Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Applied Mathematical Robustness Analysis of Maximum-Likelihood Pairwise Ranking for Comparison-Driven Intelligent Systems

Junyi Yao

j.yao@wustl.edu

Washington University in St. Louis

Zihao Zheng

Washington University in St. Louis

Jiayu Long

Washington University in St. Louis

Research Article

Keywords: pairwise ranking, maximum-likelihood estimation, robustness analysis, structured perturbation, computational science, intelligent systems

Posted Date: August 20th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10083033/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Applied Mathematical Robustness Analysis of Maximum-Likelihood Pairwise Ranking for Comparison-Driven Intelligent Systems

Junyi Yao[1*], Zihao Zheng[1] and Jiayu Long[1]

[1*]Washington University in St. Louis, St. Louis, Missouri, USA.

*Corresponding author(s). E-mail(s): j.yao@wustl.edu;
Contributing authors: z.zihaogary@wustl.edu;
jiayujacqueline@wustl.edu;

Abstract

Pairwise ranking is a widely used computational primitive in preference aggregation, recommendation, reputation estimation, and engineering decision support. Yet the robustness of maximum-likelihood ranking estimators under coordinated perturbations of comparison data remains insufficiently characterized. This paper studies pairwise ranking robustness as an applied mathematical and computational science problem for comparison-driven intelligent systems. We formulate coordinated perturbation as a budget-constrained subset-selection problem over observed pairwise comparisons and use Adaptive Subset Selection Attack (ASSA) as a scalable heuristic for probing the perturbation landscape. Using synthetic and observed preference datasets, we evaluate how structured perturbations alter output rankings across low-budget and regime-sensitive settings. The results show that fragility is strongly data dependent: small but coordinated perturbations can produce meaningful target-rank shifts in some comparison structures, while similarly sized perturbations have limited effect in others. Repeated-trial comparisons with random and greedy baselines further indicate that the dominant empirical signal is regime-dependent ranking sensitivity rather than a universal dominance relation among perturbation heuristics. Taken together, the findings position robustness analysis of pairwise ranking as a useful bridge between applied mathematics, computational science, and trustworthy machine learning for engineering-facing ranking systems.

Keywords: pairwise ranking, maximum-likelihood estimation, robustness analysis, structured perturbation, computational science, intelligent systems

1 Introduction

Pairwise ranking is a standard computational mechanism for converting relative comparisons into global orderings. It appears in recommendation, preference aggregation, reputation estimation, and comparison-driven engineering decision support, where systems often infer latent strength from relational observations rather than from direct cardinal labels. Maximum-likelihood estimators based on Bradley–Terry-style models remain especially attractive because they offer a principled statistical framework, interpretable latent parameters, and efficient optimization procedures Bradley and Terry (1952); Hunter (2004). As a result, they serve as an important modeling layer in many comparison-driven intelligent systems.

From an applied mathematics perspective, however, reliability of the inferred ranking is at least as important as nominal estimation quality. Comparison data are often sparse, noisy, and structurally heterogeneous. In many deployed systems, a relatively small number of altered pairwise outcomes may lead to a disproportionate shift in the final ordering. This creates a robustness problem that is both statistical and combinatorial: the estimator depends on the likelihood geometry of the underlying model, while the perturbation process depends on which subset of observed comparisons is modified. In practical terms, this means that system-level reliability cannot be inferred from estimation accuracy alone.

This setting differs in useful ways from conventional adversarial learning. In supervised learning, perturbations are frequently studied in continuous feature spaces and evaluated with respect to misclassification. In pairwise ranking, the perturbation space is discrete, budget constrained, and graph structured, while the output is a global ordering rather than a class label. As a result, the analysis naturally belongs to the intersection of applied mathematics, computational science, and trustworthy intelligent systems Wang et al. (2019); Liu and Moitra (2021). It also motivates a computational-science perspective in which perturbation sensitivity is treated as a property of the full inference pipeline rather than of the estimator in isolation. This broader view is consistent with recent work arguing that reliable evaluation of intelligent systems requires more than headline task performance, including diagnostic separation of distinct behavioral dimensions and clearer attention to reproducibility assumptions Yao et al. (2026a); Yao and Zheng (2026b).

This paper studies robustness of maximum-likelihood pairwise ranking under structured comparison perturbations. We formulate coordinated perturbation as a budget-constrained subset-selection problem and use Adaptive Subset Selection Attack (ASSA) as a scalable search heuristic for probing the perturbation landscape. The goal is not to claim that one heuristic always dominates all baselines, but to identify when ranking estimators become fragile, how strongly low-budget perturbations alter output orderings, and what these behaviors imply for intelligent systems that rely on comparison-driven inference.

The main contributions of this study are as follows:

1. We formulate robustness analysis for maximum-likelihood pairwise ranking as a budget-constrained subset-selection problem over observed comparison data.

