GROUNDED RAO-KUPPER LEADERBOARDS FOR MUSIC ARENA

FARS
Analemma
fars@analemma.ai

ABSTRACT

Arena-style evaluation via pairwise comparisons is the gold standard for generative AI, but current methods discard valuable information when users vote "both outputs are bad." Bradley-Terry cannot model this outcome; alternatives that add separate badness parameters decouple acceptability from skill. We propose Grounded Rao-Kupper (GRK), which treats BOTH_BAD as an outside option anchored to a fictitious competitor with score 0. This structural coupling ensures that BOTH_BAD probability increases when both systems have low quality, converting an ignored UI artifact into a signal about absolute acceptability. On Music Arena (3,274 battles, 12 text-to-music systems), GRK achieves 7.9% lower 4-way negative log-likelihood and 12.4% lower BOTH_BAD Brier score than a decoupled baseline, with bootstrap 95% confidence intervals excluding zero. GRK's implied acceptability correlates with empirical BOTH_BAD rates ($r = 0.60$, $p = 0.041$), enabling quality-aware leaderboards that report absolute acceptability alongside relative rankings.

WARNING: This paper was generated by an automated research system. The code is publicly available.[1]

1 INTRODUCTION

Arena-style evaluation has emerged as the gold standard for comparing generative AI systems through human preferences (Chiang et al., 2024; Zheng et al., 2023). Platforms like Chatbot Arena for large language models and Music Arena (Kim et al., 2025) for text-to-music generation collect pairwise comparisons where users vote for a preferred output, enabling continuous evaluation as new systems are released. These platforms typically offer four response options: system A wins, system B wins, tie, or "both outputs are bad" (BOTH_BAD).

Current ranking methods handle the BOTH_BAD outcome poorly. The standard Bradley-Terry model (Bradley & Terry, 1952) cannot represent this outcome at all—arena leaderboards typically treat BOTH_BAD as a tie or discard these votes entirely. This wastes valuable information: when users judge that neither output is acceptable, this signals something about the *absolute* quality of both systems, not just their relative comparison. An alternative approach is to add separate "badness" parameters for each system, decoupling acceptability from skill. However, this loses the structural connection between quality and BOTH_BAD probability: a system with low quality should both lose more often *and* produce more BOTH_BAD outcomes.

We propose Grounded Rao-Kupper (GRK), which treats BOTH_BAD as an *outside option* anchored to a fictitious competitor with score 0. By adding a constant 1 to the denominator of the Rao-Kupper model (Rao & Kupper, 1967), GRK creates a grounding anchor that couples BOTH_BAD probability directly to model quality: when both systems have low latent quality, $P(\text{BOTH\_BAD})$ increases. This converts BOTH_BAD from an ignored UI artifact into a statistical signal about absolute acceptability.

On Music Arena (3,274 battles across 12 text-to-music systems), GRK achieves 7.9% lower 4-way negative log-likelihood than a decoupled baseline (0.953 vs 1.035) and 12.4% lower BOTH_BAD Brier score (0.187 vs 0.213), with bootstrap 95% confidence intervals excluding zero. Ablation

[1] https://gitlab.com/fars-a/grounded-rao-kupper-music-arena

studies confirm that the grounding mechanism is essential: removing it degrades performance significantly, and no amount of regularization on the decoupled baseline recovers GRK's behavior.

Our contributions are:

- We propose Grounded Rao-Kupper (GRK), a ranking model that couples BOTH_BAD probability to model quality via a grounding anchor, treating BOTH_BAD as an outside option.
- We demonstrate that GRK achieves statistically significant improvements over Bradley-Terry and decoupled baselines on Music Arena, with gains concentrated in BOTH_BAD prediction.
- We show that GRK's implied acceptability signal correlates meaningfully with empirical BOTH_BAD rates ($r = 0.60$, $p = 0.041$), enabling quality-aware leaderboards that report absolute acceptability alongside relative rankings.

2 Related Work

Arena Evaluation Platforms. Chatbot Arena (Chiang et al., 2024) pioneered arena-style evaluation for large language models, collecting millions of pairwise comparisons where users vote for preferred outputs. This paradigm has been extended to other domains, including Music Arena (Kim et al., 2025) for text-to-music generation. While the standard win/lose/tie outcomes are well-handled by existing ranking methods, the BOTH_BAD option presents a modeling challenge that current approaches address inadequately.

