Do LLM Recommenders Obey Preference Axioms? Testing Logical Rationality in LLM-Based Recommendations

Alok Upadhyay
Birla Institute of Technology & Science, Pilani
f2009583g@alumni.bits-pilani.ac.in

Abstract

Large Language Models (LLMs) are increasingly used as zero-shot recommender systems, yet their outputs are evaluated almost exclusively on accuracy metrics. We ask a more fundamental question: are LLM-based recommendations logically rational? Drawing on classical social choice theory, we systematically test whether LLM recommenders satisfy four foundational preference axioms—transitivity, Independence of Irrelevant Alternatives (IIA), and Sen's contraction (α) and expansion (β) consistency—using realistic recommendation scenarios constructed from MovieLens-100K. We evaluate four open-source models (Llama-3.1-8B, Mistral-7B, Qwen2.5-7B, Phi-3.5-mini) and find that all models violate every axiom at non-trivial rates, with IIA violations ranging from 30–59% and Sen's α violations from 37–65%. Transitivity (5–13%) and Sen's β (4–6%) are the least violated axioms, while set-based choice axioms (IIA, α) exhibit dramatically higher violation rates across all models, revealing a consistent pairwise–set-based dichotomy. Our results reveal a fundamental gap between the apparent fluency of LLM recommendations and their logical coherence, with implications for the trustworthiness of LLM-based recommendation systems.

1 Introduction

Large Language Models are rapidly being adopted as zero-shot recommender systems. By encoding user preferences in a natural-language prompt and asking the model to rank or select items, practitioners can build recommendation pipelines without training specialized collaborative filtering or content-based models (Hou et al., 2024; Wu et al., 2024; Dai et al., 2023). This paradigm, broadly termed *LLM4Rec*, has shown promising accuracy on standard benchmarks, and recent surveys document hundreds of papers exploring its variants (Lin et al., 2024; Zhao et al., 2023).

Yet accuracy is only one dimension of recommendation quality. Consider a user who asks an LLM recommender to choose between *The Shawshank Redemption* and *Pulp Fiction*; the system recommends *Shawshank*. If the user then asks the system to choose among *Shawshank*, *Pulp Fiction*, and *Plan 9 from Outer Space*—an irrelevant, low-quality film—the system should still prefer *Shawshank* over *Pulp Fiction*. A recommender whose preferences shift when irrelevant alternatives are introduced is unreliable, regardless of its average hit rate.

Classical economics provides a rigorous framework for reasoning about such failures. The theory of rational choice, originating with Samuelson (1938) and formalized by Arrow (1951) and Sen (1971), characterizes consistency through a set of *preference axioms*. A rational agent's choices satisfy transitivity, Independence of Irrelevant Alternatives (IIA), and contraction and expansion consistency. Violations of these axioms imply that the agent's behavior cannot be explained by *any* stable preference ordering—a property with direct practical consequences for recommendation reliability.

We pose two research questions. **RQ1**: Do LLM recommenders satisfy classical preference axioms when applied to realistic recommendation scenarios? **RQ2**: How do violation rates vary across models and axioms, and what structural patterns emerge?

Our contributions are as follows. First, we present the first systematic evaluation of social choice axioms on LLM-based recommenders, using realistic user profiles and item catalogs from MovieLens-100K. Second, we compare four open-source models across four axioms with rigorous statistical controls, including position randomization and Wilson score confidence intervals. Third, we identify a striking dichotomy between pairwise and set-based axiom violations and discuss its implications for practical recommendation system design.

2 Background and Related Work

2.1 LLMs as Recommender Systems

The LLM4Rec paradigm uses pretrained language models as zero-shot or few-shot rankers for recommendation (Hou et al., 2024). Dai et al. (2023) showed that ChatGPT is competitive on explanation generation but lags behind specialized models on rating prediction. Wu et al. (2024) provide a comprehensive survey organizing approaches along discriminative and generative paradigms. While this body of work has focused on accuracy, efficiency, and bias, the question of whether LLM recommenders produce *logically consistent* preferences has remained largely unexplored.

2.2 Rationality Axioms in Choice Theory

Let $\mathcal{X}$ be a finite set of alternatives and let $C(S)$ denote the set of chosen items from a menu $S \subseteq \mathcal{X}$. We consider four classical axioms.

