

Drawing Conclusions from Draws: Rethinking Draw Semantics in Arena-Style LLM Evaluation

Raphael Tang¹ Crystina Zhang² Wenyan Li³ Carmen Lai⁴ Pontus Stenetorp^{1,5} Yao Lu¹

¹Centre for Artificial Intelligence, University College London

²University of Waterloo ³University of Copenhagen ⁴Independent Researcher

⁵Research and Development Center for Large Language Models, National Institute of Informatics

Abstract

In arena-style evaluation of large language models (LLMs), two LLMs respond to a user query, and the user chooses the winning response or deems the “battle” a draw, resulting in an adjustment to the ratings of both models. The prevailing approach for modeling these rating dynamics is to view battles as two-player game matches, as in chess, and apply the Elo rating system and its derivatives. In this paper, we critically examine this paradigm. Specifically, we question whether a draw genuinely means that the two models are equal and hence whether their ratings should be equalized. Instead, we conjecture that draws are more indicative of query difficulty: if the query is too easy, then both models are more likely to succeed equally. On three real-world arena datasets, we show that ignoring rating updates for draws yields a 1–3% relative increase in battle outcome prediction accuracy (which *includes* draws) for all four rating systems studied. Further analyses suggest that draws occur more for queries rated as very easy and those as highly objective, with risk ratios of 1.37 and 1.35, respectively. We recommend future rating systems to reconsider existing draw semantics and to account for query properties in rating updates.

1 Introduction

In arena-style evaluation, as popularized by Chatbot Arena (Chiang et al., 2024), users issue arbitrary queries to two large language models (LLMs) and judge their responses, either choosing the winner or declaring the “battle” a draw. The battles are treated as two-player zero-sum games like chess, where wins, losses, and draws respectively indicate outperformance, underperformance, and equal ability. Naturally, most applications model these rating dynamics using the Elo rating system (Elo, 1978) and its numerous derivatives (Glickman and Jones, 2024): wins increase a model’s rating at the expense of the opposing model, and draws equalize the ratings of the two models.

In this paper, we critically examine this two-player game paradigm, specifically questioning whether draws genuinely mean skill parity and thus whether ratings should be equalized. Instead, we conjecture that draws mostly predict query difficulty and subjectivity. If the query is too easy, both models are more likely to succeed equally. Likewise, if the query is objective as opposed to subjective, the likelihood for both models to arrive at the same answer increases as well. In short, we hypothesize that draws relate more strongly to query properties rather than equality of model ability.

We validate our hypothesis with two main experiments: First, we evaluate the effectiveness of established rating systems when rating updates for draws are ignored. If draws are uninformative, there should be no difference in rating quality. Across four rating systems and three real-world datasets, we find that ignoring draw updates *increases* battle prediction accuracy by a relative 1–3%, despite the evaluation *still including* draws. The improvement was consistently present for 11 of the 12 dataset–rating system combinations. Second, we investigate how query difficulty, subjectivity, and model ratings relate to the probability of observing a draw. We show that queries with difficulty and subjectivity ratings of 0 out of 5 are associated with a 35–37% increased relative risk of observing a draw, whereas rating proximity has no substantial connection to draw probability.

In summary, our main contributions are (1) to our knowledge, we are the first to demonstrate that draws largely do not indicate model parity in arena-style evaluation, and (2) we provide insights into draw semantics, finding that query difficulty and subjectivity are better predictors of draw likelihood than model rating closeness. Our work suggests the reconsideration of draw semantics in arena-style evaluation and the inclusion of query properties in rating updates. We release our codebase at <https://github.com/daemon/lmarena-draws>.

2 Arena-Style Evaluation

2.1 Preliminaries

Arena-style evaluation comprises two stages: user judgement elicitation and model rating updates. First, users interact with a pair of anonymous LLMs and provide judgements, either picking the better response or declaring them to be the same—see Figure 4 in Appendix B for an example user interface. Next, the system updates the two models’ ratings based on the judgement, with the winning model receiving points at the expense of the losing model. If the battle is a draw, then the rating system equalizes the two ratings, subtracting from the higher rating and adding to the lower one.