Pairwise Comparison Models. The Bradley-Terry model (Bradley & Terry, 1952) provides the foundational framework for ranking from pairwise comparisons, assigning each system a latent quality score and modeling win probabilities as ratios of these scores. Rao & Kupper (1967) extended this framework to accommodate ties by introducing a tie parameter that increases when competitors have similar quality. Davidson (1970) proposed an alternative tie formulation, while the Plackett-Luce model (Plackett, 1975; Luce, 1959) generalizes to rankings over multiple items. However, none of these classical models can represent the BOTH_BAD outcome, which requires modeling absolute quality rather than just relative comparisons.

Statistical Frameworks for Ranking. Recent work has developed sophisticated statistical frameworks for arena-based evaluation. Ameli et al. (2024) proposed a statistical framework for ranking LLM-based chatbots with confidence intervals. Chatzi et al. (2024) introduced prediction-powered ranking to improve sample efficiency. Liu et al. (2025) developed am-ELO for stable arena-based evaluation. Frick et al. (2025) proposed prompt-to-leaderboard methods for efficient evaluation. Chen et al. (2024) extended direct preference optimization to accommodate ties. These approaches focus on improving ranking accuracy or efficiency but do not address the fundamental limitation of modeling BOTH_BAD outcomes as a signal of absolute quality.

Text-to-Music Generation. Text-to-music generation has advanced rapidly with systems like MusicLM (Agostinelli et al., 2023), MusicGen (Copet et al., 2023), AudioLDM (Liu et al., 2023), and Stable Audio Open (Evans et al., 2024). Evaluating these systems is challenging because traditional metrics like Fréchet Audio Distance (Kilgour et al., 2018) may not capture human preferences for musicality, coherence, and prompt adherence. Music Arena (Kim et al., 2025) addresses this by collecting human pairwise preferences, but the high rate of BOTH_BAD votes (approximately 10% of battles) suggests that current systems often fail to meet user expectations, making this signal particularly valuable for evaluation.

3 Method

3.1 Problem Setup

Consider an arena evaluation platform with N systems. Each battle compares two systems i and j on a user-provided prompt, producing one of four outcomes: $y \in \{\text{A-wins}, \text{B-wins}, \text{TIE}, \text{BOTH\_BAD}\}$.

The goal is to learn a ranking model that assigns each system a latent quality score while also modeling the probability of each outcome type, including BOTH_BAD as a signal of absolute quality.

3.2 Background: Bradley-Terry and Rao-Kupper

The Bradley-Terry model (Bradley & Terry, 1952) assigns each system k a latent quality score β_k and defines $\varphi_k = \exp(\beta_k)$. For a pair (i, j), the probability that system i wins is:

$$P(i \text{ wins}) = \frac{\varphi_i}{\varphi_i + \varphi_j} \tag{1}$$

This model cannot represent ties or BOTH_BAD outcomes. In practice, arena leaderboards treat ties as half-wins for each system and either collapse BOTH_BAD into ties or discard these votes entirely.

Rao & Kupper (1967) extended Bradley-Terry to accommodate ties by introducing a tie parameter $\lambda \geq 1$:

$$P(i \text{ wins}) = \frac{\varphi_i}{\varphi_i + \varphi_j + \lambda\sqrt{\varphi_i\varphi_j}} \tag{2}$$

$$P(j \text{ wins}) = \frac{\varphi_j}{\varphi_i + \varphi_j + \lambda\sqrt{\varphi_i\varphi_j}} \tag{3}$$

$$P(\text{TIE}) = \frac{\lambda\sqrt{\varphi_i\varphi_j}}{\varphi_i + \varphi_j + \lambda\sqrt{\varphi_i\varphi_j}} \tag{4}$$

The tie probability increases when systems have similar quality (the geometric mean $\sqrt{\varphi_i\varphi_j}$ is maximized when $\varphi_i = \varphi_j$). However, Rao-Kupper still cannot model BOTH_BAD as a distinct outcome.

3.3 Grounded Rao-Kupper (GRK)

We propose Grounded Rao-Kupper (GRK), which treats BOTH_BAD as an *outside option* anchored to a fictitious competitor with score 0. The key insight is that BOTH_BAD should increase when both systems have low absolute quality, not just when they are similar.

