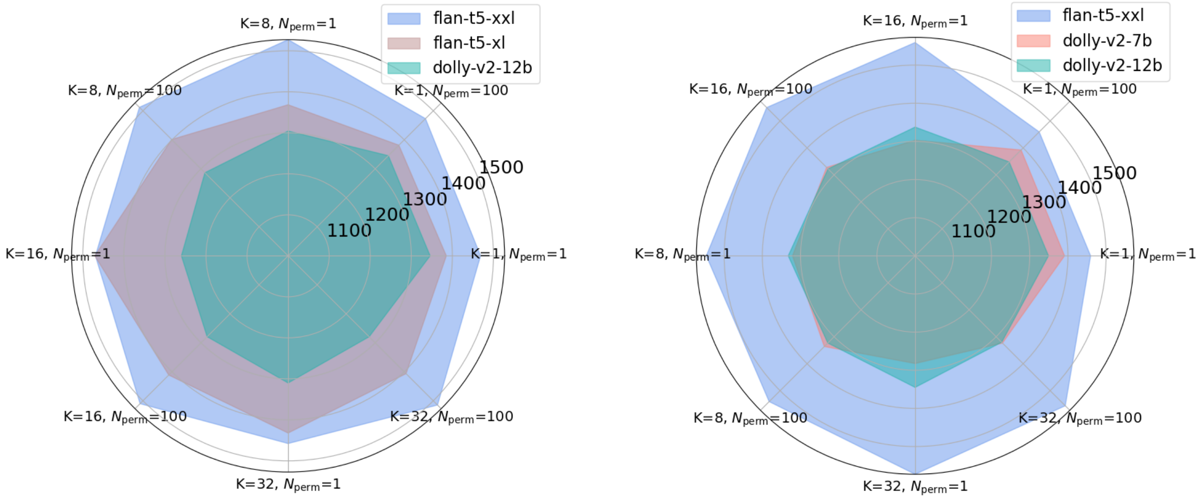
Experiment	Win Rate
Flan-t5-xxl	0.79
Dolly-v2-12b	0.21
Flan-t5-xxl	0.64
Flan-t5-xl	0.36
Dolly-v2-7b	0.51
Dolly-v2-12b	0.49



(a) Experiment: Flan-t5-xxl vs. Flan-t5-xl
Recorded Win rates: 0.64 vs 0.36

(b) Experiment: Dolly-v2-7b vs. Dolly-v2-12b
Recorded Win rates: 0.51 vs 0.49

Figure 4: Final Elo scores difference ($S_A - S_B$) as a function of K-factor and N_{perms}. In this comparison, Model A corresponds to Flan-t5-xxl and Model B corresponds to Flan-t5-xl. Positive values reflect the expected ranking where Model A is superior to Model B, while negative values indicate a discrepancy, falsely suggesting that Model B has a higher Elo score than Model A.



(a) Flan-t5-xxl vs. Flan-t5-xl and Flan-t5-xxl vs. Dolly-v2-12b
Recorded Win rates: 0.64 vs 0.36 and 0.79 vs 0.21

(b) Dolly-v2-7b vs. Dolly-v2-12b and Flan-t5-xxl vs. Dolly-v2-12b
Recorded Win rates: 0.51 vs 0.49 and 0.79 vs 0.21

Figure 5: Final Elo scores (S_A, S_B and S_C) for three different models at multiple configurations of $N_{perms} = \{1, 100\}$ and K-factor = $\{1, 8, 16, 32\}$. When the surfaces representing individual model scores intersect, it signifies that the relative ranking of the models is sensitive to these configurations. The order of models overlaps represent these models ranking based on their Elo scores.

Experimental Setup. Our study leverages human feedback data previously collected to explore data prioritization in language model evaluations. For details about our pool of prompts and models, completion generation, and annotation collection process, we refer the reader to the experimental setup section of our previous work (Boubdir et al., 2023). We focus on models from the well-established Dolly (Conover et al., 2023) and Flan (Chung et al., 2022) families, ensuring relevance to the broader NLP community. The evaluation dataset consists of 400 prompts, with 100 randomly chosen from the SODA (Kim et al., 2022) dataset and 100 from each of the Common-

SENSEQA (Talmor et al., 2019), COMMONGEN (Lin et al., 2020), and ADVERSARIALQA (Bartolo et al., 2020) subsets, all of which are part of the Public Pool of Prompts (P3) dataset (Sanh et al., 2021). This ensures a diverse set of evaluation scenarios for a comprehensive assessment of the models' capabilities. Consistent with our synthetic data methodology, tie outcomes have been excluded from this analysis to focus specifically on the implications for the robustness of Elo scores.