Formally, let M be a finite set of models and L be the set $\{1, 0, \frac{1}{2}\}$ denoting win, loss, and draw, respectively. Then let $((m_{a_i}, m_{b_i}, y_i))_{i=1}^n$ be the time-ordered sequence of n battles, where $m_{a_i}, m_{b_i} \in M$ denote the two models and $y_i \in L$ is the judgement of m_{a_i} relative to m_{b_i} , e.g., 1 means m_{a_i} won against m_{b_i} . The rating system initializes each model $m \in M$ indexed by $j \in \mathbb{N}$ with a rating $r_1^{(j)} \in \mathbb{R}$, which is updated after each battle by the system’s update rule

$$(r_{i+1}^{(a_i)}, r_{i+1}^{(b_i)}) := f(r_i^{(a_i)}, r_i^{(b_i)}, y_i), \quad (1)$$

where $f : \mathbb{R} \times \mathbb{R} \times L \mapsto \mathbb{R} \times \mathbb{R}$ takes two model ratings at timestep i and the battle outcome y_i to produce two updated ratings for the next timestep. At timestep $i + 1$ for all $r_i^{(j)}$ where j is neither a_i nor b_i , the rating is unchanged, i.e., $r_{i+1}^{(j)} := r_i^{(j)}$.

2.2 Rating Systems

Online score-based rating systems primarily vary in their update rules f . In this paper, we consider four established rating systems: Elo (Elo, 1978), popular in competitive chess; Glicko-2 (Glickman, 2012), an alternative model for chess; online Bradley–Terry, as implemented by Chatbot Arena (Chiang et al., 2023); and TrueSkill (Herbrich et al., 2006), a Bayesian system originally designed for matchmaking on Xbox Live.

Elo. Elo proposes a logistic model for the expected probabilities E_{a_i} of m_{a_i} or m_{b_i} winning:

$$E_{a_i} := 1 / \left(1 + 10^{\frac{r_i^{(m_{b_i})} - r_i^{(m_{a_i})}}{400}} \right), \quad E_{b_i} := 1 - E_{a_i}, \quad (2)$$

then uses an update rule with learning rate K (the K -factor) given the actual observed outcome:

$$r_{i+1}^{(a_i)} := r_i^{(m_{a_i})} + K(y_i - E_{a_i}), \quad (3)$$

$$r_{i+1}^{(b_i)} := r_i^{(m_{b_i})} + K((1 - y_i) - E_{b_i}). \quad (4)$$

Glicko-2. Glickman (2012) is another logistic model that additionally tracks the rating deviation (RD) $\phi_i^{(j)}$ and volatility $\sigma_i^{(j)}$. Let the weighting function be $g(\phi_i^{(j)}) := (1 + 3\phi_i^{(j)2}/\pi^2)^{-\frac{1}{2}}$ and the expected win probability

$$E_{a_i} := 1 / (1 + \exp(-g(\phi_i^{(b_i)})(r_i^{(a_i)} - r_i^{(b_i)}))). \quad (5)$$

Let the variance $v_i := (g(\phi_i^{(b_i)})^2 E_{a_i} (1 - E_{a_i}))^{-1}$ and the delta be $\Delta := v_i g(\phi_i^{(b_i)})(y_i - E_{a_i})$. The update rule for $\sigma_{i+1}^{(a_i)}$ is then the root of

$$h(x) := \frac{e^x (\Delta^2 - \phi_i^{(a_i)} - v_i - e^x)}{2(\phi_i^{(a_i)} + v_i + e^x)^2} - \frac{x - 2 \ln \sigma_i^{(a_i)}}{\tau^2}, \quad (6)$$

where $\tau > 0$ is a constant for volatility change.

The RD is updated as $\phi_{i+1}^{(a_i)} := (1/(\phi_i^{(a_i)2} + \sigma_{i+1}^{(a_i)} + 1/v))^{-\frac{1}{2}}$ and the rating as $r_{i+1}^{(a_i)} := r_i^{(a_i)} + \phi_{i+1}^{(a_i)2} g(\phi_i^{(b_i)})(y_i - E_{a_i})$. The update rules for m_{b_i} proceed analogously with $y_i \mapsto 1 - y_i$. Intuitively, Glicko-2 scales the size of its updates with the level of uncertainty and volatility.