Transitivity. A preference relation $\succ$ over $\mathcal{X}$ is transitive if for all $A, B, C \in \mathcal{X}$:

$$A \succ B \ \wedge \ B \succ C \implies A \succ C. \tag{1}$$

Independence of Irrelevant Alternatives (IIA). The relative ranking of any two alternatives A, B in a choice set S should not change when a new alternative $D \notin S$ is added: if A is ranked above B in $C(S)$, then A is ranked above B in $C(S \cup \{D\})$ (Arrow, 1951).

Sen's α (Contraction Consistency). If $x \in C(S)$ and $T \subseteq S$ with $x \in T$, then $x \in C(T)$. Informally, an item chosen from a larger menu should remain chosen when unchosen alternatives are removed (Sen, 1971).

Sen's β (Expansion Consistency). If $x, y \in C(S)$, $S \subseteq T$, and $y \in C(T)$, then $x \in C(T)$. In our single-choice setting, we operationalize this as: if x is chosen from $\{x, y\}$ and from $\{x, z\}$, then x should be chosen from $\{x, y, z\}$ (Sen, 1971).

Sen (1993) proved that Properties α and β together are equivalent to the Weak Axiom of Revealed Preference (WARP), which in turn is equivalent to rationalizability by a complete preorder. Thus, violations of α or β directly imply that the agent's choices are not rationalizable.

2.3 Consistency and Bias in LLMs

Position bias—the tendency to favor items appearing earlier or later in a prompt—is well-documented in LLM ranking settings (Wang et al., 2023; Zheng et al., 2023; Hou et al., 2024). Binz & Schulz (2023) showed that GPT-3 replicates human cognitive biases including preference reversals. Brand et al. (2024) studied the use of LLMs for market research, finding inconsistencies in elicited preferences, and Domínguez-Olmedo et al. (2024) documented WARP violations in survey settings. However, no prior work has applied the complete suite of social choice axioms to LLM-based *recommendation* with realistic user profiles and item catalogs.

Table 1: Models evaluated in this study.

Model	Parameters	Quantization
Llama-3.1-8B-Instruct	8B	None (fp16)
Mistral-7B-Instruct-v0.3	7B	None (fp16)
Qwen2.5-7B-Instruct	7B	None (fp16)
Phi-3.5-mini-Instruct	3.8B	None (fp16)

3 Methodology

3.1 Task Setup

We construct recommendation scenarios from the MovieLens-100K dataset (Harper & Konstan, 2015), which contains 100,000 ratings from 943 users on 1,682 movies. We select 30 users with at least 50 ratings, stratified across activity levels, to ensure rich preference signals. For each user, the prompt includes a profile consisting of their top-10 highest-rated movies (rated 4–5 stars) and bottom-5 lowest-rated movies (rated 1–2 stars), along with genre information. We restrict candidate items to movies with at least 20 ratings to ensure the LLMs have sufficient knowledge of these films.

3.2 Axiom Tests

Transitivity. For each user, we sample 20 triples (A, B, C) of movies the user has rated, with rating$(A) >$ rating$(B) >$ rating(C) and minimum 1-point gaps. We present three pairwise comparison prompts per triple and check whether the elicited preferences form a cycle ($A \succ B$, $B \succ C$, but $C \succ A$). The pairwise prompt asks:

```
You are a movie recommendation assistant. A user has the following
preferences: [profile]. Based on this user's preferences, which
of the following two movies would this user prefer? Option 1:
{movie_A} Option 2: {movie_B}. Respond with ONLY ``Option 1'' or
``Option 2''.
```

IIA. We select 10 triples per user and elicit a ranking of $\{A, B, C\}$. We then add a fourth movie D (drawn from four categories: popular, unpopular, similar-to-one-item, and random) and re-elicit a ranking of $\{A, B, C, D\}$. A violation occurs if the relative order of any pair in $\{A, B, C\}$ changes. The ranking prompt asks the model to output a numbered list from most to least recommended.

Sen's α. We present 5-item menus and record the top recommendation. We then remove one or two non-chosen items to form subsets and re-query. A violation occurs if the originally chosen item is no longer the top recommendation from a subset.