In line with our previous analyses, we continue to explore the influence of variations in $N_{\text{perms}} = \{1, 100\}$ and the K-factor (ranging from 1 to 36) on the robustness and reliability of Elo scores. The win rates for each model, derived from human evaluations, are summarized in Table 2. Our real-world experiments yield two distinct types of scenarios: i) one in which a model decisively outperforms the other, such as the Flan-t5-xxl vs. Flan-t5-xl pairing; and ii) another one with two models nearly evenly matched, as in the Dolly-v2-7b vs. Dolly-v2-12b case.

Key Findings. Our analysis of real-world human feedback data reveals that the stability of Elo ratings is influenced by the disparities in win rates and the choice of hyperparameters K-factor and N_{perms}. In situations where win rates show a significant discrepancy, such as in our Flan family experiment, Elo ratings remain notably consistent across different K-factors and N_{perms} configurations (see Figure 7). On the other hand, in cases like the Dolly family experiment where win rates are closely matched, the Elo rating system exhibits higher volatility at $N_{\text{perms}} = 1$ but gains stability at $N_{\text{perms}} = 100$ at relatively small K-factors (see Figure 4b).

Regarding the conservation of transitivity, our findings indicate that this property is not universally maintained in real-world human evaluations, as observed in synthetic data in section 4. The relative rankings of models that perform similarly are sensitive to the choice of hyperparameters K-factor and N_{perms}. Consequently, one should exercise caution in drawing conclusions from the Elo scores when comprehensive paired comparison data, as dictated by the combination formula 5, is not available. Our observations are in line with the trends seen in our synthetic data experiments.

6 Empirical Guidelines for Robust Elo-based Evaluation

We consolidate the following best practices for a reliable and robust Elo-based evaluation of language models:

- **Stability of Scores**: Running multiple permutations and averaging the Elo scores, preferably with $N_{\text{perm}} \geq 100$, generally yields stable and reliable outcomes.
- **Fine-Tuning the K-factor**: A smaller K-factor may reduce significant rating fluctuations when models have closely matched win rates.
- **Rapid Convergence for Clear Winners**: A larger K-factor can expedite the convergence of Elo ratings to the "true" performance levels when there is a distinct performance gap between models.
- **Transitivity is not guaranteed**: (A beats B and B beats C implies $A > C$) does not always hold in Elo scores, particularly when some of the pairwise comparisons yield closely matched win rates.

These guidelines serve as empirically-grounded recommendations to improve the robustness and interpretability of Elo-based evaluations for LLMs. Following these best practices will help in yielding more reliable conclusions on models' performance via human judgment.

7 Conclusion

This paper provides a comprehensive study on the reliability of the Elo rating system for evaluating LLMs using human feedback. We identify various factors that influence the robustness of Elo ratings and offer guidelines for their effective application in real-world scenarios. While our findings lay down an essential framework, they are by no means exhaustive. Future work could extend the present study by considering tie outcomes and adopting multi-category Bernoulli synthetic data to more closely simulate the varied landscape of human feedback. Such extensions could provide additional insights into the convergence properties of the Elo rating system in the fast-evolving landscape of language models.