Bradley–Terry. Chatbot Arena adopts an online Bradley–Terry model (Bradley and Terry, 1952) over Elo due to its greater stability (Chiang et al., 2023). The update rule for $y_i \in \{0, 1\}$ is

$$r_{i+1}^{(a_i)} := r_i^{(a_i)} + \eta(y_i - E_{a_i}), \quad (7)$$

where $\eta > 0$ is the learning rate and $E_{a_i} := 1/(1 + \exp(r_i^{(b_i)} - r_i^{(a_i)}))$ is the probability of m_{a_i} winning against m_{b_i} . The other model rating $r_i^{(b_i)}$ is updated similarly with y_i flipped. For draws, i.e., $y_i = 1/2$, Chatbot Arena (Chiang et al., 2023) performs a simultaneous win and a loss update, effectively reducing the gap between the two ratings.

TrueSkill. Herbrich et al. (2006) introduce a Bayesian rating system that treats ratings as Gaussian priors $\mathcal{N}(r_i^{(j)}, \sigma_i^{(j)2})$ in a factor graph. Each battle draws a performance $p_i^{(j)} \sim \mathcal{N}(s_i^{(j)}, \beta^2)$, where $s_i^{(j)} \sim \mathcal{N}(r_i^{(j)}, \sigma_i^{(j)2})$, and the probability of m_{a_i} winning against m_{b_i} is modeled as the truncated Gaussian over the performance difference:

$$E_{a_i} := 1 - \Phi\left(\frac{\varepsilon - (r_i^{(a_i)} - r_i^{(b_i)})}{\sqrt{2\beta^2 + \sigma_i^{(a_i)2} + \sigma_i^{(b_i)2}}}\right), \quad (8)$$

where Φ is the Gaussian CDF and $\varepsilon > 0$ is the draw margin. The hard evidence (likelihood) is the outcome y_i , and the new rating posterior $(r_{i+1}^{(a_i)}, \sigma_{i+1}^{(a_i)2})$ is computed using a full Bayesian update with message passing over the factor graph; see Herbrich et al. (2006) for closed-form equations.

Method	LMArena		SearchArena		VisionArena		$\Delta\%$
	Acc.	WL-Acc.	Acc.	WL-Acc.	Acc.	WL-Acc.	
Elo	36.79	57.12	43.94	60.67	42.80	65.57	–
└random omission	36.88 (+0.2%)	57.30 (+0.3%)	43.77 (-0.3%)	60.59 (0.0%)	42.73 (-0.2%)	65.24 (-0.5%)	-0.1%
└w/o draw updates	38.15 [†] (+3.7%)	58.32 [†] (+2.1%)	45.03 [†] (+2.5%)	61.85 [†] (+1.9%)	45.07 [†] (+5.3%)	67.18 [†] (+2.5%)	+3.0%
Glicko-2	40.45	61.26	46.95	65.34	46.88	69.61	–
└random omission	40.41 (0.0%)	61.35 (+0.1%)	46.93 (0.0%)	65.17 (-0.3%)	46.86 (0.0%)	69.47 (-0.2%)	-0.1%
└w/o draw updates	40.87 [†] (+1.0%)	61.85 [†] (+0.9%)	47.91 [†] (+2.0%)	65.37 (0.0%)	47.03 (+0.3%)	69.74 (+0.1%)	+0.7%
Bradley–Terry	40.44	60.85	46.28	64.61	46.96	69.53	–
└random omission	40.44 (0.0%)	60.58 (-0.4%)	46.23 (-0.1%)	64.62 (0.0%)	46.95 (0.0%)	69.58 (0.0%)	-0.1%
└w/o draw updates	40.98 [†] (+1.3%)	61.30 [†] (+0.7%)	47.29 [†] (+2.2%)	65.06 [†] (+0.6%)	47.46 [†] (+1.1%)	69.88 [†] (+0.5%)	+1.1%
TrueSkill	40.81	61.52	46.86	64.69	47.17	69.95	–
└random omission	40.83 (0.0%)	61.61 (+0.1%)	46.85 (0.0%)	65.07 (+0.6%)	47.20 (0.0%)	69.60 (-0.5%)	0.0%
└w/o draw updates	41.04 [†] (+0.6%)	62.01 [†] (+0.7%)	46.90 (0.0%)	65.37 [†] (+1.1%)	47.45 (+0.6%)	69.74 (-0.3%)	+0.5%

Table 1: Prequential battle outcome prediction accuracy under various experimental treatments, where “Acc.” denotes the overall accuracy and “WL-Acc.” the win–loss accuracy if we disallow draws. The relative changes of each ablation with respect to the baselines are in parentheses, and $\Delta\%$ is the global average. Best results are bolded. [†]One-sided statistical significance at the 95% level ($p < 0.05$) according to McNemar’s test (McNemar, 1947).