2. We present ASSA as a scalable perturbation-search heuristic for stress testing structured sensitivity in ranking systems.
3. We provide a compact empirical framework showing that ranking fragility is regime dependent and data dependent, with direct implications for engineering-facing decision systems built on pairwise supervision.

2 Related work

2.1 Pairwise ranking and preference learning

Pairwise comparison models infer global orderings from incomplete or relative observations. Bradley–Terry-type methods are especially influential because they estimate latent strengths through a statistically grounded likelihood formulation and support efficient iterative optimization Bradley and Terry (1952); Hunter (2004). These methods are widely used in social choice, ranking aggregation, recommendation, and preference learning Zhao and Xia (2020).

2.2 Robustness in learning-based ranking

Robustness is a central requirement for trustworthy machine learning. Much of the literature focuses on adversarial feature perturbations or poisoning in supervised prediction. Pairwise ranking introduces a different mathematical object: perturbations act on relational observations, the feasible set is combinatorial, and robustness must be measured with respect to shifts in the inferred ordering rather than class labels Wang et al. (2019).

2.3 Computational reliability in engineering-facing systems

Comparison-driven ranking is often embedded in larger engineering and computational pipelines, such as decision support, evaluation aggregation, or automated prioritization. In these settings, reliability depends not only on the estimator itself but also on how sensitive the full inference process is to structured irregularities in the comparison graph. This makes robustness analysis relevant to dependable intelligent systems, computational reliability assessment, and engineering-oriented model validation Liu and Moitra (2021). Related recent work in other intelligent-system domains has likewise emphasized that meaningful evaluation requires diagnostic metrics and explicit reporting of execution assumptions rather than aggregate performance alone Yao et al. (2026a); Yao and Zheng (2026b). More broadly, recent AI-systems research also highlights that deployment quality depends jointly on robustness, efficiency, and adaptive resource allocation across training and inference pipelines Zhang et al. (2025a,c,b).

3 Problem setting

3.1 Maximum-likelihood pairwise ranking

We consider m candidates and a set of observed pairwise preferences D. Each record indicates that candidate c_i is preferred to candidate c_j. Under a Bradley–Terry-style model, each candidate is assigned a latent strength parameter $p_i > 0$, and the probability of a preference outcome is

$$P(c_i \succ c_j) = \frac{p_i}{p_i + p_j}. \tag{1}$$

The maximum-likelihood estimator computes $\mathbf{p}^*$ from the observed comparison set D, and the final ranking $\tau(\mathbf{p}^*)$ is obtained by sorting candidates according to their estimated strengths.

3.2 Structured perturbation objective

We study a coordinated perturbation process in which a budget-limited subset of pairwise comparisons is modified to move the final ranking toward a target ordering. Let $\Delta \subset D$ denote the set of perturbed comparisons, and let $D' = D \oplus \Delta$ be the perturbed dataset. The resulting ranking is $\tau(\mathbf{p}^*(D'))$.

The perturbation objective is formulated as

$$\min_{\Delta} \quad K_d\big(\tau_t, \tau(\mathbf{p}^*(D \oplus \Delta))\big), \tag{2}$$

subject to a perturbation budget and any integrity constraints imposed by the threat model. Here $K_d(\cdot,\cdot)$ denotes a Kendall tau style ranking distance and τ_t is the target ranking.

3.3 Robustness interpretation

This formulation treats perturbation as a robustness diagnostic for learning-based ranking. The perturbation acts on a discrete subset of observed relations rather than on continuous features. Consequently, ranking fragility depends jointly on the statistical estimator, the topology of the comparison structure, and the search strategy used to locate influential perturbation subsets.

4 Perturbation search methods

4.1 Baseline strategies

We compare ASSA with three baseline perturbation strategies.

Random Flip randomly perturbs comparisons within the allowed budget.

Greedy Flip iteratively selects the perturbation that yields the largest immediate reduction in ranking distance.

Randomized Subset Attack (RSA) samples candidate subsets and retains those that improve the perturbation objective.

4.2 Adaptive Subset Selection Attack

ASSA is designed as a scalable heuristic for identifying high-impact perturbation subsets more efficiently than naive random exploration. At a high level, ASSA partitions candidate perturbations into subsets, evaluates their impact on the target objective, and adaptively retains the subsets that contribute useful improvement. The repeated shrink-and-refine process concentrates computation on structurally influential regions of the perturbation space.

In this paper, ASSA is used primarily as a stress-testing heuristic. Its main value is to expose fragility regimes in maximum-likelihood ranking systems, including regimes where several different search procedures reveal comparable instability.

4.3 Threat model

The experiments are interpreted as a white-box or near-white-box robustness analysis. This yields an upper-bound view of what a strategically informed perturbing agent could achieve under favorable knowledge assumptions. The conclusions are therefore about estimator fragility under structured pressure, not direct forecasts of deployment-time attack success in every real setting.