References

Amanda Askell, Yuntao Bai, Anna Chen, Dawn Drain, Deep Ganguli, Tom Henighan, Andy Jones, Nicholas Joseph, Ben Mann, Nova DasSarma, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Jackson Kernion, Kamal Ndousse, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, and Jared Kaplan. 2021. A general language assistant as a laboratory for alignment.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, Ben Mann, and Jared Kaplan. 2022a. Training a helpful and harmless assistant with reinforcement learning from human feedback.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, Jamie Kerr, Jared Mueller, Jeffrey Ladish, Joshua Landau, Kamal Ndousse, Kamile Lukosuite, Liane Lovitt, Michael Sellitto, Nelson Elhage, Nicholas Schiefer, Noemi Mercado, Nova DasSarma, Robert Lasenby, Robin Larson, Sam Ringer, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Tamera Lanham, Timothy Telleen-Lawton, Tom Conerly, Tom Henighan, Tristan Hume, Samuel R. Bowman, Zac Hatfield-Dodds, Ben Mann, Dario Amodei, Nicholas Joseph, Sam McCandlish, Tom Brown, and Jared Kaplan. 2022b. Constitutional ai: Harmlessness from ai feedback.

Max Bartolo, A Roberts, Johannes Welbl, Sebastian Riedel, and Pontus Stenetorp. 2020. Beat the ai: Investigating adversarial human annotation for reading comprehension. *Transactions of the Association for Computational Linguistics*, 8:662–678.

Jakob Bernoulli. 1713. *Ars conjectandi, opus posthumum. Accedit Tractatus de seriebus infinitis, et epistola gallicé scripta de ludo pilae reticularis*. Thurneysen Brothers, Basel.

Quentin Bertrand, Wojciech Marian Czarnecki, and Gauthier Gidel. 2023. On the limitations of the elo, real-world games are transitive, not additive. In *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pages 2905–2921. PMLR.

John J. Binder and Murray Findlay. 2009. The effects of the bosman ruling on national and club teams in europe. *Journal of Sports Economics*, 13:107–129.

Meriem Boubdir, Edward Kim, Beyza Ermis, Marzieh Fadaee, and Sara Hooker. 2023. Which prompts make the difference? data prioritization for efficient human llm evaluation.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, Wei Ye, Yue Zhang, Yi Chang, Philip S. Yu, Qiang Yang, and Xing Xie. 2023. A survey on evaluation of large language models.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models.

Mike Conover, Matt Hayes, Ankit Mathur, Jianwei Xie, Jun Wan, Sam Shah, Ali Ghodsi, Patrick Wendell, Matei Zaharia, and Reynold Xin. 2023. Free dolly: Introducing the world's first truly open instruction-tuned llm.

Arman Dehpanah, Muheeb Faizan Ghori, Jonathan F. Gemmell, and Bamshad Mobasher. 2021. Evaluating team skill aggregation in online competitive games. *2021 IEEE Conference on Games (CoG)*, pages 01–08.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. *arXiv preprint arXiv:2305.14314*.

Ning Ding, Yulin Chen, Bokai Xu, Yujia Qin, Zhi Zheng, Shengding Hu, Zhiyuan Liu, Maosong Sun, and Bowen Zhou. 2023. Enhancing chat language models by scaling high-quality instructional conversations.

Aram Ebtekar and Paul Liu. 2021. Elo-mmr: A rating system for massive multiplayer competitions. In *Proceedings of the Web Conference 2021*, WWW '21, page 1772–1784, New York, NY, USA. Association for Computing Machinery.

Arpad E. Elo. 1978. *The Rating of Chessplayers, Past and Present*. Arco Pub., New York.

ESL. Ranking - dota2 - esl pro tour.

Mark E Glickman. 1995. A comprehensive guide to chess ratings. *American Chess Journal*, pages 59–102.

Mark E Glickman. 1999. Parameter estimation in large dynamic paired comparison experiments. *Applied Statistics*, pages 377–394.

Mark E Glickman. 2012. Example of the glicko-2 system. *Boston University*, pages 1–6.

Ralf Herbrich, Tom Minka, and Thore Graepel. 2006. Trueskill™: A bayesian skill rating system. In *Advances in Neural Information Processing Systems*, volume 19. MIT Press.

Lars Magnus Hvattum and Halvard Arntzen. 2010. Using elo ratings for match result prediction in association football. *International Journal of Forecasting*, 26(3):460–470.

Jean Kaddour, Joshua Harris, Maximilian Mozes, Herbie Bradley, Roberta Raileanu, and Robert McHardy. 2023. Challenges and applications of large language models.