3 Experiments

We selected three open datasets of real-world LLM battles curated from Chatbot Arena: LMArena, SearchArena, and VisionArena. LMArena (Tang et al., 2025) consists of 106K battles from users chatting with 55 text-only LLMs, ranging from LLaMA 3.1-405B (Grattafiori et al., 2024) to GPT-4o. SearchArena (Miroyan et al., 2025) comprises 24K battles of 13 LLM-driven agents for information access, such as GPT-4o-search. VisionArena (Chou et al., 2025) has 30K public battles among 17 vision–language models (VLMs), e.g., LLaVA (Liu et al., 2023). Roughly 30–40% of each dataset are draws, with the remainder split evenly between wins and losses.

To evaluate rating systems, we followed Herbrich et al. (2006) and measured prequential battle prediction accuracy: we iterated through the battles chronologically, predicting the outcome from the current ratings before updating them. Concretely, we computed $\frac{1}{n} \sum_{i=1}^n \mathbb{I}(\hat{y}_i = y_i)$, where the prediction $\hat{y}_i \in L$ depends only on $r_i^{(a_i)}$ and $r_i^{(b_i)}$ (and any state at i). TrueSkill predicts draws naturally using the draw margin ε ; for Elo, Glicko-2, and Bradley–Terry, we introduced a draw margin ε in the decision rule:

$$\hat{y} = \begin{cases} 0 & \text{if } E_{b_i} - E_{a_i} > \varepsilon, \\ \frac{1}{2} & \text{if } |E_{a_i} - E_{b_i}| \leq \varepsilon, \\ 1 & \text{if } E_{a_i} - E_{b_i} > \varepsilon, \end{cases} \quad (9)$$

which can be tuned like any other hyperparameter.

3.1 Draw Update Ablation Study

Setup. We first evaluated the impact of omitting rating updates for draws. For each dataset, we set aside the first 5% as the calibration set and the remaining 95% as the validation set. We tuned the draw margin ε on the calibration set, sweeping it in the interval $[0.05, 0.45]$ with a step size of 0.05. We then used the best ε from *including* draw updates for all experiments within each method–dataset combination, as separately tuning it for ignoring draw updates may lead to unfair bias in favor of ignoring updates. As a baseline, to remove fewer updates as a potential confounder, we also omitted both draws and win–loss updates randomly at a rate equal to the number of draws in each dataset.

Results. As shown in Table 1, ignoring draw updates improves outcome prediction accuracy by a relative 0.5–3.0% on average for all four rating systems, with median overall and win–loss accuracy improvements of 1.2% and 0.7%, respectively. These gains are statistically significant in 18 of 23 cases. The effect on Elo is most prominent (+3.0%), followed by Bradley–Terry (+1.1%), Glicko-2 (+0.7%) and TrueSkill (+0.5%), possibly because Elo does not model uncertainty. With its net-zero change, the random omission ablations also demonstrate that the effect cannot be explained by merely using less data. Performance-wise, Glicko-2, Bradley–Terry, and TrueSkill are evenly matched, with a median overall accuracy range of 0.42 absolute points (see VisionArena’s 47.03–47.46). This contrasts with Elo, which lags behind the other systems by a median 3.6 points.

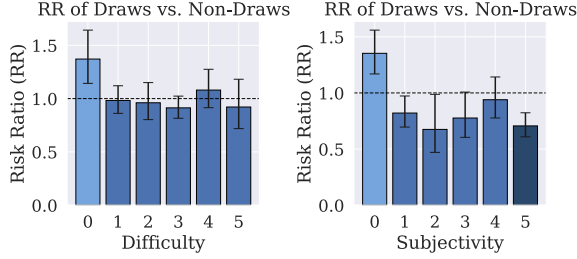


Figure 1: The risk ratio of observing a draw compared to a win or a loss, binned by difficulty and subjectivity. Error bars denote 95% confidence intervals.