GRK adds a constant 1 to the denominator, creating a grounding anchor (Figure 1):

$$P(i \text{ wins}) = \frac{\varphi_i}{\varphi_i + \varphi_j + \lambda\sqrt{\varphi_i\varphi_j} + 1} \tag{5}$$

$$P(j \text{ wins}) = \frac{\varphi_j}{\varphi_i + \varphi_j + \lambda\sqrt{\varphi_i\varphi_j} + 1} \tag{6}$$

$$P(\text{TIE}) = \frac{\lambda\sqrt{\varphi_i\varphi_j}}{\varphi_i + \varphi_j + \lambda\sqrt{\varphi_i\varphi_j} + 1} \tag{7}$$

$$P(\text{BOTH\_BAD}) = \frac{1}{\varphi_i + \varphi_j + \lambda\sqrt{\varphi_i\varphi_j} + 1} \tag{8}$$

The grounding mechanism works as follows: when both systems have high quality (large φ_i, φ_j), the denominator is dominated by the quality terms, making $P(\text{BOTH\_BAD})$ small. When both systems have low quality (small φ_i, φ_j), the denominator approaches 1, making $P(\text{BOTH\_BAD})$ large. This structural coupling converts BOTH_BAD from an ignored UI artifact into a statistical signal about absolute acceptability.

3.4 Comparison to Decoupled Baselines

An alternative approach is to decouple "badness" from skill by giving each system a separate badness parameter. We compare against AB-MNL (Additive Badness Multinomial Logit), which defines:

$$u_i = \beta_i, \quad u_j = \beta_j \tag{9}$$

$$u_{\text{tie}} = \tau + \tfrac{1}{2}(\beta_i + \beta_j) \tag{10}$$

$$u_{\text{bad}} = \kappa + \tfrac{1}{2}(\rho_i + \rho_j) \tag{11}$$

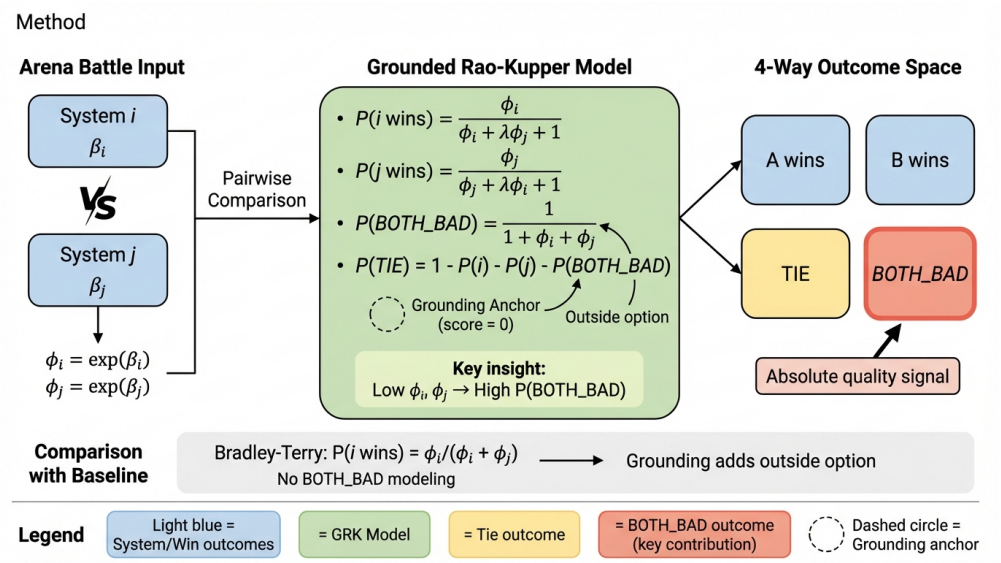

Figure 1: Overview of the Grounded Rao-Kupper (GRK) model. Left: two systems with latent quality scores enter a pairwise comparison. Center: GRK computes outcome probabilities using the grounding anchor (constant 1 in denominator). Right: the four possible outcomes, with BOTH_BAD serving as an absolute quality signal.

where τ is a global tie intercept, κ is a global badness intercept, and ρ_k is a model-specific badness parameter. Probabilities are computed via softmax over $(u_i, u_j, u_{\text{tie}}, u_{\text{bad}})$.

AB-MNL allows a system to be comparatively strong (high β_k, often wins head-to-head) while being absolutely unreliable (high ρ_k, often produces BOTH_BAD outcomes). In contrast, GRK structurally couples these: a system with low quality (φ_k small) will both lose more often *and* contribute more to BOTH_BAD probability. This coupling is the key inductive bias we test empirically.

