# Elo Uncovered: Robustness and Best Practices in Language Model Evaluation

**Meriem Boubdir**
Cohere For AI
meri.boubdir@gmail.com

**Edward Kim**
Cohere
edward@cohere.com

**Beyza Ermis**
Cohere For AI
beyza@cohere.com

**Sara Hooker**
Cohere For AI
sarahooker@cohere.com

**Marzieh Fadaee**
Cohere For AI
marzieh@cohere.com

## Abstract

In Natural Language Processing (NLP), the Elo rating system, originally designed for ranking players in dynamic games such as chess, is increasingly being used to evaluate 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. We study two fundamental axioms that evaluation methods should adhere to: **reliability** and **transitivity**. We conduct an extensive evaluation of Elo behavior across simulated and real-world scenarios, demonstrating that individual Elo computations can exhibit significant volatility. We show that both axioms are not always satisfied, raising questions about the reliability of current comparative evaluations of LLMs. If the current use of Elo scores is intended to substitute the costly head-to-head comparison of LLMs, it is crucial to ensure the ranking is as robust as possible. Guided by the axioms, our findings offer concrete guidelines for enhancing the reliability of LLM evaluation methods, suggesting a need for reassessment of existing comparative approaches.

## 1 Introduction

In the rapidly evolving field of Natural Language Processing (NLP), the task of accurately and reliably evaluating LLMs has become increasingly challenging [32, 10, 50, 26, 43]. 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 [2, 3, 4, 50, 13, 12]. 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. Pairwise comparisons, however, are fundamentally difficult to scale for large pools of models, due to the quadratic growth of comparisons required. Fortunately, this paired feedback structure [56] naturally lends itself to the Elo rating system, originally designed for ranking chess players (including those who have never before played each other) for better matchmaking [16].

Under the Elo rating system, players' skills are indicated by an *Elo rating*, where higher ratings indicate higher skill, and all players can be ranked best to worst using this scalar Elo rating. In the standard formulation (see Section 2), a player rated at 1800 has $10:1$ odds of winning against a player rated at 1400. After a match, the winner takes rating points from the loser in a zero-sum fashion [16]. Thus, with the Elo rating system, we can efficiently integrate subjective human feedback on paired *"A vs B"* language model completions into a structured and unified rating system to assess the performance of language models.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).


https://doi.org/10.52202/079017-3367

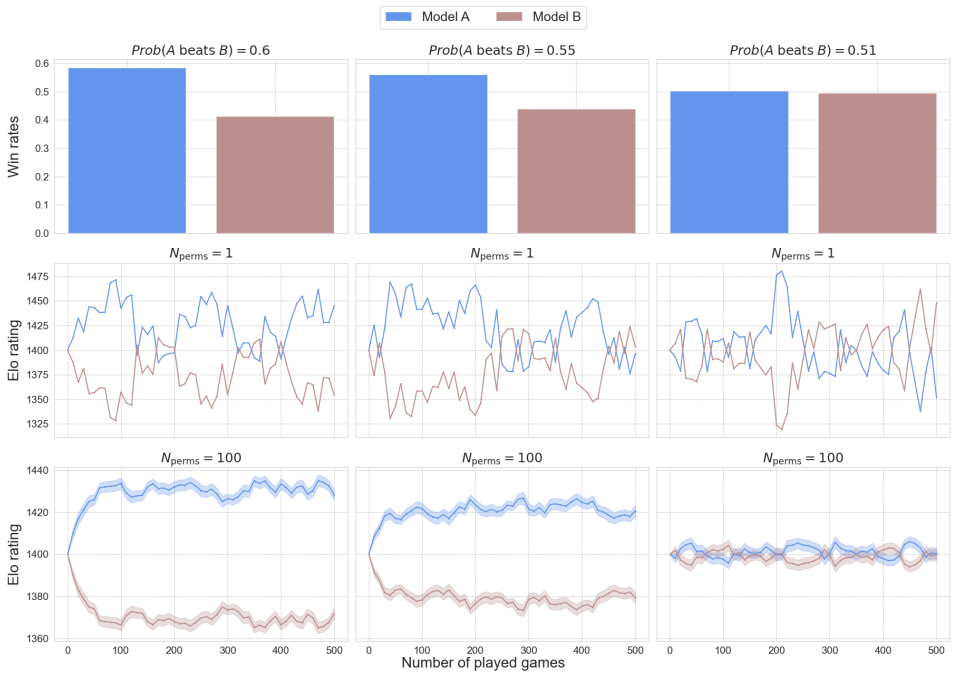



Figure 1: **Impact of win probabilities and permutation sampling on Elo ratings**: Comparing Model A and Model B across three different win probabilities ($Prob(A \text{ beats } B) = \{0.6, 0.55, 0.51\}$) with two levels of permutation sampling ($N_{\text{perms}} = 1$ and $N_{\text{perms}} = 100$). The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

The core principles of Elo rating have proven to be resilient and adaptable due to its dynamic adjustments, relative rating focus, consistency across skill levels, and simplicity and transparency. As a result, the Elo rating system has found diverse applications, from predicting sports events outcomes [8, 25, 31, 53], and facilitating matchmaking in massively multiplayer online games like StarCraft II and Dota [15, 44, 35, 17], to its recent use in the evaluation of LLMs [2, 3, 4, 50, 13, 12, 54, 33]. However, to-date there has not been a comprehensive examination of the compatibility of Elo scores and LLMs evaluation.

Unlike dynamic competitors that evolve over time, LLMs have static capabilities and operate in a time-agnostic context. In this setting, evaluations of LLMs are not constrained by a preset number of turns, as is the case with tournament timelines or predefined match sequences. Moreover, the ordering of matches can significantly influence the final Elo scores and, consequently, model rankings. This oversight is particularly concerning, given the direct impact of Elo system rankings on both research directions and real-world applications in NLP as well as its widespread adoption [2, 3, 55, 57, 29, 4, 50, 13, 12, 54, 33].

This study aims to close this research gap by adopting an axiomatic approach and scrutinizing both the reliability and limitations of the Elo rating system when applied to LLMs. We study two fundamental axioms that evaluation methods should adhere to: **reliability** and **transitivity**. 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 for LLMs are highly sensitive to the *order of comparisons* and the choice of hyperparameters. Moreover, 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 addressing Elo rating sensitivities by offering guidelines for hyperparameter selection and matchmaking scenarios.