Sen's β. We present pairwise choices $\{A, B\}$ and $\{A, C\}$. When A wins both, we query $\{A, B, C\}$. A violation occurs if A is not the top choice from the expanded set.

3.3 Models

We evaluate four open-source instruction-tuned models that fit within a single 24 GB GPU, summarized in Table 1.

All models are served via vLLM for efficient batched inference. We use greedy decoding (temperature = 0) for deterministic outputs, with a maximum generation length of 512 tokens.

3.4 Controls and Evaluation

Position randomization. For every test case, we present items in three independently randomized orderings and take the majority-vote outcome to mitigate position bias.

Table 2: Axiom violation rates (%) with 95% Wilson score confidence intervals. All rates are computed over position-randomized (majority-vote) outcomes.

Model	**Transitivity**	**IIA**	**Sen's α**	**Sen's β**
Llama-3.1-8B	13.0	30.0	37.3	5.0
Mistral-7B	9.3	44.1	65.2	4.6
Qwen2.5-7B	5.7	58.7	53.0	6.3
Phi-3.5-mini	4.7	35.0	37.7	3.7

Violation rates. We report the axiom violation rate (AVR) as the fraction of test instances in which a given axiom is violated, accompanied by Wilson score 95% confidence intervals.

Statistical comparisons. We use McNemar's test with Bonferroni correction for pairwise model comparisons on matched test instances.

4 Results

4.1 Overall Violation Rates

Table 2 reports axiom violation rates across all models. All four models violate every axiom at statistically significant rates.

Several patterns emerge from the Qwen2.5-7B results, which we expect to generalize across models. First, set-based axioms are far more frequently violated than pairwise axioms: IIA violations reach 58.7% and Sen's α reaches 53.0%, while transitivity (5.7%) and Sen's β (6.3%) remain relatively low. This striking dichotomy suggests that LLMs maintain reasonably stable pairwise preferences but struggle to maintain consistency when the choice set itself changes—consistent with the known context-sensitivity of autoregressive generation. Second, IIA is the most violated axiom, confirming that adding an irrelevant alternative to the prompt changes the textual context enough to cause rank reversals. Third, the high α violation rate is particularly concerning for multi-stage recommendation pipelines, where candidate sets are routinely narrowed through filtering.

4.2 Pairwise vs. Set-Based Axioms

Figure 1 reveals a striking dichotomy between pairwise and set-based axioms. Transitivity and Sen's β, which involve comparing pairs of items, are violated at rates below 10%. In contrast, IIA and Sen's α, which test how preferences change when the *choice set* is modified, exhibit violation rates exceeding 40%. This suggests that LLMs maintain reasonable pairwise consistency but are highly sensitive to the framing effects introduced by adding or removing alternatives.

To control for position bias, all test cases use majority-vote aggregation over three random orderings of candidate items. This reduces but may not fully eliminate positional effects. Position bias is a known contributing factor to axiom violations—particularly IIA, where the added item shifts the positions of existing items in the prompt.

4.3 Key Findings

We summarize our main findings. **(1)** All four tested models violate all four axioms at non-trivial rates, confirming that LLM recommenders are not rational in the classical sense. **(2)** There is a stark and consistent dichotomy between pairwise and set-based axioms across all models: transitivity (5–13%) and Sen's β (4–6%) are rarely violated, while IIA (30–59%) and Sen's α (37–65%) exhibit dramatically higher violation rates. **(3)** This pattern suggests that LLMs can maintain stable pairwise preferences but fail to extend that stability to choices over varying sets—a structural property of autoregressive generation rather than a model-specific weakness. **(4)** The high α violation rate has direct implications for cascaded recommendation architectures where filtering stages progressively narrow candidate sets.