5 Experimental design

5.1 Datasets

The study uses both synthetic and observed preference datasets. Synthetic data provide controlled comparison structures in which candidate count, sparsity, and budget effects can be varied systematically. The observed dataset provides a practical view of whether the same robustness patterns appear outside stylized simulations. This mixed design allows the paper to connect mathematical stress testing with application-oriented robustness interpretation.

5.2 Evaluation metrics

The primary evaluation quantities are target-rank shift, improvement in Kendall tau style distance, and perturbation success indicators where appropriate. These metrics quantify how strongly constrained changes in the comparison set alter the final ordering.

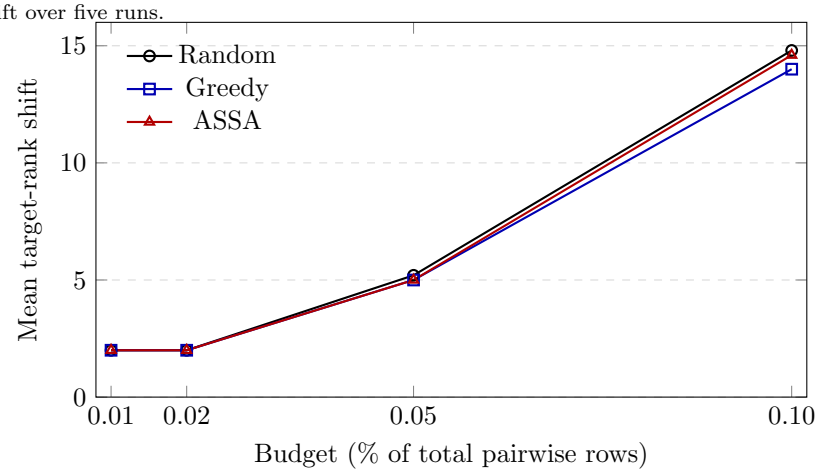
5.3 Repeated trials and budget regimes

Because several perturbation strategies are stochastic, the evaluation reports repeated-trial summaries rather than single-run outcomes. Low-budget perturbations are emphasized because they are more practically relevant for robustness auditing. The current study therefore focuses on a conservative budget grid of 0.01%, 0.02%, 0.05%, and 0.10% of all pairwise rows.

Table 1 Cross-dataset comparison of low-budget behavior. Values report mean target-rank shift over five runs.

Dataset	0.01%	0.02%	0.05%	0.10%
Local 100 (Greedy)	2.0	2.0	5.0	14.0
Local 100 (ASSA)	2.0	2.0	5.0	14.6
Synthetic BT-30 (Greedy)	0.0	0.0	0.0	1.0
Synthetic BT-30 (ASSA)	0.0	0.0	0.0	1.0

Fig. 1 Low-budget response on the local 100-candidate dataset. The plotted value is mean target-rank shift over five runs.



6 Results and discussion

6.1 Cross-dataset low-budget comparison

Table 1 shows that low-budget sensitivity is not universal. On the local 100-candidate dataset, both Greedy and ASSA induce substantial target-rank movement under modest budgets. On the synthetic Bradley–Terry-style dataset, the same budget grid produces almost no movement until the largest tested budget. This suggests that robustness is shaped strongly by data structure and comparison geometry rather than by the estimator alone.

6.2 Low-budget response on the local dataset

Figure 1 and Table 2 show a gradual increase in target-rank shift as the perturbation budget grows. On this dataset, the smallest budgets yield nearly indistinguishable outcomes across methods, while the larger tested budgets reveal much stronger sensitivity. At 0.05% and 0.10%, ASSA and Greedy both uncover substantial fragility, but the gap between them remains small relative to the overall ranking movement.

This result is useful for the paper's main claim. The dominant empirical signal is not that ASSA always wins by a large margin, but that maximum-likelihood ranking

Table 2 Low-budget pilot on the local 100-candidate dataset. Values report mean target-rank shift over five runs, with 95% confidence intervals in parentheses where nonzero.

Budget (%)	Random	Greedy	ASSA
0.01	2.0	2.0	2.0
0.02	2.0	2.0	2.0
0.05	5.2 (0.39)	5.0	5.0
0.10	14.8 (0.39)	14.0	14.6 (0.48)

can become sensitive under very small, structured changes in observed comparisons. In computational science terms, the perturbation landscape appears broad enough that multiple search strategies can reveal comparable instability once the system enters a fragile regime.

6.3 Robustness interpretation

Taken together, the results support a regime-dependent view of ranking robustness. In some settings, small coordinated perturbations already produce meaningful output shifts, while in others similarly sized perturbations have limited effect. This makes robustness an empirical property that should be measured alongside estimation accuracy when pairwise ranking is used in engineering-facing decision systems.

