RESEARCH-ARTICLE

Exploring the Potential of LLMs for Serendipity Evaluation in Recommender Systems

LI KANG, Hong Kong Baptist University, Hong Kong, Hong Kong

YUHAN ZHAO, Hong Kong Baptist University, Hong Kong, Hong Kong

LI CHEN, Hong Kong Baptist University, Hong Kong, Hong Kong

Published: 22 September 2025

RecSys '25: Nineteenth ACM Conference on Recommender Systems
September 22 - 26, 2025
Prague, Czech Republic

Conference Sponsors:
SIGCHI

https://doi.org/10.1145/3705328.3748167
ISBN: 9798400713644

Exploring the Potential of LLMs for Serendipity Evaluation in Recommender Systems

Li Kang*
Hong Kong Baptist University
Hong Kong, China
likang@comp.hkbu.edu.hk

Yuhan Zhao*
Hong Kong Baptist University
Hong Kong, China
csyhzhao@comp.hkbu.edu.hk

Li Chen
Hong Kong Baptist University
Hong Kong, China
lichen@comp.hkbu.edu.hk

Abstract

Serendipity plays a pivotal role in enhancing user satisfaction within recommender systems, yet its evaluation poses significant challenges due to its inherently subjective nature and conceptual ambiguity. Current algorithmic approaches predominantly rely on proxy metrics for indirect assessment, often failing to align with real user perceptions, thus creating a gap. With large language models (LLMs) increasingly revolutionizing evaluation methodologies across various human annotation tasks, we are inspired to explore a core research proposition: *Can LLMs effectively simulate human users for serendipity evaluation?*

To address this question, we conduct a meta-evaluation on two datasets derived from real user studies in the e-commerce and movie domains, focusing on three key aspects: the accuracy of LLMs compared to conventional proxy metrics, the influence of auxiliary data on LLM comprehension, and the efficacy of recently popular multi-LLM techniques. Our findings indicate that even the simplest zero-shot LLMs achieve parity with, or surpass, the performance of conventional metrics. Furthermore, multi-LLM techniques and the incorporation of auxiliary data further enhance alignment with human perspectives. Based on our findings, the optimal evaluation by LLMs yields a Pearson correlation coefficient of 21.5% when compared to the results of the user study. This research implies that LLMs may serve as potentially accurate and cost-effective evaluators, introducing a new paradigm for serendipity evaluation in recommender systems. Our code is publicly available at https://github.com/Leah-HKBU/SerenEva.

CCS Concepts

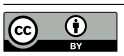
• **Information systems** → **Recommender systems**;

Keywords

Recommender Systems, Serendipity, Large Language Models

ACM Reference Format:
Li Kang, Yuhan Zhao, and Li Chen. 2025. Exploring the Potential of LLMs for Serendipity Evaluation in Recommender Systems. In *Proceedings of the Nineteenth ACM Conference on Recommender Systems (RecSys '25), September 22–26, 2025, Prague, Czech Republic.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3705328.3748167

*Equal contributions.



RecSys '25, Prague, Czech Republic

ACM ISBN 979-8-4007-1364-4/25/09
https://doi.org/10.1145/3705328.3748167

1 Introduction

Serendipity plays a pivotal role in enhancing user satisfaction within recommender systems by mitigating the effects of the information cocoon and filter bubble issues from traditional recommendation methods [7, 9, 12, 20, 29]. Despite growing recognition of serendipity's importance and subsequent algorithmic innovations, how to evaluate serendipity remains challenging. This difficulty arises from its inherently subjective nature and conceptual ambiguity, distinguishing it from traditional accuracy-oriented metrics [5, 8, 43]. The gold standard for user-centered evaluation involves carefully designed user studies that directly capture user feedback [4, 15, 25], which, however, are costly in practice. As a result, many researchers rely on predefined proxy metrics (e.g., relevance-unexpectedness) to approximate serendipity scores, which are then treated as ground truth for algorithm evaluation [20, 21, 36]. Nonetheless, the gap between these indirect measurements and actual user perceptions introduces bias into serendipity research [5, 16].

The emergence of large language models (LLMs) has revolutionized evaluation methodologies across human annotation tasks, showcasing remarkable potential in user simulation and automatic assessment [2, 3, 18, 28]. This breakthrough motivates our key research question: ***Can LLMs effectively simulate human users for serendipity evaluation?*** If the answer is yes, the LLM-based evaluation approach could effectively combine the accuracy of user studies with the efficiency of proxy metrics, potentially transforming research paradigms in serendipity. While preliminary explorations have been conducted [7, 26], to our knowledge, no comprehensive studies have addressed the effectiveness of various LLM techniques across different product domains and data types for serendipity evaluation. To systematically investigate this proposition, we have attempted to address the following three research questions:

- *RQ1: Can LLMs using basic prompt strategies surpass conventional proxy metrics in serendipity evaluation?*
- *RQ2: What auxiliary data (e.g., user age and gender) might further enhance the LLMs' understanding of serendipity?*
- *RQ3: Can advanced multi-LLM techniques help improve the accuracy of evaluation?*

To address these questions, we employ two user-study-validated serendipity datasets from distinct domains: e-commerce and movies. This allows us to identify both general and domain-specific observations, enhancing the comprehensiveness and credibility of our conclusions. Then, we propose SerenEva (**seren**dipity **eva**luation framework), a novel meta-evaluation framework that measures the alignment between LLM evaluator ratings and human judgments using correlation and error metrics.

For RQ1, we benchmark zero-shot and few-shot LLM evaluators (e.g., Qwen2.5-7B [35]) against conventional proxy metrics using basic prompting strategies. To address RQ2, we categorize and inject auxiliary data through structured prompting. For RQ3, we explore the effectiveness of multi-LLM techniques. Our findings reveal:

- Zero-shot and few-shot LLMs achieve parity with or surpass conventional metrics across various evaluation dimensions (e.g., Pearson correlation coefficient), demonstrating their viable potential as serendipity evaluators.
- The incorporation of auxiliary data (e.g., user curiosity and item similarity) markedly enhances the accuracy of LLM evaluations, although the optimal choice of auxiliary data is contingent upon the specific domain.
- Multi-LLM techniques with a score averaging strategy substantially improve evaluation performance and alignment with human judgments.

Based on these findings, the optimal evaluation by LLMs yields a Pearson correlation coefficient of over 20% when compared to the results of the user study, implying the potential of using LLMs to evaluate recommendation serendipity.