**Implications of our work** As LLMs rapidly advance, evaluation leaderboards are gaining popularity to assess the performance of newly introduced models using Elo scores. Elo can also be used in the learning framework of LLMs to produce a ranking of models and their outputs for preference training. No research has explored the nuances of using Elo scores to compare LLMs, which, unlike chess, exhibit static capabilities and operate in a time-agnostic manner. We show that Elo rating does not always satisfy two critical axioms—reliability and transitivity—leading to rankings of models that are not accurate. Our research offers guidelines for reliable and robust implementation of Elo scores when comparing LLMs. Deviation from our recommendations could result in inaccuracies when ranking LLMs, particularly in situations where model performances are closely matched, and Elo score differences are minimal (a common occurrence in many real-world scenarios).

## 2 Elo Algorithm Explained

We provide the standard mathematical formulation of the Elo algorithm [16], 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 Elo rating $R_i$. For each match between two models, we calculate the *expected score*, then update the *ratings* of both models as follows:

### 2.1 Expected Score Computation

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

$$E_A = \frac{1}{1 + 10^{(R_B - R_A)/400}} \quad \text{and} \quad E_B = \frac{1}{1 + 10^{(R_A - R_B)/400}} \tag{1}$$

In this context, the factor of $400$ [16] 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.

### 2.2 Rating Update Mechanism

Following each match, the Elo ratings are updated based on the observed win-loss outcome. The rating adjustment for each model 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 for a loss or 1 for a win. Model B's Elo rating is updated via the same method. The $K$-factor serves as a variable hyperparameter to adapt the rate of change in rating to different scenarios. A higher $K$-factor results in larger changes in the Elo score after each match-up, making the scoring more sensitive to individual results. A lower $K$-factor, in contrast, makes the Elo ratings more stable, with smaller changes after each match. In chess, the $K$-factor is usually set to 16 for masters and to 32 for novice players.

## 3 Desirable Properties of Elo

The objective of using Elo scores to rank models is to establish a comparative understanding of the performance hierarchy among them. When incorporating a new model into an already ranked list, only a limited number of pairwise annotations are required to determine its position in the ranking. The ability to infer a model's relative performance compared to all previous models in the list relies on the robustness of the scoring method and the transitive property of the ranking system. We describe these desirable properties through two axioms: *transitivity* and *reliability*.

 https://doi.org/10.52202/079017-3367

### 3.1 Axiom 1: Transitivity

A desirable property of any rating system is transitivity because it ensures consistency and logical coherence in how entities are ranked or rated. 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 $C$. If the ranking of large language models exhibits transitivity, we can deduce their comparative performance without the need for direct head-to-head evaluations between every pair of models. The central assumption in developing various leaderboards for comparing language models is that the rankings adhere to the principle of transitivity [57].

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 affect the final ranking of language models and their relative performance assessments.

### 3.2 Axiom 2: Reliability

We consider two aspects of reliability:

**Sensitivity to ordering:** 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 model ranking. In this context, each match represents the performance comparison between two models on a specific prompt. If the prompts are presented in a specific order, and one model happens to perform better on the initial set of prompts, it may gain an advantage in subsequent comparisons due to the cumulative effect of its early success. This inherent variability prompts us to investigate the extent to which match-up ordering affects the robustness of Elo ratings.

**Sensitivity to hyperparameters:** The sensitivity of hyperparameters can compromise the robustness of Elo scores leading to inconsistent rankings. Evaluating and understanding this sensitivity is crucial for building evaluation frameworks that maintain consistency across diverse models. In this work, we evaluate the sensitivity of Elo performance to one key hyperparameter, the $K$-factor. This factor acts as a scaling constant in the Elo rating system, pivotal for updating ratings after each matching. It essentially determines how quickly a model's rating converges to what can be considered its "true" skill. While conventional applications like chess use standard $K$-factor values, these may not be directly applicable in the context of evaluating LLMs due to the unique characteristics and requirements of this domain.

## 4 Synthetic Human Feedback

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 first validate the properties of Elo using synthetic data generation via Bernoulli processes to simulate various human feedback scenarios. In Section 6 we extend these evaluations to include real-world human feedback. This time-agnostic and independent setup of LLM evaluations resembles a Bernoulli process [6], a sequence of independent experiments, each yielding a simple "win" or "loss" outcome, representing one model outperforming another. We use this setting to control the characteristics of the distribution and evaluate the different desirable properties of a rating system.

In this controlled setting, our primary objectives include testing the **transitivity** axiom—whether a consistently higher-rated model outperforms those with lower ratings in all scenarios. Additionally, in studying the **reliability** axiom, we explore how the Elo scores are affected by the *order in which models are compared* and the sensitivity to *hyperparameter adjustments*, particularly the $K$-factor. This synthetic setup offers a robust platform to dissect and understand the dynamics of the Elo rating system in the context of LLM evaluations, without the constraints and limitations of relying solely on real-world human feedback.

### 4.1 The Bernoulli Analogy

Pairwise comparisons in LLM evaluation draw parallels with the foundational principles of the Bernoulli experiment in probability theory. This section studies 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". The outcomes adhere to the probability condition:

$$P(\text{success}) + P(\text{failure}) = 1 \tag{3}$$

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

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

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 $n$ distinct models $\mathcal{M}$, their pairwise comparisons can be formulated as:

$$\binom{n}{2} = \frac{n!}{2!(n-2)!} = \frac{n(n-1)}{2} \tag{5}$$

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

### 4.2 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:

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

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.

## 5 How Robust Are Elo Scores?

This section describes rigorous stress tests designed to investigate whether the two axioms, presented in Section 3, are satisfied in this evaluation framework. 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 the $K$-factor, and (3) preserve properties of transitivity. Subsequently, we provide empirically grounded guidelines for a safe and interpretable application of Elo ratings.

### 5.1 Impact of Ordering on Elo Ratings

**Experimental Setup** To quantify the effect of match-up ordering, we generate a baseline sequence of $N_{\text{games}} = 1000$ match outcomes between models $A$ and $B$ (see Equation 6), reflecting the scale typical of LLM evaluations via human feedback. We hold $N_{\text{games}}$ constant for the entirety of our study to maintain consistency. From this baseline, we derive $N_{\text{perms}}$ distinct permutations, each involving a complete reshuffling of the initial sequence to simulate various chronological orders in which the games might unfold. It is important to note that we are not generating new match outcomes for each permutation; instead, we simply reorder the existing data to explore the potential impact of different

