Novelty Assessment Report

Paper: Dropping Just a Handful of Preferences Can Change Top Large Language Model Rankings
PDF URL: https://openreview.net/pdf?id=jNiEMDsRgc
Venue: ICLR 2026 Conference Submission
Year: 2026
Report Generated: 2025-12-27

Abstract

We propose a method for evaluating the robustness of widely used LLM ranking systems---variants of a Bradley--Terry model---to dropping a worst-case very small fraction of preference data. Our approach is computationally fast and easy to adopt. When we apply our method to matchups from popular LLM ranking platforms, including Chatbot Arena and derivatives, we find that the rankings of top-performing models can be remarkably sensitive to the removal of a small fraction of preferences; for instance, dropping just 0.003% of human preferences can change the top-ranked model on Chatbot Arena. Our robustness check identifies the specific preferences most responsible for such ranking flips, allowing for inspection of these influential preferences. We observe that the rankings derived from MT-bench preferences are notably more robust than those from Chatbot Arena, likely due to MT-bench's use of expert annotators and carefully constructed prompts. Finally, we find that neither rankings based on crowdsourced human evaluations nor those based on LLM-as-a-judge preferences are systematically more sensitive than the other.

Disclaimer

This report is **AI-GENERATED** using Large Language Models and WisPaper (a scholar search engine). It analyzes academic papers' tasks and contributions against retrieved prior work. While this system identifies **POTENTIAL** overlaps and novel directions, **ITS COVERAGE IS NOT EXHAUSTIVE AND JUDGMENTS ARE APPROXIMATE**. These results are intended to assist human reviewers and **SHOULD NOT** be relied upon as a definitive verdict on novelty.

Note that some papers exist in multiple, slightly different versions (e.g., with different titles or URLs). The system may retrieve several versions of the same underlying work. The current automated pipeline does not reliably align or distinguish these cases, so human reviewers will need to disambiguate them manually.

If you have any questions, please contact: mingzhang23@m.fudan.edu.cn

Core Task Landscape

This paper addresses: **Evaluating Robustness of LLM Ranking Systems to Data Dropping**

A total of **10 papers** were analyzed and organized into a taxonomy with **7 categories**.

Taxonomy Overview

The research landscape has been organized into the following main categories:

- **Ranking System Robustness and Sensitivity Analysis**
- **Data Removal and Modification Techniques**
- **LLM Evaluation Frameworks and Methodologies**

Complete Taxonomy Tree

- Evaluating Robustness of LLM Ranking Systems to Data Dropping Survey Taxonomy
- Ranking System Robustness and Sensitivity Analysis
  - Preference-Based Ranking Robustness ★ (2 papers)
  - [0] Dropping Just a Handful of Preferences Can Change Top Large Language Model Rankings (Anon et al., 2026) View paper
  - [10] CURATRON: Complete and Robust Preference Data for Rigorous Alignment of Large Language Models (Son The Nguyen, 2024) View paper
  - Benchmark Stability Under Missing Data (2 papers)
  - [5] Towards more robust nlp system evaluation: Handling missing scores in benchmarks (Anas Himmi, 2024) View paper
  - [7] Bridging Context, Statistics, and Practice: A Multi-Dimensional Framework for Responsible LLM Evaluation and Selection (Shiva kumar Ramavath, 2025) View paper
  - Multi-Criteria Task Prioritization with Incomplete Data (1 papers)
  - [8] Mathematical model of stable task prioritization with dynamically adjustable criteria weights (Trushin, 2025) View paper
- Data Removal and Modification Techniques
  - Safety-Preserving Data Filtering (1 papers)
  - [1] Layer-aware representation filtering: Purifying finetuning data to preserve llm safety alignment (Li Hao, 2025) View paper
  - Machine Unlearning for LLMs (3 papers)
  - [2] Exact and efficient unlearning for large language model-based recommendation (Zhiyu Hu, 2025) View paper
  - [3] A survey on large language models unlearning: taxonomy, evaluations, and future directions (Le-Khac, 2025) View paper
  - [9] AdapterSwap: Continuous Training of LLMs with Data Removal and Access-Control Guarantees (Fleshman, 2024) View paper
- LLM Evaluation Frameworks and Methodologies
  - Open-Ended Capability Assessment (1 papers)
  - [4] ONEBench to Test Them All: Sample-Level Benchmarking Over Open-Ended Capabilities (Albanie, 2024) View paper
  - LLM-Based Ranking and Re-ranking Systems (1 papers)
  - [6] Make large language model a better ranker (Zhi Zheng, 2024) View paper