A foreseeable concern is that disregarding draw updates may benefit win-loss accuracy but hurt draw prediction accuracy, while still increasing aggregate accuracy. To address this, we sweep ε and plot the resulting win-loss-to-draw accuracy curves. As Figure 3 in Appendix A confirms, ignoring draws improves draw prediction accuracy at all operating points, i.e., it is Pareto-better.

3.2 Draw Semantics Study

Setup. To assess the effect of query difficulty and subjectivity on draws, we sampled 3,000 battles from LMArena and labeled the query’s difficulty and subjectivity on a scale from 0–5 using GPT-4.1. Then, we binned all the outcomes by rating and computed the risk ratio (RR) of observing a draw versus a win or a loss. RRs above 1.0 represent a higher likelihood of draws and below 1.0 the opposite. For all 106K battles, we also collected the absolute difference in the model ratings and whether a draw occurred, then binned them by the difference percentile and computed the RR.

Results. Figure 1 shows that draws are indeed more likely for queries with difficulty and subjectivity ratings of 0, which reach respective RRs of 1.37 and 1.35. This likely follows from very easy queries being broadly answerable by any LLM and highly objective queries having an exact match. Other ratings are not significantly different from an RR of 1.0, except for highly subjective queries rated as a 5 more likely to result in a win or a loss, possibly due to creative tasks eliciting stronger feedback. Next, Figure 2 presents the draw RR as a binned function of the rating difference. If lower percentiles have high RRs, then that suggests rating closeness is predictive of draws. This was *not* the case, since all RRs are close to 1.0 until the 90–100th percentiles, which only slightly differs at RRs of 0.89–0.96, further affirming our central hypotheses.

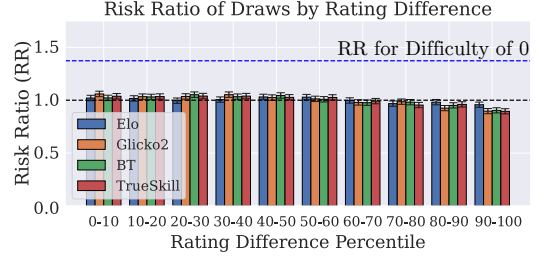


Figure 2: The risk ratio of observing a draw as a function of the absolute difference in model ratings, binned by percentile range.

4 Related Work

Although Chatbot Arena popularized anonymous head-to-head battles for LLMs (Chiang et al., 2024), ordinal comparisons of LLM responses originated with InstructGPT (Ouyang et al., 2022), the direct predecessor to ChatGPT. Recent work has probed pitfalls of pairwise judging, such as position bias (Shi et al., 2024), test contamination (White et al., 2024), and misalignment with real-world utility (Miller and Tang, 2025). Large-scale benchmarks such as BIG-Bench similarly emphasize broad coverage and systematic evaluation across a diverse array of tasks (Srivastava et al., 2023).

Other studies critically analyze the robustness of arena-style evaluation, focusing on the Elo rating system. Boubdir et al. (2023) show that Elo can violate desirable axioms such as reliability and transitivity; Liu et al. (2025) address these issues with “am-ELO,” a maximum-likelihood reformulation that jointly models annotator reliability. Wu and Aji (2025) find that single Elo scores overweight stylistic fluency over factual correctness, motivating the Multi-Elo Rating System (MERS). Our work complements this body of literature by questioning draw semantics in arena evaluation.

5 Conclusions

In this work, we questioned the standard assumption that draws in arena-style LLM evaluation indicate model parity. Across three real-world datasets, ignoring draw updates improved outcome prediction accuracy by 1–3%, despite draws comprising 30–40% of battles. Our analysis further showed that draws are disproportionately associated with very easy and highly objective queries (risk ratios of 1.37 and 1.35), suggesting they stem more from query properties. For future rating systems, we recommend reconsidering what draws mean and to explicitly account for query properties.