In summary, our contributions include investigating the feasibility of leveraging LLMs as evaluators for serendipity in recommender systems and exploring the incorporation of auxiliary data and multi-LLM techniques to enhance their understanding of serendipity. We demonstrate that LLMs, when appropriately prompted, may serve as potentially reliable and reproducible evaluators that surpass conventional proxy metrics in accuracy.

2 Experimental Setup

2.1 Problem Formulation

Let $\mathcal{U}$ ($|\mathcal{U}| = M$) and $\mathcal{V}$ ($|\mathcal{V}| = N$) denote the sets of users and items, respectively. Traditional recommender systems focus on recommending a subset of items $\mathcal{M} \subseteq \mathcal{V}$ to maximize accuracy metrics such as NDCG and Recall [40, 41]. In contrast, serendipity-oriented recommendations introduce an additional criterion: the ability to recommend unexpected yet relevant items [5]. Our research problem centers on identifying which method might approximate the ground truth obtained from user studies.

2.2 User Study as Gold Standard

We obtained two public serendipity datasets validated through rigorous user studies, with dataset statistics shown in Table 1:

- *Taobao Serendipity* [5], which was collected through a user survey on Mobile Taobao, a leading e-commerce platform in China. It captures user perceptions of serendipity ("pleasant surprise") on 5-point Likert scales, along with users' demographic profiles (e.g., age and gender) and psychological profiles including curiosity (CEI-II [13]) and Big-Five personality traits (TIPI [11]).
- *Serendipity-2018* [14], which was collected through a user survey on the MovieLens platform. The survey captures users' perceptions of serendipity through eight statements rated on 5-point Likert scales, covering aspects such as movie discovery, novelty, unexpectedness, and preference broadening. Although the dataset lacks direct user ratings on serendipity, we calculated the score by averaging three unexpectedness-related variables, as validated in [31].

Table 1: Statistics of two serendipity datasets.

Dataset	Domain	#Users	#Items	#Ratings
Taobao Serendipity	Shopping	11,383	9,985	11,383
Serendipity-2018	Movie	481	1,678	2,150

Note: The numbers of ratings refer specifically to serendipity ratings.

By examining two distinct product domains, e-commerce and movies, we aim to identify general and domain-specific insights, thereby enhancing the comprehensiveness of our conclusions. Because user studies provide real user feedback, making them the closest approximation to actual user perceptions. Therefore, we regard these results as the gold standard for evaluating various simulators (also called evaluators) in this work.

Another commonly used dataset is SerenLens [10], which includes annotations from third-party summarization. We do not use this dataset because our work focuses on using LLMs as user simulators, and user studies give us direct feedback from end users, which better matches our goals.

2.3 Proxy Metrics for Serendipity

Due to the cost and time constraints associated with user studies, many researchers have adopted proxy metrics to indirectly evaluate serendipity [17, 20, 21, 36]. These metrics do not incorporate real users' attitudes, but instead calculate serendipity scores based on their assumptions. With those scores, the evaluation transforms into a ranking task [17, 20] or regression task [36], using metrics such as NDCG or HR for the final assessment. For convenience, we use acronyms from the original papers to refer to these metrics.

SOG (Serendipity-Oriented Greedy) [17]. This method integrates multiple dimensions:

$$S_{uiB} = \sum_{\phi \in \Phi} \alpha_\phi \cdot \phi(uiB) \tag{1}$$

where $\Phi = \{\text{relevance, diversity, history dissimilarity, unpopularity}\}$ with corresponding weights α_ϕ.

SNPR (Serendipity-oriented Next POI Recommendation) [36]. In this work, serendipity is defined as a linear combination of relevance and unexpectedness:

$$\text{Serendipity}(i,u) = \lambda \text{R}(i,u) + (1-\lambda)\text{U}(i,u) \tag{2}$$

where $R(i,u)$ represents relevance, which can be inferred from user interactions, and $U(i,u)$ represents unexpectedness, calculated using multi-level dissimilarity measures.

PURS (Personalized Unexpected Recommender System) [20]. A hybrid utility function is proposed in this work to integrate the unexpectedness factor:

$$\text{Utility}_{u,i} = r_{u,i} + f(\text{unexp}_{u,i}) \cdot \text{unexp\_factor}_{u,i} \tag{3}$$

where unexpectedness derives from the embedding distance to the user interest clusters.

DESR (Directional and Explainable Serendipity Recommendation) [21]. It adopts an F-score-inspired formulation:

$$\mathrm{AD} = \frac{\mathrm{acc} \cdot \mathrm{dif}}{\mathrm{acc} + \mathrm{dif}} \tag{4}$$

where accuracy (acc) combines long- and short-term preference alignment, and dif balances diversity and historical dissimilarity.

2.4 LLM-Based Evaluation Framework

Our LLM evaluator employs constrained prompting strategies to assess the serendipity of recommendation items. The basic **prompt** is shown below.

> You are a user of a Chinese e-commerce platform, and you have received a user survey that aims to gather your opinion on the serendipity of the items recommended to you. Serendipity here means that the item recommended is a pleasant surprise.
> ## Background
> You have used the Chinese e-commerce platform, and this recommendation is based on your behavior history. You are provided with the genres of the recommended item and the items you have clicked on or purchased. Your behavior history is listed in a comma-separated format, sorted from oldest to newest.
> ## Task
> Please provide a serendipity rating for the recommended item using a 5-point Likert scale: 1 – "strongly disagree"; 2 – "disagree"; 3 – "neither agree nor disagree"; 4 – "agree"; 5 – "strongly agree".
> Rate the recommended item from the perspective of serendipity, based on your behavior history.
> ## Output Format
> Generate only the rating number, without any additional commentary or explanation.
> ## Response
> Your behavior history: [*user behavior history*]
> Recommended item: (*item info*)
> Your serendipity rating:

Additional details and the remainder of the prompt are available in our open-source codebase. All experiments were conducted using publicly available models (e.g., LLaMA2 [27], Qwen2.5 [35], GPT-4 [1]) to ensure transparency and reproducibility.