Narrative

Core task: Evaluating robustness of LLM ranking systems to data dropping. The field examines how ranking systems that compare or order large language models respond when portions of evaluation data are removed or modified. The taxonomy organizes this area into three main branches. The first, Ranking System Robustness and Sensitivity Analysis, investigates how stable rankings remain under perturbations, including preference-based methods that assess consistency when human or model preferences are altered. The second branch, Data Removal and Modification Techniques, encompasses approaches for systematically dropping, unlearning, or filtering data—ranging from exact unlearning methods like Exact Efficient Unlearning[2] to representation-level interventions such as Layer-aware Representation Filtering[1]. The third branch, LLM Evaluation Frameworks and Methodologies, addresses broader assessment

paradigms, including multi-dimensional evaluation schemes like Multi-Dimensional LLM Evaluation[7] and benchmarks such as ONEBench[4] that test model capabilities under varied conditions.

Several active lines of work explore trade-offs between evaluation completeness and robustness. One thread examines how missing or incomplete scores affect ranking validity, as seen in Handling Missing Scores[5], while another investigates whether LLMs themselves can serve as reliable rankers, exemplified by LLM Better Ranker[6]. A related concern is maintaining stable task prioritization when data availability fluctuates, addressed by Stable Task Prioritization[8]. Within this landscape, Dropping Preferences Rankings[0] sits squarely in the preference-based robustness cluster, focusing on how rankings derived from pairwise or preference judgments degrade when subsets of preferences are systematically dropped. This emphasis contrasts with nearby work like CURATRON[10], which targets data curation for training rather than post-hoc ranking stability, and differs from unlearning-focused studies like LLM Unlearning Survey[3] that prioritize removing specific knowledge rather than testing ranking resilience.

Related Works in Same Category

The following **1 sibling papers** share the same taxonomy leaf node with the original paper:

1. CURATRON: Complete and Robust Preference Data for Rigorous Alignment of Large Language Models

Authors: Son The Nguyen, Niranjan Uma Naresh, Theja Tulabandhula | **Year/Venue**: 2024 • DASH | **URL**: View paper

Abstract

This paper addresses the challenges of aligning large language models (LLMs) with human values via preference learning (PL), focusing on incomplete and corrupted data in preference datasets. We propose a novel method for robustly and completely recalibrating values within these datasets to enhance LLMsâ resilience against the issues. In particular, we devise a guaranteed polynomial time ranking algorithm that robustifies several existing models, such as the classic Bradleyâ Terryâ Luce (BTL)...

Relationship Analysis

Both papers belong to the Preference-Based Ranking Robustness category, examining the stability of Bradley-Terry model rankings when preference data is compromised. The original paper focuses on evaluating the sensitivity of existing LLM ranking systems (like Chatbot Arena) to worst-case dropping of small fractions of human/AI preferences, identifying which specific preferences cause ranking flips. In contrast, the candidate paper (CURATRON) proposes algorithmic solutions for robustly recovering rankings from preference data that is both incomplete and adversarially corrupted, providing theoretical guarantees for rank recovery under specific conditions rather than auditing existing systems.

Contributions Analysis

Overall novelty summary. The paper introduces a computationally efficient method for assessing how Bradley-Terry-based LLM ranking systems respond to worst-case removal of small preference data fractions. It resides in the 'Preference-Based Ranking Robustness' leaf, which contains only two papers total, indicating a relatively sparse research direction within the broader taxonomy of ten papers across seven leaf nodes. This positioning suggests the work addresses a focused problem—stability of preference-driven rankings under adversarial data dropping—that has received limited prior attention compared to adjacent areas like general benchmark evaluation or machine unlearning.

The taxonomy reveals neighboring research directions that contextualize this contribution. The sibling leaf 'Benchmark Stability Under Missing Data' examines ranking stability when scores are absent but does not focus on preference-based models or adversarial dropping scenarios. Adjacent branches include 'Data Removal and Modification Techniques,' which emphasizes unlearning and safety filtering rather than ranking sensitivity analysis, and 'LLM Evaluation Frameworks,' which addresses broader assessment paradigms. The paper's scope_note explicitly excludes general evaluation frameworks and data modification methods, positioning it at the intersection of ranking theory and adversarial robustness rather than comprehensive evaluation design or training-time data curation.

Among fourteen candidates examined through limited semantic search, no contributions were clearly refuted by prior work. The core methodological contribution—evaluating worst-case robustness to data dropping—examined ten candidates with zero refutable overlaps. The identification of specific influential preferences examined three candidates, again with no refutations, while empirical findings on platform sensitivity examined one candidate without overlap. These statistics suggest that within the examined scope, the approach and findings appear distinct from existing literature, though the limited search scale (fourteen total candidates) means comprehensive coverage of all related work cannot be claimed.