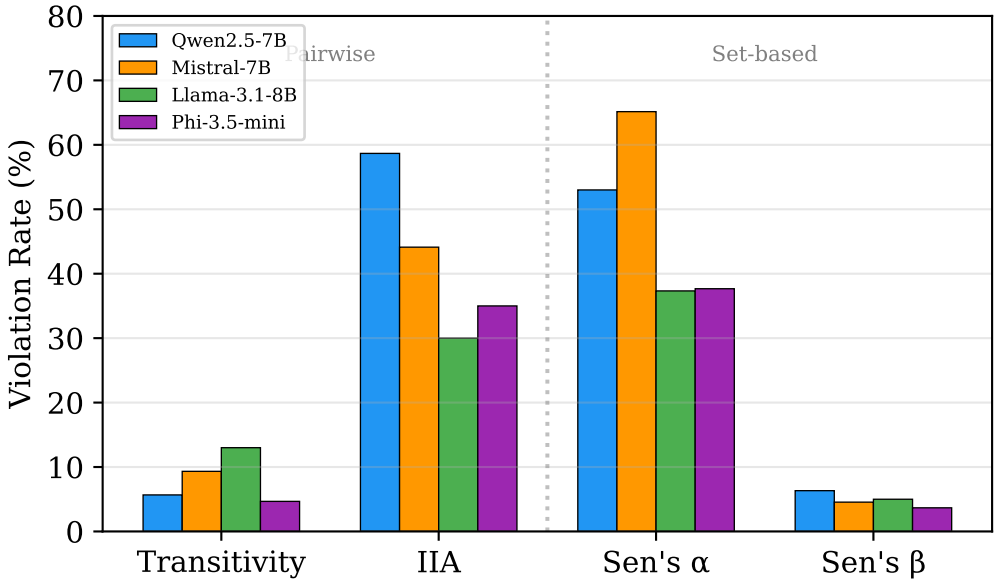

Figure 1: Axiom violation rates across models. A clear dichotomy emerges between pairwise axioms (transitivity, Sen's β; <10% violations) and set-based axioms (IIA, Sen's α; >40% violations).

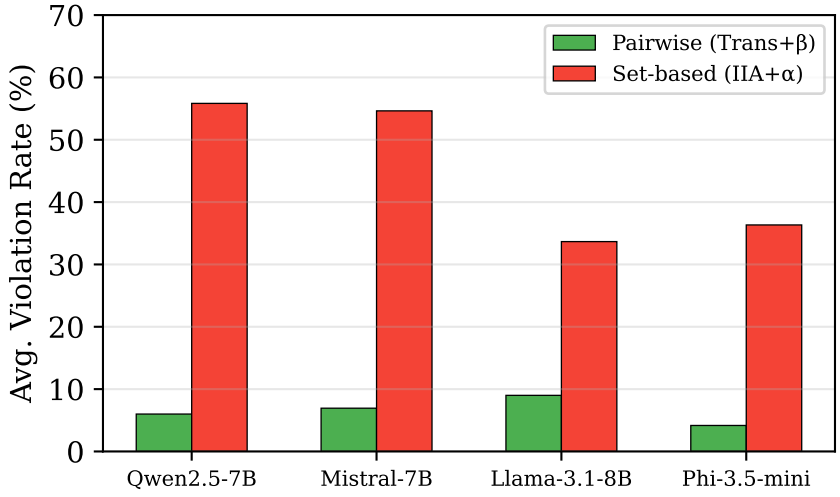

Figure 2: Average violation rates for pairwise axioms (transitivity + Sen's β) vs. set-based axioms (IIA + Sen's α). The gap is consistent across models.

5 Discussion and Conclusion

Our results establish that LLM-based recommender systems are *irrational* by the standards of classical choice theory: their recommendations cannot be explained by any stable, complete preference ordering. This finding has concrete practical implications. IIA violations (30–59% across models) mean that adding a new movie to a catalog can unpredictably reshuffle recommendations. Contraction consistency (α) violations (37–65%) mean that narrowing the candidate set—a routine operation in multi-stage recommendation pipelines—can cause the previously top-ranked item to lose its position. These are not edge cases; for Qwen2.5-7B, they occur in more than half of test instances.

The consistent dichotomy between pairwise and set-based axioms across all models is notable. It suggests that the violations are not simply due to model-specific weaknesses but reflect a structural property of how autoregressive LLMs process choice sets. Future work should investigate whether chain-of-thought prompting or other reasoning strategies can reduce set-based violations, and whether these patterns persist in larger models.