 https://doi.org/10.52202/079017-3367

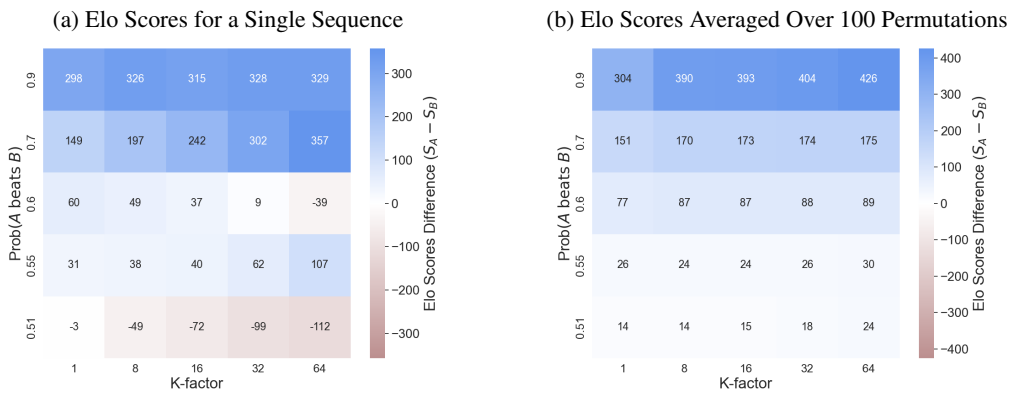



Figure 2: Final Elo scores difference ($S_A - S_B$) as a function of $K$-factor and $N_{\text{perms}}$. Positive values reflect the expected ranking where Model $A$ is superior to Model $B$, while negative values indicate a discrepancy, falsely suggesting that Model $B$ has a higher Elo score than Model $A$. We compare between a single sequence of outcomes and averages over $N_{\text{perms}} = 100$ unique permutations.

match-up sequences. For each reordered sequence, we update the Elo ratings $R_A$ and $R_B$ according to equation 2, resetting both ratings to an initial value of 1400 at the start of each permutation. Finally, we compute average Elo ratings per match across all $N_{\text{perms}}$ permutations, ensuring a robust analysis that takes into account the full range of possible match-up orders.

We repeat this process to generate baseline sequences and their respective reorderings for a set of selected winning probabilities enabling us to inspect ratings' behavior under various real-world scenarios. $N_{\text{perms}}$ is varied from a minimum of 1 to a maximum of 10k, providing a robust sample size for statistical analysis (see Figure 3). 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 $N_{\text{perms}} = 100$, 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 $N_{\text{perms}}$ 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 win probabilities around 0.5, Elo ratings frequently intertwine, making it challenging to discern a clear performance difference between the two. The instability plateaus as $N_{\text{perms}}$ 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 $N_{\text{perms}}$ for a reliable interpretation of Elo ratings in LLM evaluations. In Elo-based evaluations, the sequence of model comparisons can significantly influence the final Elo scores, particularly in scenarios with models of similar quality, where this effect is magnified.

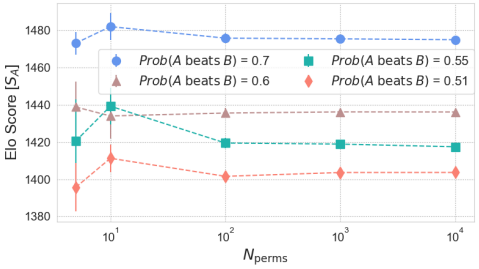



Figure 3: Variation of Model A's average Elo score with increasing number of permutations $N_{\text{perms}}$ for different probabilities of Model A winning ($P(A \text{ beats } B)$). Error bars indicate standard errors of the mean.

### 5.2 Sensitivity to Hyperparameters

**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

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. $\approx$ represents models with similar performance; $\gg$ indicates that a model significantly outperforms the other one.

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

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 2 to assess the stability and expected ranking between the two models.

**Key Findings** As shown in Figure 2, 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-world scenario.

### 5.3 Transitive Properties of Elo Scores

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

$$A > B \quad \text{and} \quad B > C \implies A > C \tag{7}$$

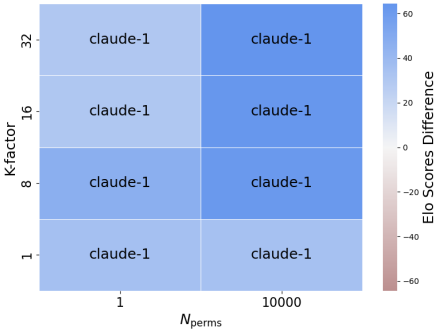
To test the transitivity property, we design four distinct scenarios that model real-world conditions:

- ♔ Model $A$ beats model $B$ and model $B$ beats model $C$ both with high win probabilities ($P_{\text{win}} = 0.75$).
- ♖ Model $A$ beats model $B$ with a high win probability ($P_{\text{win}} = 0.75$), model $B$ beats model $C$ with a win probability close to 0.5 ($P_{\text{win}} = 0.51$).
- ♗ Model $A$ beats model $B$ with a win probability close to 0.5 ($P_{\text{win}} = 0.51$), model $B$ beats model$C$ with a high win probability ($P_{\text{win}} = 0.75$).
- ♘ Model $A$ beats model $B$ with a win probability of 0.54, model $B$ beats model $C$ 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 [54, 33]. We then analyze whether Elo scores maintain the expected model hierarchies.

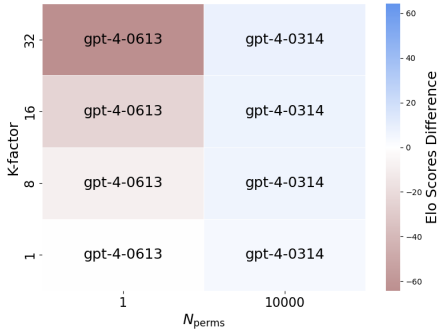
**Key Findings** The outcomes from all four scenarios, detailed in Table 1, demonstrate the performance of Elo-based rankings across various configurations. In scenarios where there is a clear disparity between models (e.g.,♔), Elo ratings accurately reflect the expected hierarchy. However, in


https://doi.org/10.52202/079017-3367




(a) Experiment: Claude-1 vs. Claude-2.1

**Recorded Win rates**: 0.59 vs 0.41




(b) Experiment: GPT-4-0314 vs. GPT-4-0613

**Recorded Win rates**: 0.51 vs 0.49

Figure 4: Elo score differences ($S_A - S_B$) across varying K-factors and $N_{\text{perms}}$. Positive values in the heatmap indicate that the expected ranking is maintained (Model A outperforming Model B), while negative values suggest a ranking inversion, where Model B appears to outperform Model A, contrary to the actual win rates. Each cell's label indicates the model with the higher Elo score.

more complex cases such as ♖ and ♗, where one model significantly outperforms a second, which in turn is closely matched with a third, the rankings become less stable, challenging the assumption of transitivity. We observe once again that varying the number of permutations ($N_{\text{perms}} = 1$ vs. $N_{\text{perms}} = 100$) and the $K$-factor plays a critical role in stability. In the ♖ and ♗ scenarios, with $N_{\text{perms}} = 100$ and $K = 1$, we notice discrepancies in the models' rankings. This contrasts with $K = 16$, where rankings are more consistent and accurate. The slower updates from $K = 1$ suggest this setting may be too conservative to capture transitive relations quickly, leading to inconsistencies.