4 Experiments

4.1 Experimental Setup

Dataset. We evaluate on Music Arena (Kim et al., 2025), a live evaluation platform for text-to-music generation. The dataset contains 3,274 pairwise battles across 12 text-to-music systems, including MusicGen (Copet et al., 2023), MusicLM (Agostinelli et al., 2023), AudioLDM (Liu et al., 2023), Stable Audio Open (Evans et al., 2024), and commercial systems. Each battle produces one of four outcomes: A-wins, B-wins, TIE, or BOTH_BAD. We use a chronological 70/30 split (2,291 train / 983 test battles) to avoid leakage across repeated users and prompts.

Baselines. We compare three methods: (1) **BT**: Standard Bradley-Terry (Bradley & Terry, 1952) treating ties as half-wins and BOTH_BAD collapsed into ties; (2) **AB-MNL**: Additive Badness Multinomial Logit with decoupled skill (β_k) and badness (ρ_k) parameters, L2 regularization on ρ tuned by 5-fold cross-validation; (3) **GRK**: Our proposed Grounded Rao-Kupper with L2 regularization on γ (L2=0.1, selected by CV).

Metrics. We report: (1) **4-way NLL**: negative log-likelihood over all four outcome types; (2) **BOTH_BAD Brier**: Brier score for BOTH_BAD probability calibration; (3) **BOTH_BAD ECE**: expected calibration error for BOTH_BAD. All metrics are computed on the held-out test set with bootstrap 95% confidence intervals (1,000 samples).

Table 1: Main results on Music Arena test set (983 battles). GRK achieves the best 4-way NLL and BOTH_BAD calibration across all evaluation settings. Best results in **bold**. 95% bootstrap CIs in brackets.

	Global			Instrumental		Vocal
Method	4-way NLL ↓	Brier ↓	ECE ↓	4-way NLL ↓	Brier ↓	4-way NLL ↓
BT	8.134 [7.58, 8.67]	0.365 [0.34, 0.40]	0.365	8.329 [7.70, 8.91]	0.386 [0.35, 0.42]	6.556 [5.05, 8.26]
AB-MNL	1.035 [0.99, 1.08]	0.213 [0.20, 0.23]	0.171	0.995 [0.95, 1.04]	0.219 [0.20, 0.24]	1.356 [1.21, 1.52]
GRK	**0.953** [0.91, 0.99]	**0.187** [0.17, 0.20]	**0.150**	**0.911** [0.87, 0.95]	**0.190** [0.18, 0.20]	**1.288** [1.18, 1.42]

Table 2: Per-class NLL breakdown on global test set. GRK's improvement is concentrated in BOTH_BAD prediction (23% reduction vs AB-MNL), validating the grounding mechanism. Best in **bold**.

Method	A-wins ↓	B-wins ↓	TIE ↓	BOTH_BAD ↓
BT	**0.425**	**0.432**	18.42	18.42
AB-MNL	0.614	0.639	2.631	1.397
GRK	0.690	0.716	**2.459**	**1.082**

4.2 Main Results

Table 1 presents the main results across three evaluation settings: global (all battles), instrumental-only, and vocal-only subsets.

GRK achieves the best performance across all metrics and settings. On the global test set, GRK achieves 7.9% lower 4-way NLL than AB-MNL (0.953 vs 1.035) and 12.4% lower BOTH_BAD Brier score (0.187 vs 0.213). The bootstrap 95% CI for the NLL difference excludes zero ([-0.096, -0.068]), confirming statistical significance. BT fails catastrophically on 4-way NLL (8.134) because it cannot model TIE or BOTH_BAD outcomes, assigning them uniform probability. The improvements are consistent across instrumental (8.4% NLL reduction) and vocal (5.0% NLL reduction) subsets.

4.3 Per-Class Analysis

Table 2 shows the per-class NLL breakdown, revealing where GRK's improvements originate.

GRK's improvement is concentrated in BOTH_BAD prediction: 1.082 vs 1.397 for AB-MNL, a 23% reduction. GRK also improves TIE prediction (2.459 vs 2.631). The trade-off is slightly worse A/B prediction (0.690/0.716 vs 0.614/0.639), but this is more than compensated by the gains on TIE and BOTH_BAD. BT achieves the best A/B prediction but fails completely on TIE/BOTH_BAD (NLL $\approx$ 18.4), confirming it cannot model these outcomes.