2.5 Meta-Evaluation Protocol (SerenEva)

One of the most widely used methods for assessing the efficacy of an evaluator is meta-evaluation [37]. To facilitate the meta-evaluation of serendipity in our case, we propose SerenEva, which includes an evaluation function $h(\cdot)$:

$$h = \frac{1}{|\mathcal{U}|} \sum_{u \in \mathcal{U}} r(s_*, s_{real}) \tag{5}$$

where s_{real} denotes the real user feedback. s_* represents serendipity scores derived from different evaluation methods (e.g., s_{LLM} stands for evaluation results based on LLMs). h measures the discrepancy between a particular evaluation method's evaluation result s_* and the user feedback s_{real} across various indicators r. To ensure a rigorous and comprehensive comparison, we introduce three major metrics: Pearson correlation coefficient, mean absolute error (MAE), and root mean squared error (RMSE). The Pearson correlation coefficient can capture linear relationships. MAE provides robustness against outliers, and RMSE imposes stricter penalties for larger deviations. If RMSE is much larger than MAE, it indicates that the LLM performs well for the majority of users but exhibits large errors for some users.

SerenEva requires the unification of input format and scope to align with the user feedback format. While LLM predictions can be obtained directly through constrained prompts, proxy metrics require certain adjustments. Since the ground truth in both datasets is based on a 5-point Likert scale (1 – "strongly disagree"; 2 – "disagree"; 3 – "neither agree nor disagree"; 4 – "agree"; 5 – "strongly agree"), we apply the following transformation to achieve this format:

$$\mathrm{score} = \mathrm{round}\left(\frac{\mathrm{output} - \mathrm{min\_output}}{\mathrm{max\_output} - \mathrm{min\_output}} \times 4 + 1\right) \tag{6}$$

max_output and min_output represent the maximum and minimum values of the metric's predictions across all outputs, respectively. output is the raw prediction value from the metric, and round$(\cdot)$ denotes rounding to the nearest integer. The final score is thus mapped to the 5-point Likert scale, ensuring compatibility with the ground truth rating.

Due to space constraints, we present only the experimental results that demonstrate statistically significant improvements over the strongest baseline(s), as determined by a two-sided t-test with $p < 0.05$. Additionally, all reported Pearson correlation coefficients are statistically significant at the $p < 0.05$ level. To ensure the stability and reproducibility of our findings, we used a low temperature setting (0.00001) and averaged the results over five runs.

3 RQ1: LLM as a Competitive Evaluator

In this section, we explore whether LLMs can surpass conventional proxy metrics in evaluating serendipity, even in the simplest settings. To this end, we exclusively rely on the user's historical behavior data, focusing specifically on the last 10 items interacted with by the user prior to the target item in both datasets. The results are presented in Table 2.

3.1 Limitations of Conventional Proxy Metrics in Capturing User Feedback on Serendipity

As expected, proxy metrics surpass random baselines, indicating their ability to partially capture user serendipity perception. Notably, the **SOG** metric [17] demonstrates the best performance among all proxy metrics, likely due to its comprehensive integration of relevance, popularity, and diversity. However, SOG still shows a significant gap compared to real user feedback, likely due to its use of a fixed weighting strategy for different components.

Furthermore, most metrics exhibit consistent performance, except SNPR [36]. On the Serendipity-2018 dataset, despite poor Pearson correlation results, SNPR performs well in terms of MAE and RMSE. This inconsistency stems from SNPR's design, which heavily weights relevance (70%) in its calculations. When recommendations

Table 2: Performance comparison of conventional proxy metrics and LLMs. The best results are highlighted in bold, and the second-best results are underlined. SerenPrompt_1 and SerenPrompt_2 represent Discrete Style 1 and 2 from [7], respectively.

Method	Taobao Serendipity			Serendipity-2018		
	Pearson(%)	MAE	RMSE	Pearson(%)	MAE	RMSE
Random	-0.5771	1.6484	2.0637	-0.1329	1.4400	1.7847
Proxy Metrics						
SOG	4.6095	1.5231	1.8736	5.7226	1.2746	1.5725
PURS	2.6744	1.6540	2.0356	3.3270	1.2604	1.5608
DESR	3.8458	1.6132	1.9675	3.4840	1.1170	1.4059
SNPR	0.9033	1.5969	1.9955	0.0509	1.1062	1.3824
LLM-based Baseline						
LLM4Seren	4.1338	1.4286	1.7603	-7.6891	1.1084	1.4113
SerenPrompt_1	2.2306	1.5063	1.8764	6.0190	1.0165	1.3215
SerenPrompt_2	9.2422	1.4271	1.8103	8.3274	1.0773	1.3936
LLM-based inference (Zero-Shot)						
LLaMA2-7B	0.2766	1.6311	1.9910	-0.0802	1.1464	1.4653
LLaMA2-13B	1.4615	1.6236	1.9809	0.6997	1.5564	1.8794
Qwen2.5-7B	7.5001	1.5081	1.8401	4.1820	1.3485	1.5764
Qwen2.5-14B	7.1590	1.4153	1.7926	6.2096	1.2507	1.5655
Qwen2.5-72B	10.5165	**1.3263**	**1.6601**	10.3220	1.1885	1.4678
GPT-4	10.8139	1.5316	1.8631	10.8859	1.0126	1.3292
LLM-based inference (Few-Shot)						
LLaMA2-7B	1.9549	1.5694	1.9687	3.0739	1.0690	1.3945
LLaMA2-13B	1.9945	1.5696	1.9689	6.7423	1.4150	1.6949
Qwen2.5-7B	10.5019	1.3836	1.7749	6.4900	1.0068	1.2819
Qwen2.5-14B	10.2701	1.4018	1.7836	11.4769	0.9742	1.2825
Qwen2.5-72B	11.9836	1.3756	1.7734	8.3117	**0.8591**	**1.1830**
GPT-4	**12.1231**	1.5111	1.8405	**14.5545**	1.0112	1.3436

are serendipitous, they may contain unexpected items that exhibit lower relevance scores. These cases tend to become outliers in SNPR's assessment, particularly affecting Pearson correlation calculations, while MAE and RMSE remain more robust as they measure absolute differences rather than linear relationships.

3.2 Potential of LLMs as Evaluators

In zero-shot settings, LLMs such as Qwen2.5-14B and GPT-4 surpass conventional metrics across both datasets by approximately 100% in Pearson correlation compared to the best proxy metric (**SOG**). We attribute this to LLMs' robust background knowledge and user simulation capabilities, underscoring their potential as competent evaluators in the serendipity evaluation context.