Based on the available signals, the work appears to occupy a relatively underexplored niche within LLM evaluation research. The sparse taxonomy leaf and absence of refutable prior work among examined candidates suggest novelty, though the limited search scope (top-K semantic matches plus citations) means adjacent or parallel efforts outside this sample remain possible. The focus on adversarial data dropping for preference-based rankings distinguishes it from broader robustness studies that address missing data or general perturbations without targeting worst-case scenarios.

This paper presents **3 main contributions**, each analyzed against relevant prior work:

Contribution 1: Method for evaluating robustness of LLM ranking systems to worst-case data dropping

Description: The authors develop a computationally efficient algorithm that assesses whether LLM leaderboard rankings remain stable when a small fraction of preference data is removed. The method extends approximate maximum influence perturbation techniques to identify influential preferences and verify ranking changes without exhaustive combinatorial search.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. Strong Preferences Affect the Robustness of Preference Models and Value Alignment

URL: View paper

Brief Assessment

Strong Preferences Robustness[19] focuses on theoretical sensitivity analysis of preference model probabilities (Bradley-Terry, Plackett-Luce) to changes in other preference probabilities, not on evaluating ranking system robustness to data removal. The candidate examines how preference probabilities mathematically depend on each other, while the original develops a computational algorithm (AMIP-based) to identify influential data points that change rankings.

---

2. An empirical evaluation of deep semi-supervised learning

URL: View paper

Brief Assessment

Semi-Supervised Learning Evaluation[16] focuses on evaluating deep semi-supervised learning algorithms for classification tasks across different datatypes. It does not address Bradley-Terry ranking models, LLM leaderboards, or robustness to data perturbations in preference-based systems.

3. Preference-Based Dynamic Ranking Structure Recognition

URL: View paper

Brief Assessment

Dynamic Ranking Recognition[23] focuses on detecting structural changes in ranking groups over time using temporal penalties and dynamic programming, not on robustness evaluation of Bradley-Terry models to data perturbations or worst-case data dropping.

4. AdvO-RAN: Adversarial Deep Reinforcement Learning in AI-Driven Open Radio Access Networks

URL: View paper

Brief Assessment

AdvO-RAN[20] focuses on adversarial deep reinforcement learning in Open Radio Access Networks, not on evaluating robustness of ranking systems or Bradley-Terry models to data perturbations. The candidate's mention of Bradley-Terry models appears in a different context (predictor definition) unrelated to robustness evaluation of ranking systems.

5. Fusing Rewards and Preferences in Reinforcement Learning

URL: View paper

Brief Assessment

Fusing Rewards Preferences[18] focuses on reinforcement learning algorithms that combine rewards and preferences for policy optimization in control tasks. It does not address robustness evaluation of ranking systems or Bradley-Terry model stability under data perturbations.

6. Generalized Bradley-Terry Models and Multi-Class Probability Estimates.

URL: View paper

Brief Assessment

Generalized Bradley-Terry[22] focuses on extending the Bradley-Terry model for multi-class probability estimates from pairwise comparisons, not on robustness evaluation of ranking systems to data perturbations or worst-case data dropping scenarios.

7. Learning to Rank Chain-of-Thought: An Energy-Based Approach with Outcome Supervision

URL: View paper

Brief Assessment

Ranking Chain-of-Thought[17] focuses on training a lightweight energy-based model to rank and select the best chain-of-thought reasoning paths from LLM outputs using outcome supervision. It does not address robustness evaluation of Bradley-Terry ranking models to data perturbations or worst-case data dropping scenarios.

8. LLM-Driven Active Listwise Tournaments for Portfolio Selection in Large Asset Universes

URL: View paper

Brief Assessment

Active Listwise Tournaments[21] focuses on portfolio selection using listwise queries in financial asset universes, not on evaluating robustness of Bradley-Terry ranking models to data perturbations or worst-case data dropping.

9. Provably Robust DPO: Aligning Language Models with Noisy Feedback

URL: View paper

Brief Assessment

Provably Robust DPO[15] addresses robustness of policy learning under noisy preference labels in DPO training, not robustness evaluation of Bradley-Terry ranking systems to data removal. The original paper evaluates leaderboard stability; the candidate develops noise-robust training algorithms.