Hyunwoo Kim, Jack Hessel, Liwei Jiang, Peter West, Ximing Lu, Youngjae Yu, Pei Zhou, Ronan Le Bras, Malihe Alikhani, Gunhee Kim, Maarten Sap, and Yejin Choi. 2022. Soda: Million-scale dialogue distillation with social commonsense contextualization. *ArXiv*, abs/2212.10465.

Christoph Leitner, Achim Zeileis, and Kurt Hornik. 2010. Forecasting sports tournaments by ratings of (prob)abilities: A comparison for the euro 2008. *International Journal of Forecasting*, 26(3):471–481. Sports Forecasting.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Cosgrove, Christopher D. Manning, Christopher Ré, Diana Acosta-Navas, Drew A. Hudson, Eric Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue Wang, Keshav Santhanam, Laurel Orr, Lucia Zheng, Mert Yuksekgonul, Mirac Suzgun, Nathan Kim, Neel Guha, Niladri Chatterji, Omar Khattab, Peter Henderson, Qian Huang, Ryan Chi, Sang Michael Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas Icard, Tianyi Zhang, Vishrav Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Koreeda. 2022. Holistic evaluation of language models.

Bill Yuchen Lin, Wangchunshu Zhou, Ming Shen, Pei Zhou, Chandra Bhagavatula, Yejin Choi, and Xiang Ren. 2020. CommonGen: A constrained text generation challenge for generative commonsense reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1823–1840, Online. Association for Computational Linguistics.

Yen-Ting Lin and Yun-Nung Chen. 2023. LLM-Eval: Unified multi-dimensional automatic evaluation for open-domain conversations with large language models.

Liquipedia. Elo rating - liquipedia starcraft brood war wiki.

Tom Minka, Ryan Cleven, and Yordan Zaykov. 2018. Trueskill 2: An improved bayesian skill rating system. Technical Report MSR-TR-2018-8, Microsoft.

Luiza Pozzobon, Beyza Ermis, Patrick Lewis, and Sara Hooker. 2023. On the challenges of using black-box apis for toxicity evaluation in research.

April M. Reid. Elo rating system for video games explained.

Victor Sanh, Albert Webson, Colin Raffel, Stephen H. Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, Manan Dey, M Saiful Bari, Canwen Xu, Urmish Thakker, Shanya Sharma Sharma, Eliza Szczechla, Taewoon Kim, Gunjan Chhablani, Nihal Nayak, Debajyoti Datta, Jonathan Chang, Mike Tian-Jian Jiang, Han Wang, Matteo Manica, Sheng Shen, Zheng Xin Yong, Harshit Pandey, Rachel Bawden, Thomas Wang, Trishala Neeraj, Jos Rozen, Abheesht Sharma, Andrea Santilli, Thibault Fevry, Jason Alan Fries, Ryan Teehan, Stella Biderman, Leo Gao, Tali Bers, Thomas Wolf, and Alexander M. Rush. 2021. Multitask prompted training enables zero-shot task generalization.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, Agnieszka Kluska, Aitor Lewkowycz, Akshat Agarwal, Alethea Power, Alex Ray, Alex Warstadt, Alexander W. Kocurek, Ali Safaya, Ali Tazarv, Alice Xiang, Alicia Parrish, Allen Nie, Aman Hussain, Amanda Askell, Amanda Dsouza, Ambrose Slone, Ameet Rahane, Anantharaman S. Iyer, Anders Johan Andreassen, Andrea Madotto, Andrea Santilli, Andreas Stuhlmüller, Andrew M. Dai, Andrew La, Andrew Lampinen, Andy Zou, Angela Jiang, Angelica Chen, Anh Vuong, Animesh Gupta, Anna Gottardi, Antonio Norelli, Anu Venkatesh, Arash Gholamidavoodi, Arfa Tabassum, Arul Menezes, Arun Kirubarajan, Asher Mullokandov, Ashish Sabharwal, Austin Herrick, Avia Efrat, Aykut Erdem, Ayla Karakaş, B. Ryan Roberts, Bao Sheng Loe, Barret Zoph, Bartłomiej Bojanowski, Batuhan Özyurt, Behnam Hedayatnia, Behnam Neyshabur, Benjamin Inden, Benno Stein, Berk Ekmekci, Bill Yuchen Lin, Blake Howald, Bryan Orinion, Cameron Diao, Cameron Dour, Catherine Stinson, Cedrick Argueta, Cesar Ferri, Chandan Singh, Charles Rathkopf, Chenlin Meng, Chitta Baral, Chiyu Wu, Chris Callison-Burch, Christopher Waites, Christian Voigt, Christopher D Manning, Christopher Potts, Cindy Ramirez, Clara E. Rivera, Clemencia Siro, Colin Raffel, Courtney Ashcraft, Cristina Garbacea, Damien Sileo, Dan Garrette, Dan Hendrycks, Dan Kilman, Dan Roth, C. Daniel Freeman, Daniel Khashabi, Daniel Levy, Daniel Moseguí González, Danielle Perszyk, Danny Hernandez, Danqi Chen, Daphne Ippolito, Dar Gilboa, David Dohan, David Drakard, David Jurgens, Debajyoti Datta, Deep Ganguli, Denis Emelin, Denis Kleyko, Deniz Yuret, Derek Chen, Derek Tam, Dieuwke Hupkes, Diganta Misra, Dilyar Buzan, Dimitri Coelho Mollo, Diyi Yang, Dong-Ho Lee, Dylan Schrader, Ekaterina Shutova, Ekin Dogus Cubuk, Elad Segal, Eleanor Hagerman, Elizabeth