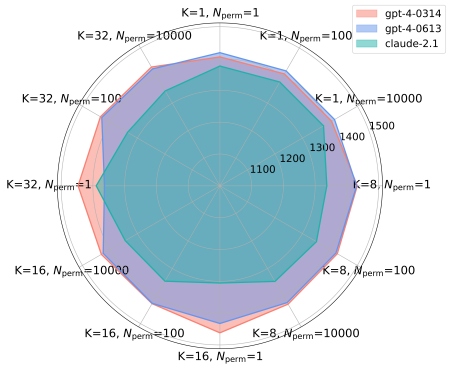
## 6 Validation on Real-World Human Feedback

Building on the insights gained from synthetic data experiments, this section extends the validation of the Elo rating system to real-world human feedback. Our objectives are twofold: first, to ascertain how the properties demonstrated using synthetic data generalize to real human annotations, and second, to evaluate the Elo rating system's utility for assessing LLMs in practical settings.

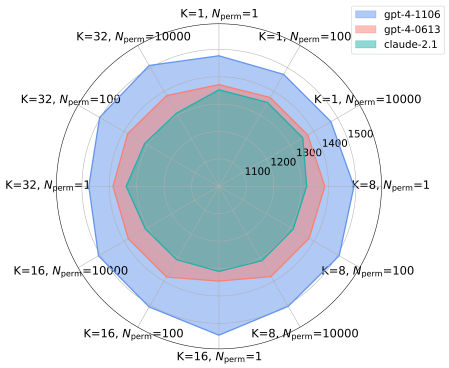
**Experimental Setup** We use the *LMSYS - Chatbot Arena* dataset [34], an open-source collection of human preference data derived from unique users' interactions with two distinct models responding to a set of user-defined prompts. To align with our methodology from synthetic data analysis, tie outcomes have been excluded from this analysis to focus specifically on the implications of win-loss dynamics. We select pairs of models (A vs. B) from the initial dataset that feature at least 300 non-tie comparisons. This threshold ensures statistical robustness and allows us to include cases where win rates are closely contested, which can lead to more sensitive ratings. These pairs predominantly involve models from the GPT-4 family [42] and the Claude family [1]. A comprehensive list of model pairs is included in Appendix C under Table 3, and a subset discussed here is shown in Table 2. The recorded win rates primarily exhibit skewed preferences, with the exception of the GPT-4-0314 vs. GPT-4-0613 pairing, indicating comparable performance levels (see Table 2). Given the variable number of evaluations per pair in the original dataset, we standardize this by sampling a fixed number, $N_{\text{sample}}$, for each pair to align with the controlled conditions used in synthetic analyses. When sampling to $N_{\text{sample}}$, we ensure that the resulting win rates accurately represent the original dataset's findings, providing a faithful evaluation of recorded model performance. This standardization facilitates a more reliable comparison and assessment of the Elo rating

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

| Experiment | Win Rates |
|---|---|
| GPT-4-0314 | 0.51 |
| GPT-4-0613 | 0.49 |
| Claude-1 | 0.59 |
| Claude-2.1 | 0.41 |
| GPT-4-0314 | 0.65 |
| Claude-2.1 | 0.35 |
| GPT-4-0613 | 0.61 |
| Claude-2.1 | 0.39 |
| GPT-4-1106-preview | 0.86 |
| GPT-4-0613 | 0.33 |




(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

Figure 5: Elo scores ($S_A$, $S_B$ and $S_C$) for three models at different configurations of $N_{perms} = \{1, 100, 10000\}$ and $K$-factor = $\{1, 8, 16, 32\}$. The intersections of score lines in 5a indicate fluctuating relative rankings, highlighting inconsistency especially pronounced among models with close performance levels. In contrast, 5b demonstrates more stable relative rankings in conditions where win rates are more skewed.

system under real-world conditions. In line with our previous analyses, we continue to explore the influence of variations in $N_{\text{perms}} = \{1, 100, 10000\}$ and the $K$-factor (ranging from 1 to 36) on Elo score robustness and reliability. We examine scenarios where one model decisively outperforms another (e.g., Claude-1 vs. Claude-2.1) and cases where models are nearly evenly matched (e.g., GPT-4-0314 vs. GPT-4-0613).

**Key Findings** Our analysis of real-world human feedback data confirms that the stability of Elo ratings is influenced by disparities in win rates, analogous to win probabilities in synthetic data, and by the choice of hyperparameters $K$-factor and $N_{\text{perms}}$. In cases where the models show a clear difference in performance as indicated by their win rates, such as in the Claude-1 vs. Claude-2.1 experiment, Elo ratings remain notably consistent across different $K$-factors and $N_{\text{perms}}$ configurations (see Figure 4a). On the other hand, in cases like the GPT-4-0314 vs. GPT-4-0613 experiment where win rates are closely matched, the Elo rating system exhibits higher volatility at $N_{\text{perms}} = 1$ but gains stability with larger $N_{\text{perms}}$ settings (100 and 10000), especially at lower $K$-factors (see Figure 4b). The magnitude of Elo score differences in these experiments illustrates that larger $K$-factor and $N_{\text{perms}}$ values can amplify or reduce the perceived performance gap between models, reflecting the critical role of these parameters in evaluation sensitivity.

Regarding the conservation of transitivity, our findings indicate that this property is not universally maintained across real-world human evaluations and synthetic scenarios (see Section 5). The relative rankings of models with similar performance levels are particularly sensitive to the choice of hyperparameters. Consequently, one should exercise caution in drawing conclusions from the Elo scores, especially in the absence of extensive paired comparison data as required by the combination formula 5. These observations are consistent with the trends from our synthetic data experiments.

## 7 Related Work

Several works have proposed improvements to the Elo rating system. Variants such as Glicko [18, 19, 20] and TrueSkill™ [24, 39] have incorporated more complex statistical methods into the original Elo framework, to address some of the limitations of the Elo rating system, particularly in the context of games with more than two players or teams, or games with more complex outcomes


https://doi.org/10.52202/079017-3367

than just win or loss. There is also ongoing research into the efficacy of these systems in diverse and dynamic environments [11, 7]. Prior work has demonstrated some limitations of Elo in maintaining transitivity, especially in non-transitive cyclic games such as rock-paper-scissors and StarCraft II [7, 52]. However, our work diverges by focusing on the reliability of Elo applied to large language model systems. To date, there has not been a comprehensive evaluation in this context.