10. Minimax Hypothesis Testing for the Bradley-Terry-Luce Model

URL: View paper

Brief Assessment

Bradley-Terry-Luce Testing[14] focuses on hypothesis testing to determine whether pairwise comparison data conforms to a BTL model, not on evaluating robustness of existing rankings to data removal. The original paper addresses stability of LLM leaderboard rankings when preferences are dropped, while the candidate develops statistical tests for model validity.

Contribution 2: Identification of specific preferences responsible for ranking flips

Description: The method not only detects non-robustness but also pinpoints the exact preference data points (prompt-response pairs) that drive changes in model rankings, enabling qualitative inspection of these influential evaluations.

This contribution was assessed against **3 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

1. Identifying Influential Nodes Using a Shell-Based Ranking and Filtering Method in Social Networks.

URL: View paper

Brief Assessment

Shell-Based Ranking Filtering[13] focuses on identifying influential nodes in social networks using shell decomposition methods, not on identifying influential data points in preference-based ranking systems or LLM evaluation.

2. Detect influential points of feature rankings.

URL: View paper

Brief Assessment

The candidate paper's full text is not available (marked as 'n/a'), making it impossible to assess whether it addresses identifying influential data points that cause ranking changes in preference models or any related methodology.

3. Towards Robust Alignment of Language Models: Distributionally Robustifying Direct Preference Optimization

URL: View paper

Brief Assessment

Distributionally Robust DPO[11] focuses on enhancing DPO's resilience to noise in training datasets through distributionally robust optimization, not on identifying specific preference data points that cause ranking changes in evaluation systems.

Contribution 3: Empirical findings on sensitivity of popular LLM evaluation platforms

Description: The authors apply their robustness check to multiple LLM arenas and discover that top model rankings are surprisingly fragile, with extremely small fractions of data (as low as 0.003%) sufficient to alter rankings. They also find that MT-bench is notably more robust due to expert annotators and carefully designed prompts.

This contribution was assessed against **1 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. ROPO: Robust Preference Optimization for Large Language Models

URL: View paper

Brief Assessment

ROPO[24] focuses on preference alignment robustness to noisy training data in RLHF, not on the sensitivity of evaluation platforms or leaderboards to data removal. The candidate addresses noise-tolerant training methods, while the original contribution concerns the fragility of ranking systems.

Appendix: Text Similarity Detection

No high-similarity text segments were detected across any compared papers.

References

- [0] Dropping Just a Handful of Preferences Can Change Top Large Language Model Rankings View paper
- [1] Layer-aware representation filtering: Purifying finetuning data to preserve llm safety alignment View paper
- [2] Exact and efficient unlearning for large language model-based recommendation View paper
- [3] A survey on large language models unlearning: taxonomy, evaluations, and future directions View paper
- [4] ONEBench to Test Them All: Sample-Level Benchmarking Over Open-Ended Capabilities View paper
- [5] Towards more robust nlp system evaluation: Handling missing scores in benchmarks View paper
- [6] Make large language model a better ranker View paper
- [7] Bridging Context, Statistics, and Practice: A Multi-Dimensional Framework for Responsible LLM Evaluation and Selection View paper
- [8] Mathematical model of stable task prioritization with dynamically adjustable criteria weights View paper
- [9] AdapterSwap: Continuous Training of LLMs with Data Removal and Access-Control Guarantees View paper
- [10] CURATRON: Complete and Robust Preference Data for Rigorous Alignment of Large Language Models View paper
- [11] Towards Robust Alignment of Language Models: Distributionally Robustifying Direct Preference Optimization View paper
- [12] Detect influential points of feature rankings. View paper
- [13] Identifying Influential Nodes Using a Shell-Based Ranking and Filtering Method in Social Networks. View paper
- [14] Minimax Hypothesis Testing for the Bradley-Terry-Luce Model View paper
- [15] Provably Robust DPO: Aligning Language Models with Noisy Feedback View paper
- [16] An empirical evaluation of deep semi-supervised learning View paper
- [17] Learning to Rank Chain-of-Thought: An Energy-Based Approach with Outcome Supervision View paper
- [18] Fusing Rewards and Preferences in Reinforcement Learning View paper
- [19] Strong Preferences Affect the Robustness of Preference Models and Value Alignment View paper
- [20] AdvO-RAN: Adversarial Deep Reinforcement Learning in AI-Driven Open Radio Access Networks View paper
- [21] LLM-Driven Active Listwise Tournaments for Portfolio Selection in Large Asset Universes View paper
- [22] Generalized Bradley-Terry Models and Multi-Class Probability Estimates. View paper
- [23] Preference-Based Dynamic Ranking Structure Recognition View paper
- [24] ROPO: Robust Preference Optimization for Large Language Models View paper