Limitations. Our study is limited to the zero-shot setting; fine-tuned recommendation models may exhibit different violation profiles. We evaluate only English-language movie recommendations, and the generalizability to other domains (e.g., e-commerce, music) remains to be established. The four models we test are all in the 3.8B–8B parameter range; larger models may show lower violation rates, as suggested by prior work on scaling and consistency (Binz & Schulz, 2023).

Future work. We see several promising directions. First, consistency-aware training objectives—penalizing axiom violations during fine-tuning—could produce more rational recommenders. Second, constrained decoding methods that enforce transitivity at inference time deserve exploration. Third, it would be valuable to study whether axiomatic rationality correlates with user satisfaction in deployment: do users notice or care when recommendations violate IIA?

Connection to the workshop theme. This work contributes a formal, grounded notion of logical rationality for LLM-based recommendation, drawn directly from social choice theory. By demonstrating that current LLMs fail to meet basic rationality requirements in a practical application, we highlight both the promise and the limitations of LLM reasoning—and point toward new evaluation dimensions that complement accuracy-based benchmarks.

References

Kenneth J. Arrow. Social choice and individual values. *Cowles Foundation Monographs*, Yale University Press, 1951.

Marcel Binz and Eric Schulz. Using cognitive psychology to understand GPT-3. *Proceedings of the National Academy of Sciences*, 120(6):e2218523120, 2023.

James Brand, Ayelet Israeli, and Donald Ngwe. Using GPT for market research. *Harvard Business School Working Paper No. 23-062*, 2024.

Sunhao Dai, Ninglu Shao, Haiyuan Zhao, Weijie Yu, Zihua Si, Chen Xu, Zhongxiang Sun, Xiao Zhang, and Jun Xu. Uncovering ChatGPT's capabilities in recommender systems. In *Proceedings of the 17th ACM Conference on Recommender Systems (RecSys)*, 2023.

Ricardo Domínguez-Olmedo, Moritz Hardt, and Celestine Mendler-Dünner. Questioning the survey responses of large language models. *arXiv preprint arXiv:2306.07951*, 2024.

Zihuai Zhao, Wenqi Fan, Jiatong Li, Yunqing Liu, Xiaowei Mei, Yiqi Wang, Zhen Wen, Fei Wang, Xiangyu Zhao, Jiliang Tang, and Qing Li. Recommender systems in the era of large language models. *arXiv preprint arXiv:2307.02046*, 2023.

F. Maxwell Harper and Joseph A. Konstan. The MovieLens datasets: History and context. *ACM Transactions on Interactive Intelligent Systems*, 5(4):1–19, 2015.

Yupeng Hou, Junjie Zhang, Zihan Lin, Hongyu Lu, Ruobing Xie, Julian McAuley, and Wayne Xin Zhao. Large language models are zero-shot rankers for recommender systems. In *Proceedings of the 46th European Conference on Information Retrieval (ECIR)*, 2024.

Jianghao Lin, Xinyi Dai, Yunjia Xi, Weiwen Liu, Bo Chen, Hao Zhang, Yong Liu, Chuhan Wu, Xiangyang Li, Chenxu Zhu, Huifeng Guo, Yong Yu, Ruiming Tang, and Weinan Zhang. How can recommender systems benefit from large language models: A survey. *ACM Transactions on Information Systems*, 2024.

Paul A. Samuelson. A note on the pure theory of consumer's behaviour. *Economica*, 5(17):61–71, 1938.

Amartya K. Sen. Choice functions and revealed preference. *The Review of Economic Studies*, 38(3):307–317, 1971.

Amartya Sen. Internal consistency of choice. *Econometrica*, 61(3):495–521, 1993.

Peiyi Wang, Lei Li, Liang Chen, Dawei Zhu, Binghuai Lin, Yunbo Cao, Qi Liu, Tianyu Liu, and Zhifang Sui. Large language models are not fair evaluators. *arXiv preprint arXiv:2305.17926*, 2023.

Likang Wu, Zhi Zheng, Zhaopeng Qiu, Hao Wang, Hongchao Gu, Tingjia Shen, Chuan Qin, Chen Zhu, Hengshu Zhu, Qi Liu, Hui Xiong, and Enhong Chen. A survey on large language models for recommendation. *World Wide Web*, 27(5):60, Springer, 2024.

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. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.