Barnes, Elizabeth Donoway, Ellie Pavlick, Emanuele Rodolà, Emma Lam, Eric Chu, Eric Tang, Erkut Erdem, Ernie Chang, Ethan A Chi, Ethan Dyer, Ethan Jerzak, Ethan Kim, Eunice Engefu Manyasi, Evgenii Zheltonozhskii, Fanyue Xia, Fatemeh Siar, Fernando Martínez-Plumed, Francesca Happé, Francois Chollet, Frieda Rong, Gaurav Mishra, Genta Indra Winata, Gerard de Melo, Germán Kruszewski, Giambattista Parascandolo, Giorgio Mariani, Gloria Xinyue Wang, Gonzalo Jaimovitch-Lopez, Gregor Betz, Guy Gur-Ari, Hana Galijasevic, Hannah Kim, Hannah Rashkin, Hannaneh Hajishirzi, Harsh Mehta, Hayden Bogar, Henry Francis Anthony Shevlin, Hinrich Schuetze, Hiromu Yakura, Hongming Zhang, Hugh Mee Wong, Ian Ng, Isaac Noble, Jaap Jumelet, Jack Geissinger, Jackson Kernion, Jacob Hilton, Jaehoon Lee, Jaime Fernández Fisac, James B Simon, James Koppel, James Zheng, James Zou, Jan Kocon, Jana Thompson, Janelle Wingfield, Jared Kaplan, Jarema Radom, Jascha Sohl-Dickstein, Jason Phang, Jason Wei, Jason Yosinski, Jekaterina Novikova, Jelle Bosscher, Jennifer Marsh, Jeremy Kim, Jeroen Taal, Jesse Engel, Jesujoba Alabi, Jiacheng Xu, Jiaming Song, Jillian Tang, Joan Waweru, John Burden, John Miller, John U. Balis, Jonathan Batchelder, Jonathan Berant, Jörg Frohberg, Jos Rozen, Jose Hernandez-Orallo, Joseph Boudeman, Joseph Guerr, Joseph Jones, Joshua B. Tenenbaum, Joshua S. Rule, Joyce Chua, Kamil Kanclerz, Karen Livescu, Karl Krauth, Karthik Gopalakrishnan, Katerina Ignatyeva, Katja Markert, Kaustubh Dhole, Kevin Gimpel, Kevin Omondi, Kory Wallace Mathewson, Kristen Chiafullo, Ksenia Shkaruta, Kumar Shridhar, Kyle McDonell, Kyle Richardson, Laria Reynolds, Leo Gao, Li Zhang, Liam Dugan, Lianhui Qin, Lidia Contreras-Ochando, Louis-Philippe Morency, Luca Moschella, Lucas Lam, Lucy Noble, Ludwig Schmidt, Luheng He, Luis Oliveros-Colón, Luke Metz, Lütfi Kerem Senel, Maarten Bosma, Maarten Sap, Maartje Ter Hoeve, Maheen Farooqi, Manaal Faruqui, Mantas Mazeika, Marco Baturan, Marco Marelli, Marco Maru, Maria Jose Ramirez-Quintana, Marie Tolkiehn, Mario Giulianelli, Martha Lewis, Martin Potthast, Matthew L Leavitt, Matthias Hagen, Mátyás Schubert, Medina Orduna Baitemirova, Melody Arnaud, Melvin McElrath, Michael Andrew Yee, Michael Cohen, Michael Gu, Michael Ivanitskiy, Michael Starritt, Michael Strube, Michał Swędrowski, Michele Bevilacqua, Michihiro Yasunaga, Mihir Kale, Mike Cain, Mimee Xu, Mirac Suzgun, Mitch Walker, Mo Tiwari, Mohit Bansal, Moin Aminnaseri, Mor Geva, Mozhdeh Gheini, Mukund Varma T, Nanyun Peng, Nathan Andrew Chi, Nayeon Lee, Neta Gur-Ari Krakover, Nicholas Cameron, Nicholas Roberts, Nick Doiron, Nicole Martinez, Nikita Nangia, Niklas