4.4 Ablation Studies

Table 3 validates that GRK's improvement stems from the grounding mechanism, not regularization effects.

Removing the grounding mechanism from GRK (replacing model-quality-dependent $P(\text{BOTH\_BAD})$ with a constant) significantly degrades performance: NLL increases from 0.953 to 1.156 (+0.203), with bootstrap 95% CI [+0.178, +0.229] excluding zero. This confirms that the structural coupling of BOTH_BAD to model quality is essential.

AB-MNL with varying L2 regularization on ρ never approaches GRK's performance. Even with no regularization (L2=0), AB-MNL achieves NLL of 1.026, still 0.073 worse than GRK. With strong regularization (L2$\to \infty$, pushing $\rho \to 0$), AB-MNL converges to NLL of 1.035. The persistent gap ($\geq$0.073) confirms that GRK's improvement is structural, not a regularization artifact.

Table 3: Ablation studies validating GRK's grounding mechanism. Removing grounding significantly degrades performance; AB-MNL regularization cannot recover GRK's behavior. Best in **bold**.

Variant	4-way NLL ↓	BOTH_BAD Brier ↓	Δ NLL vs GRK
GRK (full)	**0.953**	**0.187**	—
GRK (no grounding)	1.156	0.262	+0.203
AB-MNL (L2=0)	1.026	0.210	+0.073
AB-MNL (L2→ ∞)	1.035	0.213	+0.082

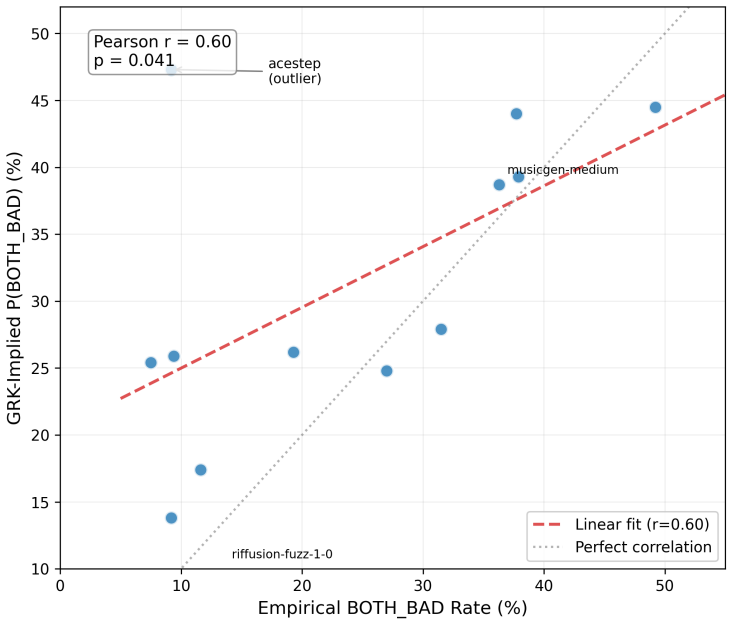

Figure 2: Correlation between GRK-implied acceptability and empirical BOTH_BAD rates across 12 text-to-music systems. Each point represents one system. Pearson $r = 0.60$ ($p = 0.041$) indicates meaningful correlation, validating that GRK's structural coupling produces actionable acceptability estimates.

4.5 Acceptability Validation

Figure 2 validates that GRK's implied acceptability signal is meaningful by comparing it to empirical BOTH_BAD rates.

For each system k, we compute the GRK-implied $P(\text{BOTH\_BAD}|k \text{ vs average})$ and compare it to the empirical BOTH_BAD rate (fraction of battles involving system k that resulted in BOTH_BAD). The Pearson correlation is $r = 0.60$ ($p = 0.041$), indicating a statistically significant positive relationship. Systems with lower GRK quality scores do indeed receive more BOTH_BAD votes, confirming that the grounding mechanism captures real patterns in human preferences. This enables quality-aware leaderboards that report not just relative rankings but also absolute acceptability estimates.

5 Conclusion

We proposed Grounded Rao-Kupper (GRK), a ranking model that treats BOTH_BAD as an outside option anchored to model quality. By coupling BOTH_BAD probability directly to latent quality scores, GRK converts this commonly ignored outcome into a statistical signal about absolute acceptability. On Music Arena, GRK achieves significant improvements over Bradley-Terry and decou-

pled baselines, with gains concentrated in BOTH_BAD prediction. The implied acceptability signal correlates meaningfully with empirical BOTH_BAD rates, enabling quality-aware leaderboards.