The experiments also show a computational tradeoff. Random perturbation is effectively instantaneous, Greedy completes in milliseconds per run on the reported local setting, and ASSA requires about one second per run on the same budget grid. ASSA therefore remains useful as a stress-testing heuristic, but claims about effectiveness should be interpreted jointly with computational cost.

6.4 Implications for comparison-driven intelligent systems

The results have practical implications for systems that rely on inferred orderings as an intermediate computational layer. When rankings are used to support prioritization, recommendation, or automated decision flow, even modest sensitivity to structured comparison perturbations can propagate downstream and alter system behavior. For this reason, robustness auditing should be viewed as part of model validation rather than as a purely adversarial add-on. In particular, low-budget perturbation analysis can serve as a compact diagnostic for determining whether a ranking pipeline is sufficiently stable for engineering-facing deployment.

7 Limitations

Several limitations bound the scope of the present study. First, the analysis focuses on maximum-likelihood pairwise ranking and does not automatically extend to every ranking architecture used in intelligent systems. Second, much of the current evidence relies on synthetic structure and a strong-information perturbation model. Third,

ASSA is a heuristic rather than an exact combinatorial optimizer, so the reported perturbations should be interpreted as informative robustness diagnostics rather than globally optimal certificates.

8 Conclusion

This paper studied robustness of maximum-likelihood pairwise ranking under structured comparison perturbations and formulated the problem as a budget-constrained subset-selection task over relational observations. The resulting perspective connects likelihood-based ranking with applied mathematical sensitivity analysis and computational robustness assessment. Across synthetic and observed preference settings, the evidence indicates that pairwise ranking can exhibit meaningful regime-dependent fragility under small coordinated perturbations. These findings support the use of structured perturbation analysis as a practical tool for evaluating ranking reliability in machine learning and comparison-driven intelligent systems.

Declarations

Funding No external funding was received for this study.

Competing interests The authors declare that they have no competing interests.

Author contributions Junyi Yao led the conceptualization, writing, and project coordination. Zihao Zheng contributed to methodology development, experiment design, and technical revision. Jiayu Long contributed to analysis framing, interpretation of results, and manuscript refinement. All authors reviewed and approved the final manuscript.

Data availability The processed data analyzed in this study are available in the project materials associated with this manuscript. Additional details on dataset construction and preprocessing are described in the manuscript, and supporting materials can be provided by the corresponding author upon reasonable request.

Code availability The scripts used for preprocessing, experiment execution, result summarization, and figure generation are available in the project materials associated with this manuscript. An archival code package can be provided by the corresponding author upon reasonable request.

References

Bradley RA, Terry ME (1952) Rank analysis of incomplete block designs: I. The method of paired comparisons. *Biometrika* 39(3–4):324–345. https://doi.org/10.2307/2334029

Hunter DR (2004) MM algorithms for generalized Bradley–Terry models. *The Annals of Statistics* 32(1):384–406. https://doi.org/10.1214/aos/1079120141

Wang Z, Xu Q, Ma K, Jiang Y, Cao X, Huang Q (2019) Adversarial preference learning with pairwise comparisons. In: *Proceedings of the ACM International Conference on Multimedia*, pp 656–664. https://doi.org/10.1145/3343031.3350908

Zhao Z, Xia L (2020) Learning mixtures of Plackett–Luce models with features from top-ℓ orders. arXiv preprint arXiv:2006.03869. https://arxiv.org/abs/2006.03869

Liu A, Moitra A (2021) Robust voting rules from algorithmic robust statistics. arXiv preprint arXiv:2112.06380. https://arxiv.org/abs/2112.06380

Yao J, Zheng Z, Li B (2026a) Measuring whether LLM tutors teach or solve: A diagnostic for educational impact. arXiv preprint arXiv:2606.16206. https://arxiv.org/abs/2606.16206

Yao J, Zheng Z (2026b) Beyond agent architecture: Execution assumptions and reproducibility in LLM-based trading systems. arXiv preprint arXiv:2606.08285. https://arxiv.org/abs/2606.08285

Zhang H, Li Z, Bao R, Gao Y, Xiao X, Zhang H, Zhang S, Huang B, Wu Y, Wang T et al (2025a) HyperAdaLoRA: Accelerating LoRA rank allocation during training via hypernetworks without sacrificing performance. arXiv preprint arXiv:2510.02630.

Zhang H, Lyu M, Chen Z, Xing X, Ao Y, Lin Y (2025b) Pdtrim: Targeted pruning for prefill-decode disaggregation in inference. arXiv preprint arXiv:2509.04467.

Zhang H, Lyu M, Huang B, Ao Y, Lin Y (2025c) TrimTokenator-LC: Towards adaptive visual token pruning for large multimodal models with long contexts. arXiv preprint arXiv:2512.22748.