Deckers, Niklas Muennighoff, Nitish Shirish Keskar, Niveditha S. Iyer, Noah Constant, Noah Fiedel, Nuan Wen, Oliver Zhang, Omar Agha, Omar Elbaghdadi, Omer Levy, Owain Evans, Pablo Antonio Moreno Casares, Parth Doshi, Pascale Fung, Paul Pu Liang, Paul Vicol, Pegah Alipoormolabashi, Peiyuan Liao, Percy Liang, Peter W Chang, Peter Eckersley, Phu Mon Htut, Pinyu Hwang, Piotr Miłkowski, Piyush Patil, Pouya Pezeshkpour, Priti Oli, Qiaozhu Mei, Qing Lyu, Qinlang Chen, Rabin Banjade, Rachel Etta Rudolph, Raefer Gabriel, Rahel Habacker, Ramon Risco, Raphaël Millière, Rhythm Garg, Richard Barnes, Rif A. Saurous, Riku Arakawa, Robbe Raymaekers, Robert Frank, Rohan Sikand, Roman Novak, Roman Sitelew, Ronan Le Bras, Rosanne Liu, Rowan Jacobs, Rui Zhang, Russ Salakhutdinov, Ryan Andrew Chi, Seungjae Ryan Lee, Ryan Stovall, Ryan Teehan, Rylan Yang, Sahib Singh, Saif M. Mohammad, Sajant Anand, Sam Dillavou, Sam Shleifer, Sam Wiseman, Samuel Gruetter, Samuel R. Bowman, Samuel Stern Schoenholz, Sanghyun Han, Sanjeev Kwatra, Sarah A. Rous, Sarik Ghazarian, Sayan Ghosh, Sean Casey, Sebastian Bischoff, Sebastian Gehrmann, Sebastian Schuster, Sepideh Sadeghi, Shadi Hamdan, Sharon Zhou, Shashank Srivastava, Sherry Shi, Shikhar Singh, Shima Asaadi, Shixiang Shane Gu, Shubh Pachchigar, Shubham Toshniwal, Shyam Upadhyay, Shyamolima Shammie Debnath, Siamak Shakeri, Simon Thormeyer, Simone Melzi, Siva Reddy, Sneha Priscilla Makini, Soo-Hwan Lee, Spencer Torene, Sriharsha Hatwar, Stanislas Dehaene, Stefan Divic, Stefano Ermon, Stella Biderman, Stephanie Lin, Stephen Prasad, Steven Piantadosi, Stuart Shieber, Summer Misherghi, Svetlana Kiritchenko, Swaroop Mishra, Tal Linzen, Tal Schuster, Tao Li, Tao Yu, Tariq Ali, Tatsunori Hashimoto, Te-Lin Wu, Théo Desbordes, Theodore Rothschild, Thomas Phan, Tianle Wang, Tiberius Nkinyili, Timo Schick, Timofei Kornev, Titus Tunduny, Tobias Gerstenberg, Trenton Chang, Trishala Neeraj, Tushar Khot, Tyler Shultz, Uri Shaham, Vedant Misra, Vera Demberg, Victoria Nyamai, Vikas Raunak, Vinay Venkatesh Ramasesh, vinay uday prabhu, Vishakh Padmakumar, Vivek Srikumar, William Fedus, William Saunders, William Zhang, Wout Vossen, Xiang Ren, Xiaoyu Tong, Xinran Zhao, Xinyi Wu, Xudong Shen, Yadollah Yaghoobzadeh, Yair Lakretz, Yangqiu Song, Yasaman Bahri, Yejin Choi, Yichi Yang, Yiding Hao, Yifu Chen, Yonatan Belinkov, Yu Hou, Yufang Hou, Yuntao Bai, Zachary Seid, Zhuoye Zhao, Zijian Wang, Zijie J. Wang, Zirui Wang, and Ziyi Wu. 2023. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. CommonsenseQA: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4149–4158, Minneapolis, Minnesota. Association for Computational Linguistics.