In few-shot settings with five examples, most LLMs (e.g., Qwen family and GPT-4) demonstrate superior performance compared to that in zero-shot settings. Remarkably, even the smaller parameter model Qwen2.5-7B achieves an impressive performance score of 10.50% on the Taobao dataset, approaching the performance of Qwen2.5-72B in zero-shot scenarios. This phenomenon inspires us to consider that, in the future, we can effectively enhance the evaluation capabilities of LLMs by incorporating limited user study data. More importantly, these results underscore the potential of smaller parameter models to deliver high-quality outcomes in few-shot learning scenarios. They offer significant advantages in terms of time and cost efficiency, making them a compelling and practical choice for various applications.

Regrettably, the LLaMA family models do not exhibit exceptional performance, especially on the Taobao dataset. This limitation may stem from the dataset's focus on the Chinese community and its inclusion of product descriptions in Chinese. Previous research has pointed out LLaMA2's limited proficiency in processing Chinese [33, 38], which may explain its subpar performance in this scenario.

3.3 LLM-based Methods

SerenPrompt [7] aims to leverage the inference capability of LLMs to predict the serendipity level. By utilizing its prompts, we explore the application of this method for evaluation purposes. LLM4Seren [26] introduces a straightforward LLM-based approach for assessing recommendation serendipity and validates the evaluation capability of LLMs and the **SOG** metric, which serves as a proxy metric we have discussed.

Table 3: Auxiliary Data

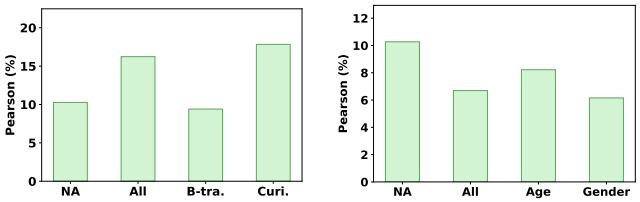
Data type	Description
	User data
Psychological Data	Curiosity and Big-Five personality traits
Demographic Data	Age and Gender
Profile Data	Long-term profile and Short-term profile respectively containing long and recent interactions
	Item data
Popularity	The item's popularity
Similarity	The item's similarity to others
	Interaction data
Interaction Length	The length of historical interactions used
Interaction Type	The type of interaction, such as clicks, ratings.

Our results, based on Qwen2.5-14B [35], show that SerenPrompt performs well, likely due to its accommodation of both unexpectedness and relevance in prompts, which are two crucial elements of serendipity. However, there is still a performance gap compared to our method. The core reason for this gap might be that SerenPrompt optimizes the recommendation task rather than being specifically designed as an evaluator, thus resulting in a natural disparity in final performance. Conversely, LLM4Seren performs poorly, possibly due to its lack of an explicit serendipity definition in prompts, leading to suboptimal evaluation capability. These observations drive us to carefully design and explore the use of LLMs as evaluators of recommendation serendipity.

4 RQ2: The Role of Auxiliary Data in Enhancing LLM-based Serendipity Evaluation

In this section, we leverage Qwen2.5-14B [35] to investigate how different types of auxiliary data can influence LLM-based serendipity evaluation performance. Our goal is to identify which type(s) of data might be useful for enhancing the model's evaluation accuracy. Throughout this section, our default configuration uses five-shot prompting with the 10 most recent user interactions prior to the target item, with "NA" denoting the baseline condition without auxiliary data. Based on the work of [31, 34, 36], we categorize the auxiliary data into three distinct types: user data, item data, and interaction data, as summarized in Table 3. User data comprises psychological, demographic, and profile data, capturing a comprehensive view of individual characteristics and behaviors. Item data includes key attributes of items, such as popularity and item similarity. Interaction data focuses on two key dimensions: (1) the quantity of interaction data used, and (2) the types of interactions, such as clicks, purchases, and ratings.

Furthermore, we also explore the impact of category-level data [31] that captures user preferences for broader item categories (such as weekly interaction frequency, hourly interaction frequency, time since last interaction, and category interaction frequency). However, our experiments revealed that adding such data types does not significantly improve evaluation performance. Due to space limitations, we thus omit this part of the experimental results.



(a) **Psychological Data** *(b)* **Demographic Data**

Figure 1: Comparison w.r.t. Pearson correlation coefficient for different user attributes in the Taobao dataset (B-tra.: Big Five Personality Traits, Curi.: Curiosity, All: Combined data within each figure).

4.1 User Data

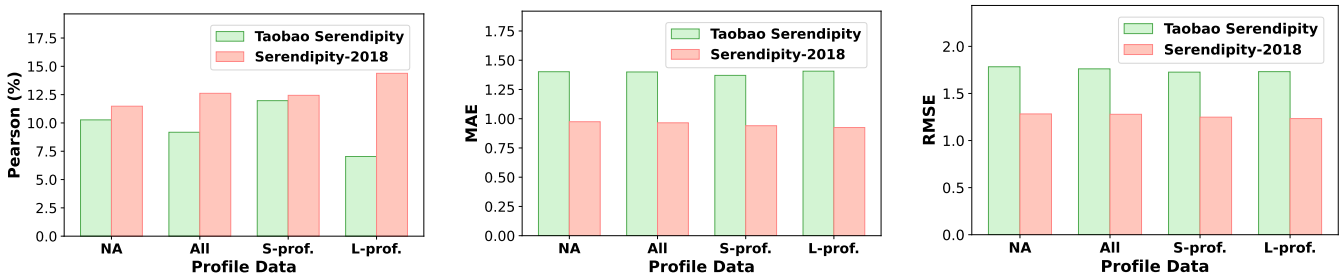
Since the Serendipity-2018 dataset lacks demographic and psychological user data, we mainly investigated their effects using the Taobao dataset. The results are presented in Figure 1. Notably, incorporating curiosity significantly improved performance, achieving a Pearson correlation coefficient of 17.83%. This is intuitive, as curiosity influences user perceptions of serendipity: more curious users are more inclined to explore items with lower relevance but higher unexpectedness [5, 6, 39].

Conversely, attributes like the Big-Five personality traits, age, and gender did not help enhance performance and even reduced it, which contradicts previous studies linking certain Big-Five traits (e.g., extraversion and neuroticism) to serendipity [31, 32]. We propose two possible reasons for this finding: (1) These attributes represent complex user characteristics, which current LLMs may not fully understand and simulate. (2) The relationship between these attributes and serendipity is less apparent than that of curiosity. This subtle relationship might be challenging for LLMs to capture without additional data support and instructions.