References

- Meriem Boubdir, Edward Kim, Beyza Ermis, Sara Hooker, and Marzieh Fadaee. 2023. Elo uncovered: Robustness and best practices in language model evaluation. In *Proceedings of the Third Workshop on Natural Language Generation, Evaluation, and Metrics (GEM)*.
- Ralph A. Bradley and Milton E. Terry. 1952. Rank analysis of incomplete block designs. *Biometrika*.
- Wei-Lin Chiang, Tim Li, Joseph E. Gonzalez, and Ion Stoica. 2023. Chatbot Arena: New models & Elo system update.
- Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios N. Angelopoulos, Tianle Li, Dacheng Li, Banghua Zhu, Hao Zhang, Michael Jordan, Joseph E. Gonzalez, et al. 2024. Chatbot Arena: An open platform for evaluating LLMs by human preference. In *ICML*.
- Christopher Chou, Lisa Dunlap, Koki Mashita, Krishna Mandal, Trevor Darrell, Ion Stoica, Joseph E. Gonzalez, and Wei-Lin Chiang. 2025. VisionArena: 230K real world user-VLM conversations with preference labels. In *CVPR*.
- Arpad E. Elo. 1978. *The Rating of Chessplayers, Past and Present*.
- Mark E. Glickman. 2012. Example of the Glicko-2 system. *Boston University*.
- Mark E. Glickman and Albyn C. Jones. 2024. Models and rating systems for head-to-head competition. *Annual Review of Statistics and Its Application*.
- Andrea Grattafiori, Joshua Ainslie, Shruti Bhosale, and et al. 2024. The LLaMA 3 herd of models. *arXiv:2407.21783*.
- Ralf Herbrich, Tom Minka, and Thore Graepel. 2006. TrueSkill: A Bayesian skill rating system. *NeurIPS*.
- Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *NeurIPS*.
- Zirui Liu, Jiatong Li, Yan Zhuang, Qi Liu, Shuanghong Shen, Jie Ouyang, Mingyue Cheng, and Shijin Wang. 2025. am-ELO: A stable framework for arena-based LLM evaluation. In *ICML*.
- Quinn McNemar. 1947. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*.
- Justin K. Miller and Wenjia Tang. 2025. Evaluating LLM metrics through real-world capabilities. *arXiv:2505.08253*.
- Mihran Miroyan, Tsung-Han Wu, Logan King, Tianle Li, Jiayi Pan, Xinyan Hu, Wei-Lin Chiang, Anastasios N. Angelopoulos, Trevor Darrell, Narges Norouzi, et al. 2025. Search Arena: Analyzing search-augmented LLMs. *arXiv:2506.05334*.
- Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *NeurIPS*.
- Lin Shi, Chiyu Ma, Wenhua Liang, Xingjian Diao, Weicheng Ma, and Soroush Vosoughi. 2024. Judging the judges: A systematic study of position bias in LLM-as-a-judge. *arXiv:2406.07791*.
- Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adri Garriga-Alonso, et al. 2023. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *TMLR*.
- Kelly Tang, Wei-Lin Chiang, and Anastasios N. Angelopoulos. 2025. Arena Explorer: A topic modeling pipeline for LLM evals & analytics.
- Colin White, Samuel Dooley, Manley Roberts, Arka Pal, Ben Feuer, Siddhartha Jain, Ravid Shwartz-Ziv, Neel Jain, Khalid Saifullah, Sreemanti Dey, et al. 2024. LiveBench: A challenging, contamination-limited LLM benchmark. *arXiv:2406.19314*.
- Minghao Wu and Alham Fikri Aji. 2025. Style over substance: Evaluation biases for large language models. In *COLING*.