Nelson Vadori and Rahul Savani. 2023. Ordinal potential-based player rating.

Ben P. Wise. 2021. Elo ratings for large tournaments of software agents in asymmetric games. *ArXiv*, abs/2105.00839.

Yuxiang Wu, Zhengyao Jiang, Akbir Khan, Yao Fu, Laura Ruis, Edward Grefenstette, and Tim Rocktäschel. 2023. Chatarena: Multi-agent language game environments for large language models. https://github.com/chatarena/chatarena.

Yao Zhao, Rishabh Joshi, Tianqi Liu, Misha Khalman, Mohammad Saleh, and Peter J. Liu. 2023. Slic-hf: Sequence likelihood calibration with human feedback.

A Extension to Multiple Outcomes

For scenarios where outcomes can extend beyond wins and losses, such as a tie possibility, the multinomial distribution becomes relevant. For outcomes win, loss, and tie, the distribution is given by:

$$\begin{aligned} & P\left(n_{\text{win}}, n_{\text{loss}}, n_{\text{tie}}; N, p_{\text{win}}, p_{\text{loss}}, p_{\text{tie}}\right) \\ & = \frac{N!}{n_{\text{win}}! n_{\text{loss}}! n_{\text{tie}}!} p_{\text{win}}^{n_{\text{win}}} p_{\text{loss}}^{n_{\text{loss}}} p_{\text{tie}}^{n_{\text{tie}}} \end{aligned} \tag{8}$$

Sampling from the appropriate distribution is fundamental to simulating synthetic human feedback: the binomial distribution for binary feedback and the multinomial for multi-category feedback.

B Impact of Ordering on Elo Ratings: Skewed Win Rates

We summarize our findings on the impact of match sequences on Elo ratings for winning probabilities $Prob(A \text{ beats } B) \geq 0.65$.