In addition, regarding user behavior history that contains the user's previous interaction data, though it can be useful for capturing user preferences, excessively long history can impair the reasoning capability of LLMs [19, 22]. To address this, following [34], we employ an LLM to summarize long-term interactions into a *long-term user profile*, while interactions from the past 2–4 weeks are aggregated to construct the *short-term profile*. Injecting the two types of user profiles into the LLM-based evaluator, respectively, reveals the following observation (see Figure 2): For the Taobao dataset, incorporating the short-term profile improves the Pearson correlation coefficient. In contrast, for the Serendipity-2018 dataset, the long-term profile yields a higher Pearson correlation coefficient. The results align with the discussion from [31] that in shopping domains like Taobao, short-term relevance might play a more important role in driving serendipity, while in movie domains, long-term unexpectedness tends to have a greater influence.

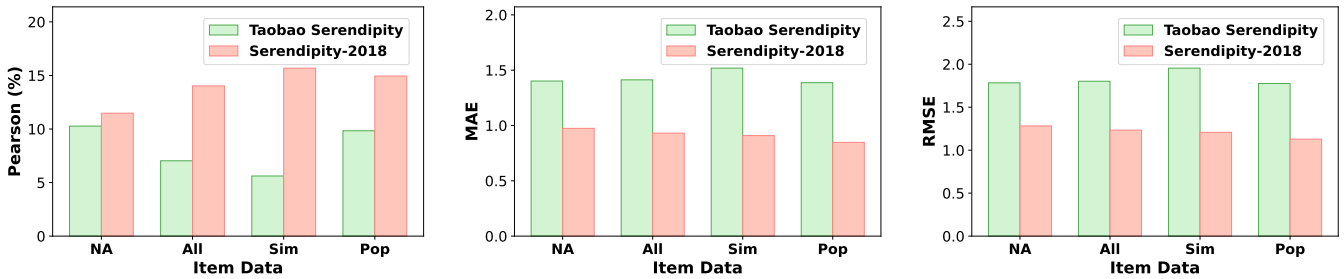
4.2 Item Data

For item data, we primarily consider item popularity and similarity. Popularity is measured as the percentage of users who rated the movie in Serendipity-2018, while in Taobao, it is binary-coded (1 for items in the platform's HOT function, 0 otherwise). Similarity is represented by the minimum collaborative-based Jaccard



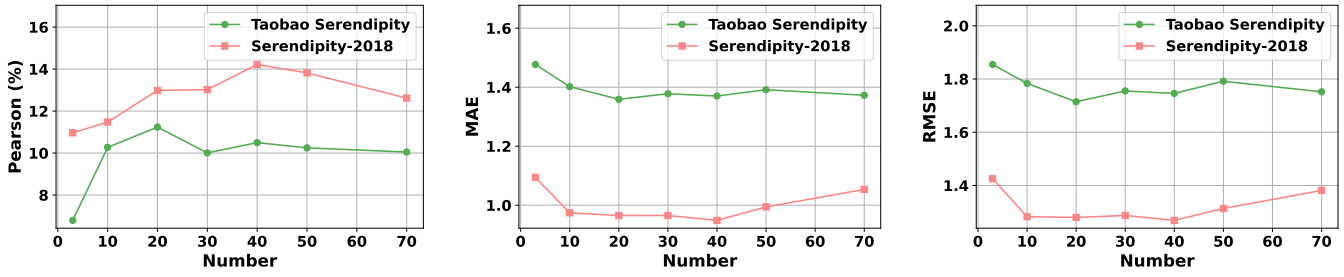
(a) Pearson correlation coefficient (b) MAE (c) RMSE

Figure 2: Comparison w.r.t. Pearson correlation coefficient, MAE, and RMSE for different profile types ("S-prof." stands for short-term user profile, and "L-prof." stands for long-term profile). Here, "ALL" denotes the condition with both long-term and short-term profiles.



(a) Pearson correlation coefficient (b) MAE (c) RMSE

Figure 3: Comparison w.r.t. Pearson correlation coefficient, MAE, and RMSE for different item data types ("Sim" stands for item similarity and "Pop" for popularity). Here, "ALL" denotes the condition with both popularity and similarity data.



(a) Pearson correlation coefficient (b) MAE (c) RMSE

Figure 4: Comparison w.r.t. Pearson correlation coefficient, MAE, and RMSE with varying interaction history length across the two datasets.

distance between the target item and the user's historical interactions. Figure 3 shows the different effects of item attributes on LLM evaluation performance across the two datasets.

- For *popularity*, its incorporation yields minimal performance improvements on the Taobao dataset, but significantly enhances performance on the Serendipity-2018 dataset. This disparity stems from inherent domain differences: the movie domain naturally encourages exploration due to minimal risk and cost, making users more open to unpopular items, while e-commerce scenarios involve monetary investment that may limit such exploratory behavior.
- For *similarity*, a similar trend is observed. On e-commerce platforms, an item that a user interacted with several months ago, even if it is very similar to the target item, might not be relevant for the current purchase decision. In contrast, for movies, user preferences tend to be more stable, and a movie with high similarity that was watched a few months ago might still have reference value.

4.3 Interaction Data

In this section, we emphasize understanding how user-item interaction data can impact the performance of evaluation.

Table 4: The performance comparison regarding different interaction types. The best results are highlighted in bold.

Interaction Type	Pearson (%)	MAE	RMSE
Taobao Serendipity			
NA	**11.2366**	**1.3590**	**1.7145**
Click	10.9889	1.3883	1.7831
Purchase	5.0663	1.3591	1.7375
Click & Purchase	9.5309	1.3966	1.7814
Serendipity-2018			
NA	11.4769	0.9742	1.2825
Rating	**12.2663**	**0.9353**	**1.2591**

- Regarding *interaction history length*, we examined the impact of varying k when selecting the users' top-k most recent interactions. Figure 4 shows that LLMs generally achieve superior performance with shorter interaction sequences. This phenomenon can be attributed to two possible reasons: (1) existing LLMs' limited capability in processing and reasoning over lengthy interaction sequences, and (2) more recent interactions having stronger correlation with the user's current preference and intention. We also experimented with time-window-based filtering (e.g., interactions in recent days), but the experiments did not identify clear performance improvements. This might be due to the substantial variance in user interaction frequencies. Some users only recorded a few interactions within a one-day window, while others accumulated thousands over two weeks. This heterogeneity poses a significant challenge for LLM inference.
- For *interaction type*, user interaction with items can occur in different forms. For example, the Taobao dataset includes interactions like clicks and purchases, while the Serendipity-2018 dataset contains user ratings on a 1–5 scale. Due to the sparse nature of purchase behaviors, we increase the interaction history length from the default 10 to 20 for the Taobao dataset. The results are shown in Table 4. It is apparent that in the Taobao dataset, excluding interaction type information produces the best results. This is plausible, as serendipity may have complex relationships with interaction types that vary among users. Without specific fine-tuning, LLMs may struggle to capture these relationships, thereby likely leading to unsatisfactory outcomes. Conversely, in Serendipity-2018, incorporating rating information enhances evaluation performance. This may be because ratings, unlike clicks, offer more precise indicators of user preferences.