Independent from Elo, numerous studies have explored how sensitivity to hyperparameters can undermine the generalization of findings [41, 36, 23, 27, 9] in machine learning. This forms part of a wider body of work that considers which factors influence reliability and reproducibility [21, 22, 5, 14]. Notable directions includes studies on the impact of random seeds [40, 37, 51], model design choices [46, 48, 43, 28, 47], the use of data parallelism [45], hardware [58] and test set construction [49, 30, 38]. Our work is complementary to these efforts, providing a rigorous evaluation of the impact of key hyperparameters and experimental settings on Elo performance.

## 8 Empirical Guidelines for Robust Elo-based Evaluation of LLMs

In this section, we distill essential practices for enhancing the reliability of Elo-based evaluation of language models. These guidelines, derived from our empirical findings, differ notably from some conventional Elo settings and have significant implications for current real-world applications:

- **Achieving Score Stability**: To obtain stable and reliable Elo ratings, it's recommended to run numerous permutations, ideally with $N_{\text{perms}} \geq 100$. This approach significantly improves the consistency of outcomes over single or fewer permutations commonly used.
- **Adjusting the $K$-factor**: A smaller K-factor may reduce significant rating fluctuations when models have closely matched win rates.
- **Rapid Convergence for Clear Winners**: When there is a clear performance disparity between models, a higher K-factor accelerates the alignment of Elo ratings with the models' "true" performance levels. This is in stark contrast to traditional uses of Elo ratings, where a one-size-fits-all K-factor is frequently applied.
- **Transitivity is not guaranteed**: The assumption that ($A$ beats $B$ and $B$ beats $C$ implies $A > C$) is not consistently valid in Elo ratings. This is particularly invalid when models have similar performance levels, challenging a common assumption in many Elo-based evaluations.

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.

## 9 Conclusion and Limitations

This paper presents a comprehensive study on the reliability of the Elo rating system for evaluating LLMs through human feedback within an axiomatic framework. We identify various factors that influence the robustness of Elo ratings and provide guidelines for their effective application in real-world scenarios. While our findings establish an essential foundation, 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 yield additional insights into the convergence properties of the Elo rating system in the fast-evolving field of language models.

## 10 Impact Statement

The implications of our work are significant in fields relying on LLMs for decision-making, content generation, and more. Improving the evaluation methods of LLMs contributes to the development of AI systems that are more reliable and trustworthy. This research also holds the potential to influence evaluation practices in other sectors that employ the Elo rating system, broadening its relevance and utility. However, it also emphasizes the need for cautious, informed application of Elo ratings to prevent misinterpretation or reliance on Elo-based rankings, particularly when the performance of models is comparable. As LLMs become more integrated into societal frameworks, ensuring the robustness and reliability of their evaluation mechanisms is paramount to fostering ethical, beneficial AI advancements.

## References


[1] Anthropic. Model card and evaluations for claude models. https://www-cdn.anthropic.com/5c49cc247484cecf107c699baf29250302e5da70/ModelCardClaudev2_with_appendix_v1.pdf, 2023.

[2] 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. A general language assistant as a laboratory for alignment, 2021.

[3] 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. Training a helpful and harmless assistant with reinforcement learning from human feedback, 2022.

[4] 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. Constitutional ai: Harmlessness from ai feedback, 2022.

[5] Lorena A. Barba. Terminologies for reproducible research, 2018.

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

[7] Quentin Bertrand, Wojciech Marian Czarnecki, and Gauthier Gidel. On the limitations of the elo, real-world games are transitive, not additive. In Francisco Ruiz, Jennifer Dy, and Jan-Willem van de Meent, editors, *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pages 2905–2921. PMLR, 25–27 Apr 2023.

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

[9] Xavier Bouthillier, Pierre Delaunay, Mirko Bronzi, Assya Trofimov, Brennan Nichyporuk, Justin Szeto, Nazanin Mohammadi Sepahvand, Edward Raff, Kanika Madan, Vikram Voleti, Samira Ebrahimi Kahou, Vincent Michalski, Tal Arbel, Chris Pal, Gael Varoquaux, and Pascal Vincent. Accounting for variance in machine learning benchmarks. In A. Smola, A. Dimakis, and I. Stoica, editors, *Proceedings of Machine Learning and Systems*, volume 3, pages 747–769, 2021.

[10] 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. A survey on evaluation of large language models, 2023.

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

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


 https://doi.org/10.52202/079017-3367

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

[14] Chris Drummond. Replicability is not reproducibility: Nor is it good science. In *Evaluation Methods for Machine Learning Workshop at the 26th International Conference on Machine Learning, ICML*, 2009.

[15] Aram Ebtekar and Paul Liu. 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, 2021. Association for Computing Machinery.

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

[17] ESL. Ranking - dota2 - esl pro tour.

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

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

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

[21] Steven N. Goodman, Daniele Fanelli, and John P. A. Ioannidis. What does research reproducibility mean? *Science Translational Medicine*, 8(341):341ps12–341ps12, 2016.

[22] Odd Erik Gundersen and Sigbjørn Kjensmo. State of the art: Reproducibility in artificial intelligence. In *AAAI*, 2018.

[23] Peter Henderson, Riashat Islam, Philip Bachman, Joelle Pineau, Doina Precup, and David Meger. Deep reinforcement learning that matters. *CoRR*, abs/1709.06560, 2017.

[24] Ralf Herbrich, Tom Minka, and Thore Graepel. Trueskill™: A bayesian skill rating system. In B. Schölkopf, J. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems*, volume 19. MIT Press, 2006.

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

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

[27] Rudolf Kadlec, Ondrej Bajgar, and Jan Kleindienst. Knowledge base completion: Baselines strike back. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*, pages 69–74, Vancouver, Canada, August 2017. Association for Computational Linguistics.

[28] Wei-Yin Ko, Daniel D'souza, Karina Nguyen, Randall Balestriero, and Sara Hooker. Fair-ensemble: When fairness naturally emerges from deep ensembling, 2023.

[29] Andreas Köpf, Yannic Kilcher, Dimitri von Rütte, Sotiris Anagnostidis, Zhi-Rui Tam, Keith Stevens, Abdullah Barhoum, Nguyen Minh Duc, Oliver Stanley, Richárd Nagyfi, Shahul ES, Sameer Suri, David Glushkov, Arnav Dantuluri, Andrew Maguire, Christoph Schuhmann, Huu Nguyen, and Alexander Mattick. Openassistant conversations – democratizing large language model alignment, 2023.