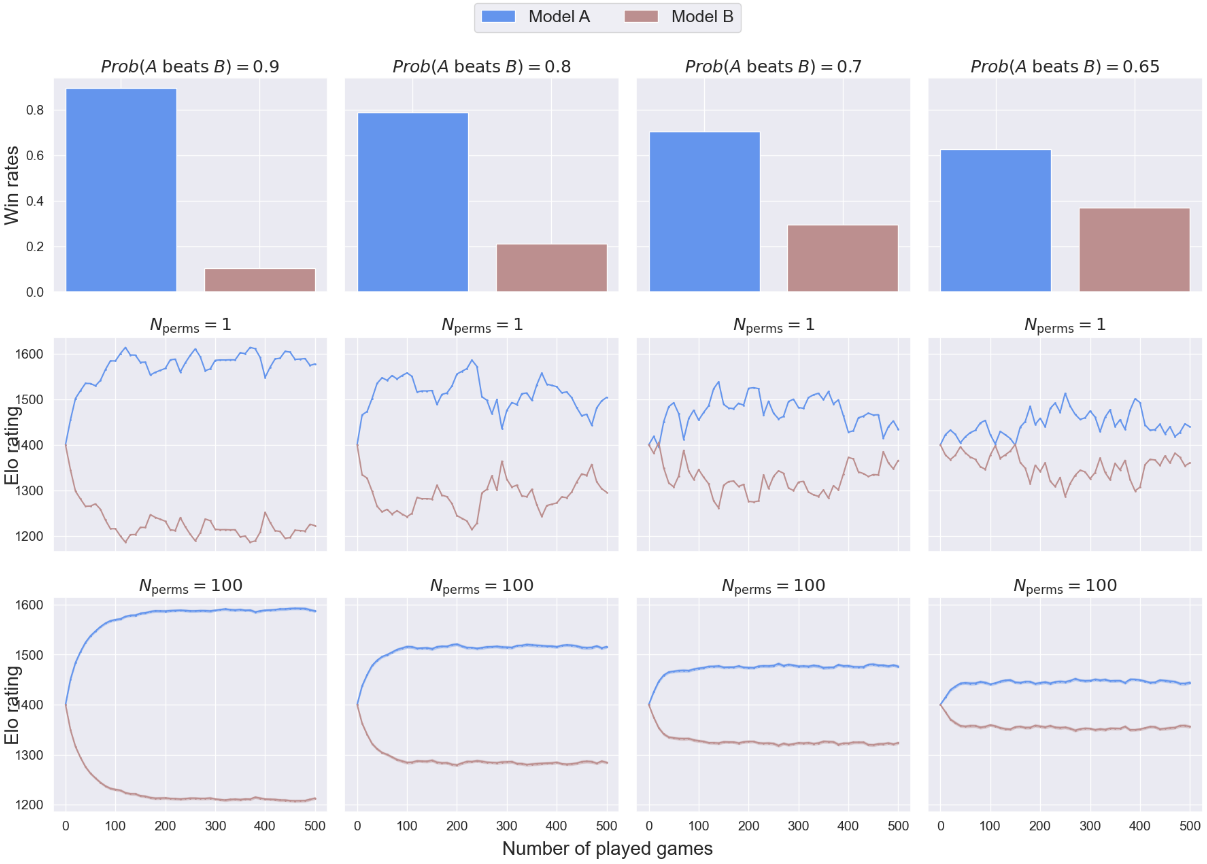


Figure 6: **Impact of win probabilities and permutation sampling on Elo ratings**: Comparing Model A and Model B across three different win probabilities ($Prob(A \text{ beats } B) = 0.9, 0.8, 0.7, 0.65$) with two levels of permutation sampling ($N_{\text{perms}} = 1$ and $N_{\text{perms}} = 100$). The top row displays the observed win rates, the middle row illustrates Elo ratings with a single permutation, and the bottom row shows the mean and standard error of the mean (SEM) of Elo ratings across 100 permutations.

C Experiment Flan-t5-xxl vs. Dolly-v2-12b Results

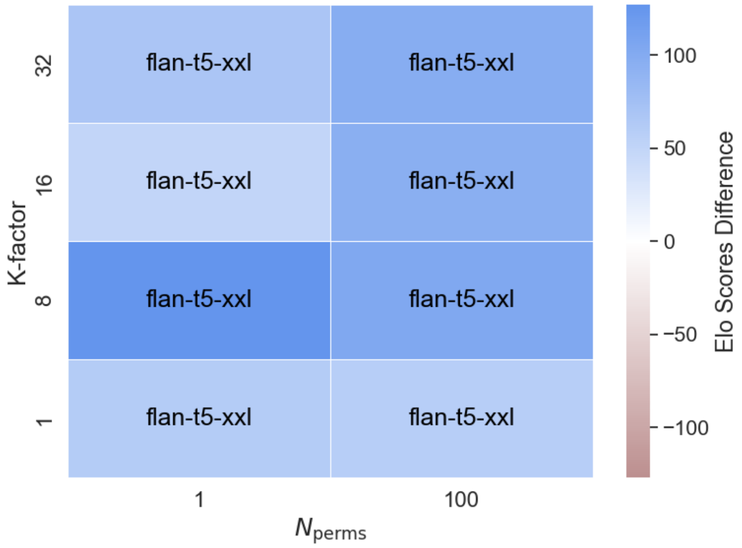


Figure 7: Experiment: Flan-t5-xxl vs. Dolly-v2-12b
Recorded Win rates: 0.79 vs 0.21