4.4 Summary about Auxiliary Data

Our comprehensive analysis yields two interesting findings: (1) Auxiliary data can improve the evaluation of recommendation serendipity by LLMs, but the effectiveness of the data type considered is dependent on the domain. For instance, user data such as curiosity might be more beneficial in the Taobao dataset, and item data such as popularity is crucial for Serendipity-2018. (2) Given LLMs' current limitations in simulating human behavior [30, 42], indiscriminately incorporating all available auxiliary data may not align with user study assessments [31]. This highlights the necessity and importance of our work in thoroughly investigating the roles of different types of auxiliary data for LLM-based evaluators.

5 RQ3: The Power of Multi-LLM

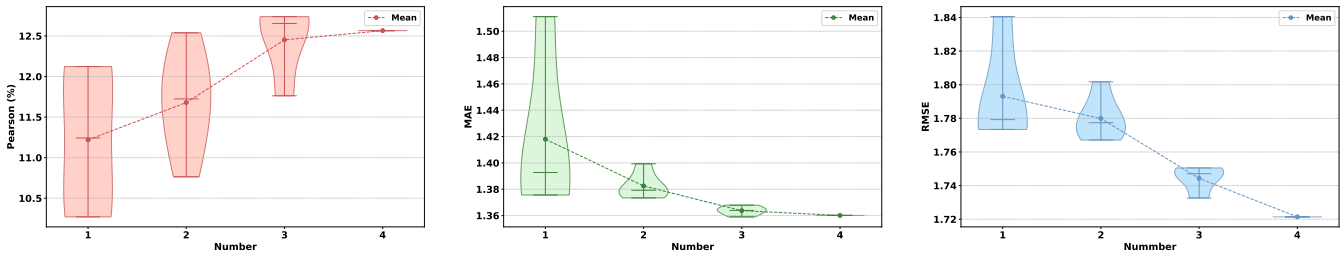
Multi-LLM aims to combine multiple LLMs through specific aggregation rules to determine the final output, which has demonstrated promising results in various evaluation tasks (e.g., explainability evaluation) [24, 37]. Therefore, in our work, we further explore the potential of multi-LLM ensembles to improve serendipity evaluation. Based on our findings from Section 3, we exclude models from the LLaMA family due to their previously demonstrated limitations. Specifically, we employ score averaging as the ensemble strategy, by which the final serendipity score is computed as the arithmetic mean of predictions from multiple LLMs. The results, as depicted in Figure 5 and Figure 6, reveal two major insights:

(1) An ensemble of multiple LLMs has the potential to further enhance serendipity evaluation accuracy. As the number of LLMs used increases, multi-LLM techniques show performance improvements. This suggests that a multi-LLM approach can effectively compensate for the limitations in knowledge and reasoning abilities inherent to a single LLM.
(2) The degree of improvement can be influenced by the specific combination of LLMs used in the ensemble. This is partly because some LLMs, such as those in the Qwen family, possess similar knowledge and reasoning capabilities; thus, integrating them may not yield substantial enhancements. This highlights the critical importance of strategic model selection in constructing effective ensembles.

Furthermore, we conducted a grid search across all possible combinations by integrating Multi-LLM and auxiliary data. Due to space limitations, we only present the best results, as shown in Table 5, where "NA" represents the baseline condition of only using Qwen2.5-14B with the 10 most recent user interactions prior to the target item. Remarkably, we achieved impressive performance in both datasets. Specifically, the Pearson correlation coefficients reach 20.23% and 21.51% on the Taobao and Serendipity-2018 datasets, respectively. These optimal results obtained from the auxiliary data for the two datasets align with our previous analysis (see Section 4), and the multi-LLMs are Qwen2.5-14B, Qwen2.5-72B, and GPT-4. This impressive performance suggests that LLMs may serve as potential evaluators in the future, offering a promising balance between efficiency and effectiveness.

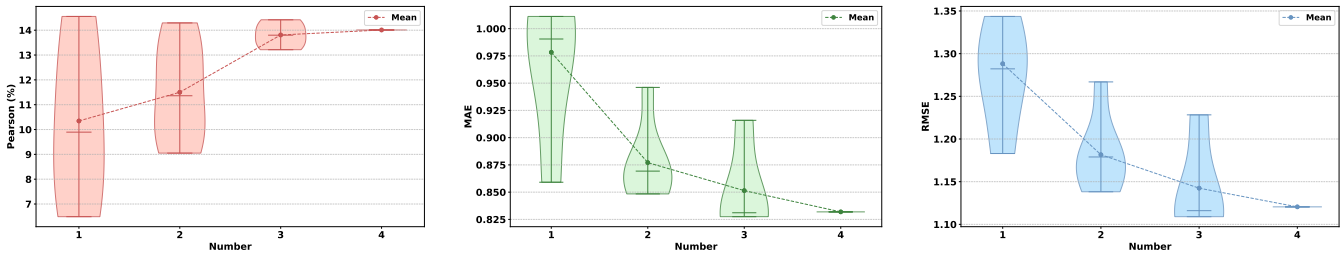
6 Related Work

Serendipity has increasingly captured the interest of the recommender systems community, due to its potential to address the information cocoon and filter bubble issues and enhance user satisfaction. Research has predominantly focused on developing algorithms to strengthen recommendation serendipity. For example, SerRec [23] introduces a novel transfer learning approach, initially using a large dataset to transfer relevance scores, which are then refined for serendipity scores using a smaller dataset. SerenCDR [9] is pioneering in enhancing cross-domain serendipity by utilizing a deep learning model and auxiliary loss to strengthen serendipity



(a) **Pearson correlation coefficient** *(b)* **MAE** *(c)* **RMSE**

Figure 5: Comparison w.r.t. Pearson correlation coefficient, MAE, and RMSE regarding different ensemble sizes on the Taobao.