[30] Angeliki Lazaridou, Adhiguna Kuncoro, Elena Gribovskaya, Devang Agrawal, Adam Liska, Tayfun Terzi, Mai Gimenez, Cyprien de Masson d'Autume, Sebastian Ruder, Dani Yogatama, Kris Cao, Tomas Kocisky, Susannah Young, and Phil Blunsom. Pitfalls of static language modelling, 2021.

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

[32] 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. Holistic evaluation of language models, 2022.

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

[34] 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.

[35] Liquipedia. Elo rating - liquipedia starcraft brood war wiki.

[36] Mario Lucic, Karol Kurach, Marcin Michalski, Sylvain Gelly, and Olivier Bousquet. Are gans created equal? a large-scale study, 2018.

[37] Pranava Madhyastha and Rishabh Jain. On model stability as a function of random seed. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 929–939, Hong Kong, China, November 2019. Association for Computational Linguistics.

[38] Gábor Melis, Chris Dyer, and P. Blunsom. On the state of the art of evaluation in neural language models. *ArXiv*, abs/1707.05589, 2018.

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

[40] Prabhat Nagarajan, Garrett Warnell, and Peter Stone. The impact of nondeterminism on reproducibility in deep reinforcement learning. In *2nd Reproducibility in Machine Learning Workshop at ICML*, July 2018.

[41] Roman Novak, Yasaman Bahri, Daniel A. Abolafia, Jeffrey Pennington, and Jascha Sohl-Dickstein. Sensitivity and generalization in neural networks: an empirical study. In *International Conference on Learning Representations*, 2018.

[42] OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis,

 https://doi.org/10.52202/079017-3367

Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. Gpt-4 technical report, 2024.

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

[44] April M. Reid. Elo rating system for video games explained.

[45] Christopher J. Shallue, Jaehoon Lee, Joseph Antognini, Jascha Sohl-Dickstein, Roy Frostig, and George E. Dahl. Measuring the effects of data parallelism on neural network training, 2019.

[46] Gil I. Shamir, Dong Lin, and Lorenzo Coviello. Smooth activations and reproducibility in deep networks, 2020.

[47] Luísa Shimabucoro, Sebastian Ruder, Julia Kreutzer, Marzieh Fadaee, and Sara Hooker. Llm see, llm do: Guiding data generation to target non-differentiable objectives, 2024.

[48] Robert R. Snapp and Gil I. Shamir. Synthesizing irreproducibility in deep networks. *CoRR*, abs/2102.10696, 2021.

[49] Anders Søgaard, Sebastian Ebert, Jasmijn Bastings, and Katja Filippova. We need to talk about random splits. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1823–1832, Online, April 2021. Association for Computational Linguistics.

[50] 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

 https://doi.org/10.52202/079017-3367

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. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*, 2023.

[51] Cecilia Summers and Michael J. Dinneen. Nondeterminism and instability in neural network optimization. In *International Conference on Machine Learning*, 2021.

[52] Nelson Vadori and Rahul Savani. Ordinal potential-based player rating, 2023.

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

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

[55] Yining Ye, Xin Cong, Yujia Qin, Yankai Lin, Zhiyuan Liu, and Maosong Sun. Large language model as autonomous decision maker, 2023.

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

[57] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric. P Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena, 2023.

[58] Donglin Zhuang, Xingyao Zhang, Shuaiwen Leon Song, and Sara Hooker. Randomness in neural network training: Characterizing the impact of tooling. *ArXiv*, abs/2106.11872, 2021.

## A Extension to Multiple Outcomes

For scenarios where outcomes can extend beyond wins and losses, such as a tie option, one could make use of the multinomial distribution for the synthetic data generation process. For the three outcomes; win, loss, and tie, one sample according to the distribution:

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

## 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 $Prob(A \text{ beats } B) \geq 0.65$.

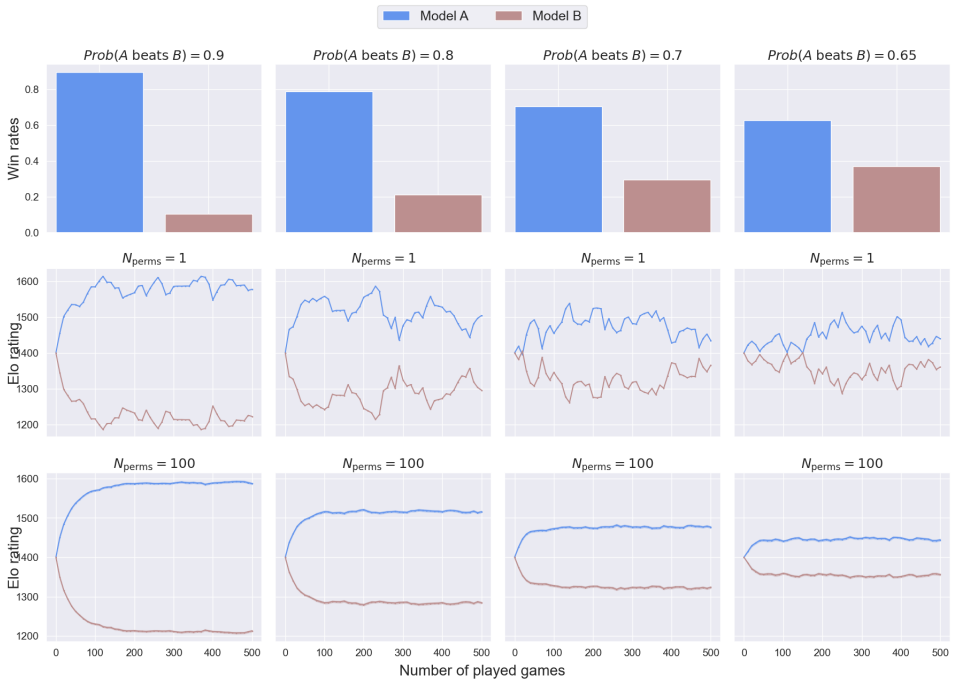



Figure 6: **Impact of win probabilities and permutation sampling on Elo ratings**: Comparing Model A and Model B across three different win probabilities ($Prob(A \text{ beats } B) = 0.9, 0.8, 0.7, 0.65$) with two levels of permutation sampling ($N_{\text{perms}} = 1$ and $N_{\text{perms}} = 100$). The 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 Chatbot Arena Human Preference Data Preparation

For the experimental validation of the Elo rating system using real-world data, we utilize the LMSYS dataset from [34]. We first vizualize the first 100 unique paired comparisons sorted in descending order by the number of recorded evaluations. The distribution of tie vs. non-tie outcomes is shown in Figure 7. To refine the dataset for our analysis, we exclude tie results, focusing exclusively on win-loss dynamics. The remaining dataset is further filtered to identify pairs with at least 300 non-tie comparisons. This threshold of 300 allows us to encompass a broad spectrum of comparison scenarios, ranging from skewed win rates ($\text{Model}_\text{A} \gg \text{Model}_\text{B}$) to closely matched ($\text{Model}_\text{A} \simeq \text{Model}_\text{B}$). These selected paired comparisons are depicted in Figure 8.