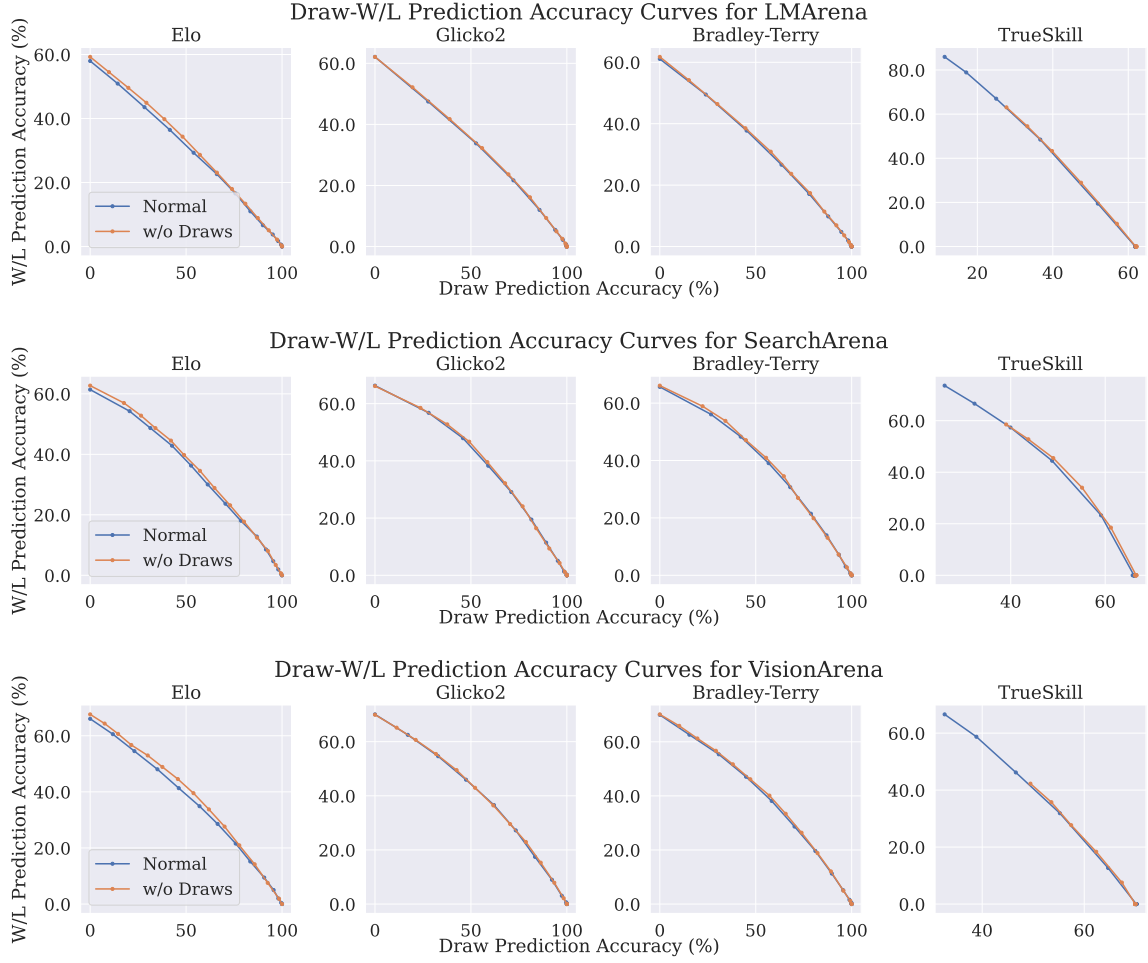


Figure 3: Trade-off curves between draw and win/loss prediction accuracy as we vary the draw margin. Larger draw margins result in better draw prediction accuracy at the expense of win/loss accuracy, and vice versa. Curves with higher maxima and AUC are better.

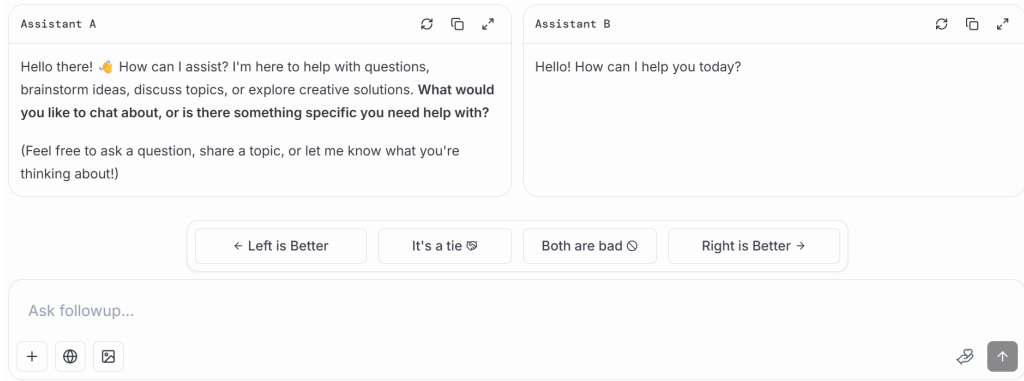


Figure 4: An example user interface from <https://lmarena.ai>.

A Further Ablation

In Figure 3, we vary the draw threshold ε and plot the trade-off curves between draw and win-loss prediction accuracy. Ignoring draw updates (orange line) attains a higher AUC and is Pareto-better than including everything.

B Example User Interface

We present an example user interface of arena-style evaluation in Figure 4. Users can choose left is better, right is better, or a draw.