(a) **Pearson correlation coefficient** *(b)* **MAE** *(c)* **RMSE**

Figure 6: Comparison w.r.t. Pearson correlation coefficient, MAE, and RMSE regarding different ensemble sizes on the Serendipity-2018 dataset.

Table 5: The performance of LLM-based serendipity evaluation by integrating multi-LLM ensemble and auxiliary data. The best results are highlighted in bold, and the second-best results are underlined.

Method	Taobao Serendipity			Serendipity-2018		
	Pearson(%)	MAE	RMSE	Pearson(%)	MAE	RMSE
NA	10.2701	1.4018	1.7836	11.4769	0.9742	1.2825
Auxiliary Data	<u>18.2535</u>	<u>1.3871</u>	<u>1.7505</u>	<u>20.2453</u>	<u>0.8256</u>	<u>1.1164</u>
Multi-LLM & Auxiliary Data	**20.2293**	**1.2960**	**1.6238**	**21.5122**	**0.7486**	**1.0563**

learning within each domain. SerenPrompt [7] is the first to explore the use of prompting LLMs for serendipitous recommendations.

Despite these advancements, evaluating serendipity remains challenging due to its inherently subjective nature. Chen [5] et al. conduct an extensive user study on the Taobao platform, capturing user perceptions of serendipity directly. Binst [4] reviews current evaluation methodologies, identifying that existing metrics lack standardized and validated methods. Tokutake et al. [26] explore the application of LLMs in assessing serendipity. However, their study primarily focuses on validating the evaluation capabilities of LLMs and the effectiveness of the **SOG** metric. While preliminary explorations have been conducted, to our knowledge, no comprehensive and systematic study has addressed the reproducibility of various LLM techniques across diverse product domains and data types for serendipity evaluation.

7 Conclusion

This study investigates the potential of large language models (LLMs) for serendipity evaluation in recommender systems, a task complicated by its subjective and ambiguous nature. Through systematic meta-evaluation on two user-study-validated datasets, we show that even basic zero-shot and few-shot LLMs can match or surpass conventional proxy metrics. Incorporating auxiliary data and multi-LLM strategies further improves alignment with human judgments, with optimal configurations achieving Pearson correlation coefficients above 20%. These findings suggest that LLM-based evaluation has the potential to combine the accuracy of user studies with the efficiency of proxy metrics, offering a promising direction for serendipity evaluation in recommender systems. In the future, we aim to explore more advanced LLM techniques and investigate the explainability of serendipity.

Acknowledgments

This work is supported by Hong Kong Baptist University IG-FNRA Project (RC-FNRA-IG/21-22/SCI/01), Key Research Partnership Scheme (KRPS/23-24/02), and NSFC/RGC Joint Research Scheme (N_HKBU214/24).

References

[1] Josh Achiam, Adler, et al. 2023. GPT-4 Technical Report. *arXiv* (2023).

[2] Zahra Ashktorab, Michael Desmond, Qian Pan, James M Johnson, Martin Santillan Cooper, Elizabeth M Daly, Rahul Nair, Tejaswini Pedapati, Swapnaja Achintalwar, and Werner Geyer. 2024. Aligning Human and LLM Judgments: Insights from EvalAssist on Task-Specific Evaluations and AI-assisted Assessment Strategy Preferences. *arXiv preprint arXiv:2410.00873* (2024).

[3] Anna Bavaresco, Raffaella Bernardi, Leonardo Bertolazzi, Desmond Elliott, Raquel Fernández, Albert Gatt, Esam Ghaleb, Mario Giulianelli, Michael Hanna, Alexander Koller, et al. 2024. LLMs Instead of Human Judges? A Large-Scale Empirical Study Across 20 NLP Evaluation Tasks. *arXiv preprint arXiv:2406.18403* (2024).

[4] Brett Binst. 2024. How to Evaluate Serendipity in Recommender Systems: the Need for a Serendiptionnaire. In *RecSys*. 1335–1341.

[5] Li Chen, Yonghua Yang, Ningxia Wang, Keping Yang, and Quan Yuan. 2019. How Serendipity Improves User Satisfaction with Recommendations? A Large-Scale User Evaluation. In *WWW*. 240–250.

[6] Zhe Fu and Xi Niu. 2023. Modeling Users' Curiosity in Recommender Systems. *TKDD* 18, 1 (2023), 1–23.

[7] Zhe Fu and Xi Niu. 2024. The Art of Asking: Prompting Large Language Models for Serendipity Recommendations. In *ICTIR*. 157–166.

[8] Zhe Fu, Xi Niu, and Mary Lou Maher. 2023. Deep Learning Models for Serendipity Recommendations: A Survey and New Perspectives. *Comput. Surveys* 56, 1 (2023), 1–26.

[9] Zhe Fu, Xi Niu, Xiangcheng Wu, and Ruhani Rahman. 2025. A Deep Learning Model for Cross-Domain Serendipity Recommendations. *TORS* 3, 3 (2025), 1–21.

[10] Zhe Fu, Xi Niu, and Li Yu. 2023. Wisdom of Crowds and Fine-grained Learning for Serendipity Recommendations. In *SIGIR*. 739–748.

[11] Samuel D Gosling, Peter J Rentfrow, and William B Swann Jr. 2003. A Very Brief Measure of the Big-Five Personality Domains. *Journal of Research in personality* 37, 6 (2003), 504–528.

[12] Tonmoy Hasan and Razvan Bunescu. 2023. Topic-Level Bayesian Surprise and Serendipity for Recommender Systems. In *RecSys*. 933–939.

[13] Todd B Kashdan, Matthew W Gallagher, Paul J Silvia, Beate P Winterstein, William E Breen, Daniel Terhar, and Michael F Steger. 2009. The Curiosity And Exploration Inventory-II: Development, Factor Structure, And Psychometrics. *Journal of research in personality* 43, 6 (2009), 987–998.

[14] Denis Kotkov, Joseph A Konstan, Qian Zhao, and Jari Veijalainen. 2018. Investigating Serendipity in Recommender Systems Based on Real User Feedback. In *SAC*. 1341–1350.

[15] Denis Kotkov, Alan Medlar, and Dorota Glowacka. 2023. Rethinking Serendipity in Recommender Systems. In *CHIIR*. 383–387.