https://doi.org/10.52202/079017-3367

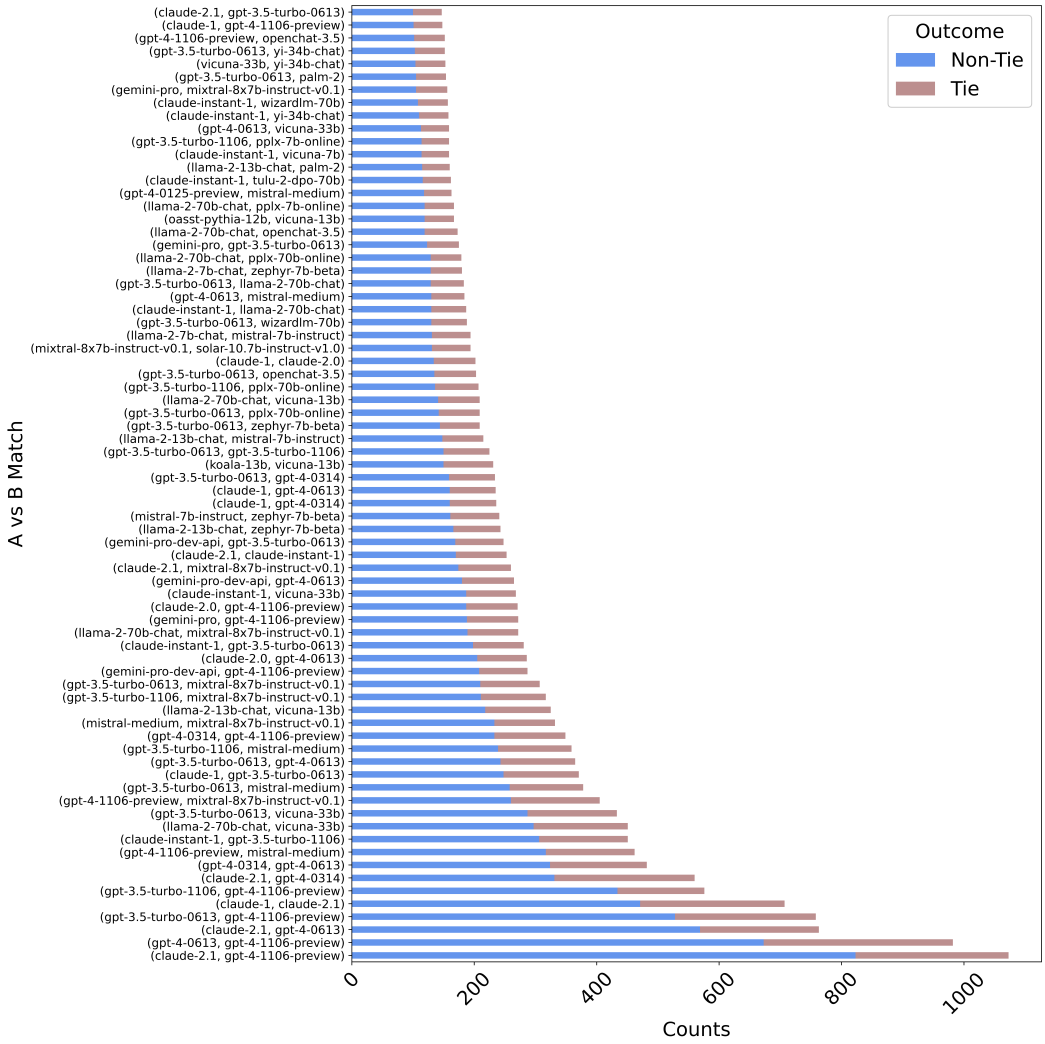



Figure 7: Initial Distribution of Tie vs. Non-Tie Outcomes: A visual overview of the first 100 paired comparisons from the LMSYS dataset ordered by evaluation sizwe