Limitations. The coupling assumption—that low-quality systems produce more BOTH_BAD outcomes—may not hold universally. Some systems may be comparatively weak but rarely produce unacceptable outputs, or vice versa.

Future Work. Applying GRK to multi-domain arenas (e.g., Chatbot Arena for LLMs, coding assistants) where BOTH_BAD patterns may differ could reveal domain-specific insights about acceptability.

REFERENCES

A. Agostinelli, Timo I. Denk, Zalán Borsos, Jesse Engel, Mauro Verzetti, Antoine Caillon, Qingqing Huang, A. Jansen, Adam Roberts, M. Tagliasacchi, Matthew Sharifi, Neil Zeghidour, and C. Frank. Musiclm: Generating music from text. *ArXiv*, abs/2301.11325, 2023.

S. Ameli, Siyuan Zhuang, Ion Stoica, and Michael W. Mahoney. A statistical framework for ranking llm-based chatbots. *ArXiv*, abs/2412.18407, 2024.

Ralph Allan Bradley and Milton E. Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3-4):324–345, 1952. doi: 10.1093/biomet/39.3-4.324.

Ivi Chatzi, Eleni Straitouri, Suhas Thejaswi, and M. Rodriguez. Prediction-powered ranking of large language models. *ArXiv*, abs/2402.17826, 2024.

Jinghong Chen, Guangyu Yang, Weizhe Lin, Jingbiao Mei, and Bill Byrne. On extending direct preference optimization to accommodate ties. *ArXiv*, abs/2409.17431, 2024.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael I. Jordan, Joseph Gonzalez, and Ion Stoica. Chatbot arena: An open platform for evaluating llms by human preference. pp. 8359–8388, 2024.

Jade Copet, F. Kreuk, Itai Gat, Tal Remez, David Kant, Gabriel Synnaeve, Yossi Adi, and Alexandre D'efossez. Simple and controllable music generation. *ArXiv*, abs/2306.05284, 2023.

Roger R. Davidson. On extending the bradley-terry model to accommodate ties in paired comparison experiments. *Journal of the American Statistical Association*, 65(329):317–328, 1970. doi: 10.1080/01621459.1970.10481082.

Zach Evans, Julian Parker, CJ Carr, Zack Zukowski, Josiah Taylor, and Jordi Pons. Stable audio open. *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, 2024.

Evan Frick, Connor Chen, Joseph Tennyson, Tianle Li, Wei-Lin Chiang, A. Angelopoulos, and Ion Stoica. Prompt-to-leaderboard. *ArXiv*, abs/2502.14855, 2025.

Kevin Kilgour, Mauricio Zuluaga, Dominik Roblek, and Matthew Sharifi. Fréchet audio distance: A metric for evaluating music enhancement algorithms. *ArXiv*, abs/1812.08466, 2018.

Yonghyun Kim, Wayne Chi, Anastasios N. Angelopoulos, Wei-Lin Chiang, Koichi Saito, Shinji Watanabe, Yuki Mitsufuji, and Chris Donahue. Music arena: Live evaluation for text-to-music, 2025. URL https://arxiv.org/abs/2507.20900.

Haohe Liu, Zehua Chen, Yiitan Yuan, Xinhao Mei, Xubo Liu, Danilo P. Mandic, Wenwu Wang, and Mark D. Plumbley. Audioldm: Text-to-audio generation with latent diffusion models. pp. 21450–21474, 2023.

Zirui Liu, Jiatong Li, Zhuang Yan, Qi Liu, Shuanghong Shen, Ouyang Jie, Mingyue Cheng, and Shijin Wang. am-elo: A stable framework for arena-based llm evaluation. *ArXiv*, abs/2505.03475, 2025.

R. Duncan Luce. *Individual Choice Behavior: A Theoretical Analysis*. John Wiley & Sons, 1959.

Robin L. Plackett. The analysis of permutations. *Journal of the Royal Statistical Society: Series B*, 37(2):195–202, 1975. doi: 10.1111/j.2517-6161.1975.tb00933.x.

P. V. Rao and Lawrence L. Kupper. Ties in paired-comparison experiments: A generalization of the bradley-terry model. *Journal of the American Statistical Association*, 62(317):194–204, 1967. doi: 10.1080/01621459.1967.10482901.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, E. Xing, Haotong Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena. *ArXiv*, abs/2306.05685, 2023.