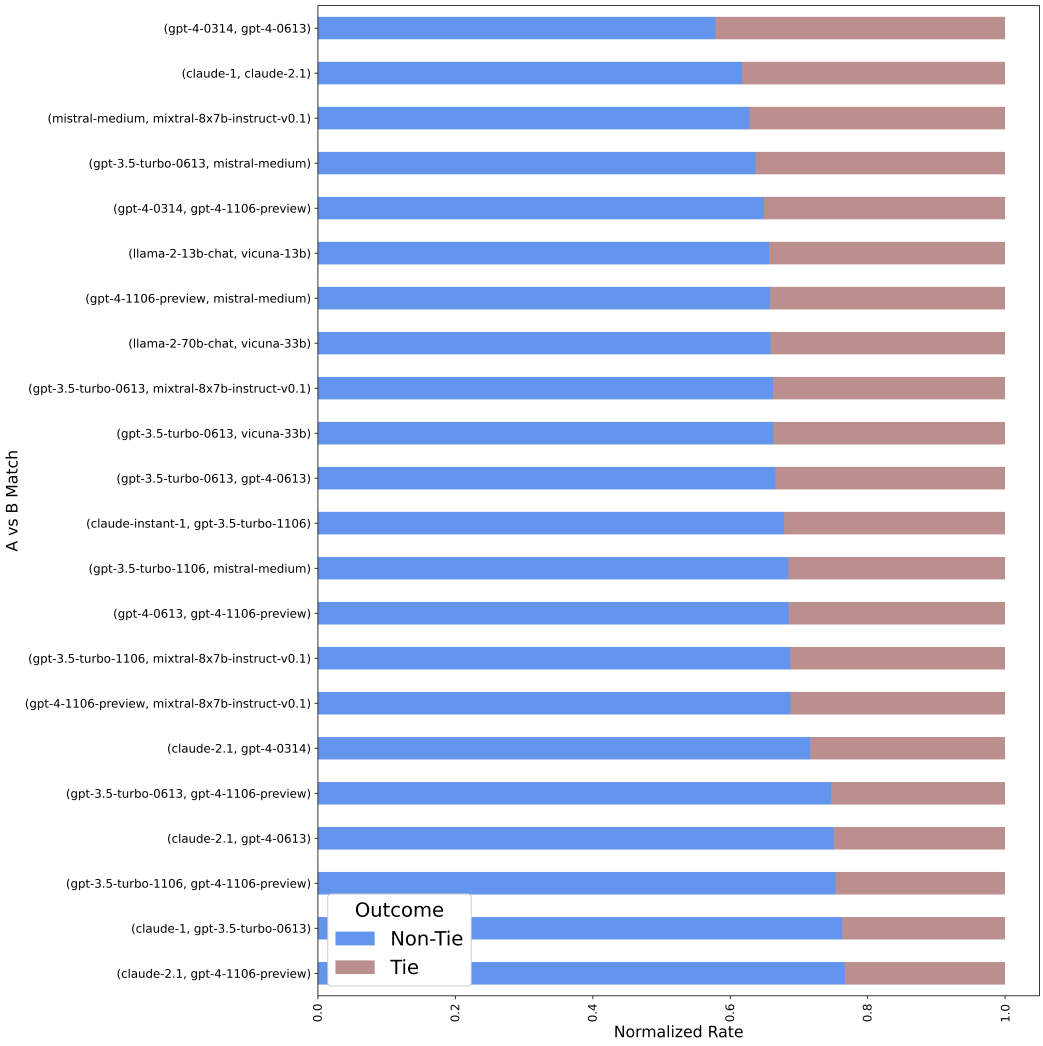



Figure 8: Normalized Tie vs. Non-Tie outcomes for A vs B model comparisons with at least 300 evaluations.


https://doi.org/10.52202/079017-3367

Table 3: Win Rates for Paired Model Evaluations: This table presents the initial match counts and win rates for model comparisons from [34] where each pair has at least 300 matches, excluding ties. Results from a fixed sample size of 300 are shown to demonstrate model performance under controlled sampling conditions.

| Experiment | Original Size | Win Rates (%) | Sample Size | Sampled Win Rates (%) |
|---|---|---|---|---|
| gpt-4-0314<br>gpt-4-0613 | 324 | 50.93<br>49.07 | 300 | 51.00<br>49.00 |
| gpt-4-1106-preview<br>gpt-4-0613 | 673 | 67.01<br>32.99 | 300 | 67.00<br>33.00 |
| gpt-4-1106-preview<br>gpt-3.5-turbo-0613 | 528 | 81.25<br>18.75 | 300 | 81.33<br>18.67 |
| gpt-4-1106-preview<br>gpt-3.5-turbo-1106 | 434 | 86.18<br>13.82 | 300 | 86.33<br>13.67 |
| claude-1<br>claude-2.1 | 471 | 58.60<br>41.40 | 300 | 58.67<br>41.33 |
| gpt-4-0314<br>claude-2.1 | 331 | 65.26<br>34.74 | 300 | 65.33<br>34.67 |
| gpt-4-0613<br>claude-2.1 | 569 | 61.16<br>38.84 | 300 | 61.00<br>39.00 |
| gpt-4-1106-preview<br>claude-2.1 | 823 | 75.21<br>24.79 | 300 | 75.33<br>24.67 |
| claude-instant-1<br>gpt-3.5-turbo-1106 | 306 | 56.86<br>43.14 | 300 | 57.00<br>43.00 |
| gpt-4-1106-preview<br>mistral-medium | 317 | 71.92<br>28.08 | 300 | 72.00<br>28.00 |

## NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction accurately outline the core investigation into the Elo rating system's reliability and transitivity properties forLLMs, aligning with the empirical results presented, which validate and support these claims.

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: This study probes the Elo rating system, highlighting its vulnerabilities without considering the impact of "tie" outcomes on Elo scores convergence. It shows its limitations for the most simplistic setting of pairwise comparison A vs B and proposes for future work an extension of tie outcomes to better model real-world scenarios.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**


https://doi.org/10.52202/079017-3367

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [NA]

Justification: Our study does not derive theoretical results. Our research primarily employs stress testing on synthetic data and subsequent validation through real-world experiments to evaluate the robustness of Elo scores.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: We detail the process of generating synthetic human feedback data and refer to a similar experimental setup used to collect real-world human feedback. Although results based on human judgment are inherently difficult to replicate exactly, the main focus of our study is on the reliability and robustness of the Elo rating system. This system requires only a sequence of outcomes (wins, losses, and ties, represented as 1, 0, or 0.5, respectively) to compute scores, facilitating the reproduction of our analysis based on these simplified inputs.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.

(c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: We provide thorough details on generating synthetic human feedback data, a key component of our experiments. We also outline our method for using LMSYS human preference data, which is publicly accessible. We also outline the steps for computing Elo scores based on existing literature. These comprehensive instructions should enable other researchers to effectively replicate our experiments and verify our results.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: In our robustness study of the Elo rating system, we provide detailed descriptions of our experimental setup, including how synthetic data is generated. We explore and visualize the effects of a wide range of hyperparameters, such as the number of permutations ($N_{\text{perms}}$) and the K-factor, to assess their impact on Elo score stability. We compare language models in the 7B to 12B parameters range, selected due to hardware computational limitations.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.


https://doi.org/10.52202/079017-3367

- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Our paper includes error bars to represent the uncertainty in experimental outcomes, specifically using standard errors of the mean (SEM) across different permutations and hyperparameter settings.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We generated completions in batches sized between 8 and 50, depending on the size of each model evaluated, using an Nvidia A100 GPU with 40GB memory for efficient computation. Inference was performed in a Bfloat16 setting to reduce memory usage, as deatailed in referenced work.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

Justification: Our research adheres to the NeurIPS Code of Ethics in all respects. We ensured fair compensation for human evaluators involved in the model pairwise comparisons. The pool of evaluation prompts includes open-source datasets (SODA and the Public Pool of Prompts (P3)), and were selected to avoid any harmful content.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: Our paper discusses several positive societal impacts, including the enhancement of AI system reliability and trustworthiness through more robust evaluation methods for large language models (LLMs). It highlights how these improved methods contribute to the development of safer and more accurate language models. We also address potential negative impacts by cautioning against the misinterpretation and over-reliance on Elo ratings, which might lead to suboptimal performance assessments. This underscores the need for exhaustive and fair evaluation mechanisms in AI development.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: We do not release any models or datasets, so there are no associated risks.

Guidelines:

- The answer NA means that the paper poses no such risks.

 https://doi.org/10.52202/079017-3367

- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We use open-source datasets and properly credit their creators by citing the original papers. The specific versions and licenses of these datasets are acknowledged, ensuring compliance with their terms of use.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: We do not release new assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [Yes]

Justification: Details regarding the instructions for human feedback experiments and evaluation format are based on established methodologies in the field. As for the compensation, while we cannot provide specific details, we can guarantee that fair compensation practices were followed.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [No]

Justification: No potential risk within our experimental setup.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.


https://doi.org/10.52202/079017-3367