[16] Denis Kotkov, Alan Medlar, Triin Kask, and Dorota Glowacka. 2024. The dark matter of serendipity in recommender systems. In *CHIIR*. 108–118.

[17] Denis Kotkov, Jari Veijalainen, and Shuaiqiang Wang. 2020. How Does Serendipity Affect Diversity in Recommender Systems? A Serendipity-Oriented Greedy Algorithm. *Computing* 102 (2020), 393–411.

[18] Haitao Li, Qian Dong, Junjie Chen, Huixue Su, Yujia Zhou, Qingyao Ai, Ziyi Ye, and Yiqun Liu. 2024. LLMs-as-Judges: A Comprehensive Survey on LLM-Based Evaluation Methods. *arXiv preprint arXiv:2412.05579* (2024).

[19] Jiaqi Li, Mengmeng Wang, Zilong Zheng, and Muhan Zhang. 2023. LooGLE: Can Long-Context Language Models Understand Long Contexts? *arXiv preprint arXiv:2311.04939* (2023).

[20] Pan Li, Maofei Que, Zhichao Jiang, Yao Hu, and Alexander Tuzhilin. 2020. PURS: Personalized Unexpected Recommender System for Improving User Satisfaction. In *RecSys*. 279–288.

[21] Xueqi Li, Wenjun Jiang, Weiguang Chen, Jie Wu, Guojun Wang, and Kenli Li. 2020. Directional and Explainable Serendipity Recommendation. In *WWW*. 122–132.

[22] Xiang Liu, Peijie Dong, Xuming Hu, and Xiaowen Chu. 2024. Longgenbench: Long-context Generation Benchmark. *arXiv preprint arXiv:2410.04199* (2024).

[23] Gaurav Pandey, Denis Kotkov, and Alexander Semenov. 2018. Recommending Serendipitous Items Using Transfer Learning. In *CIKM*. 1771–1774.

[24] Bhrij Patel, Souradip Chakraborty, Wesley A Suttle, Mengdi Wang, Amrit Singh Bedi, and Dinesh Manocha. 2024. AIME: AI System Optimization via Multiple LLM Evaluators. *arXiv preprint arXiv:2410.03131* (2024).

[25] Pearl Pu, Li Chen, and Rong Hu. 2011. A User-Centric Evaluation Framework for Recommender Systems. In *Proceedings of the fifth ACM conference on Recommender systems*. 157–164.

[26] Yu Tokutake and Kazushi Okamoto. 2024. Can Large Language Models Assess Serendipity in Recommender Systems? *JACIII* 28, 6 (2024), 1263–1272.

[27] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models. *arXiv preprint arXiv:2307.09288* (2023).

[28] Yu-Min Tseng, Wei-Lin Chen, Chung-Chi Chen, and Hsin-Hsi Chen. 2024. Are Expert-Level Language Models Expert-Level Annotators? *arXiv preprint arXiv:2410.03254* (2024).

[29] Jianling Wang, Haokai Lu, Yifan Liu, He Ma, Yueqi Wang, Yang Gu, Shuzhou Zhang, Ningren Han, Shuchao Bi, Lexi Baugher, et al. 2024. LLMs for User Interest Exploration in Large-Scale Recommendation Systems. In *RecSys*. 872–877.

[30] Lei Wang, Jingsen Zhang, Hao Yang, Zhi-Yuan Chen, Jiakai Tang, Zeyu Zhang, Xu Chen, Yankai Lin, Hao Sun, Ruihua Song, et al. 2025. User Behavior Simulation with Large Language Model-Based Agents. *TOIS* 43, 2 (2025), 1–37.

[31] Ningxia Wang and Li Chen. 2023. How Do Item Features and User Characteristics Affect Users' Perceptions of Recommendation Serendipity? A Cross-Domain Analysis. *UMUAI* 33, 3 (2023), 727–765.

[32] Ningxia Wang, Chen Li, et al. 2020. The Impacts of Item Features and User Characteristics on Users' Perceived Serendipity of Recommendations. In *UMAP*.

[33] Chris Wendler, Veniamin Veselovsky, Giovanni Monea, and Robert West. 2024. Do Llamas Work in English? On the Latent Language of Multilingual Transformers. In *ACL*. 15366–15394.

[34] Yunjia Xi, Muyan Weng, Wen Chen, Chao Yi, Dian Chen, Gaoyang Guo, Mao Zhang, Jian Wu, Yuning Jiang, Qingwen Liu, et al. 2025. Bursting Filter Bubble: Enhancing Serendipity Recommendations with Aligned Large Language Models. *arXiv preprint arXiv:2502.13539* (2025).

[35] An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. 2024. Qwen2.5 Technical Report. *arXiv preprint arXiv:2412.15115* (2024).

[36] Mingwei Zhang, Yang Yang, Rizwan Abbas, Ke Deng, Jianxin Li, and Bin Zhang. 2021. SNPR: A Serendipity-Oriented Next POI Recommendation Model. In *CIKM*. 2568–2577.

[37] Xiaoyu Zhang, Yishan Li, Jiayin Wang, Bowen Sun, Weizhi Ma, Peijie Sun, and Min Zhang. 2024. Large Language Models as Evaluators for Recommendation Explanations. In *RecSys*. 33–42.

[38] Jun Zhao, Zhihao Zhang, Luhui Gao, Qi Zhang, Tao Gui, and Xuanjing Huang. 2024. Llama Beyond English: An Empirical Study on Language Capability Transfer. *arXiv preprint arXiv:2401.01055* (2024).

[39] Pengfei Zhao and Dik Lun Lee. 2016. How Much Novelty Is Relevant? It Depends on Your Curiosity. In *SIGIR*. 315–324.

[40] Yuhan Zhao, Rui Chen, Li Chen, Shuang Zhang, Qilong Han, and Hongtao Song. 2025. From Pairwise to Ranking: Climbing the Ladder to Ideal Collaborative Filtering with Pseudo-Ranking. In *AAAI*, Vol. 39. 13392–13400.

[41] Yuhan Zhao, Rui Chen, Qilong Han, Hongtao Song, and Li Chen. 2024. Unlocking the Hidden Treasures: Enhancing Recommendations with Unlabeled Data. In *RecSys*. 247–256.

[42] Lixi Zhu, Xiaowen Huang, and Jitao Sang. 2024. How Reliable Is Your Simulator? Analysis on the Limitations of Current LLM-Based User Simulators for Conversational Recommendation. In *WWW*. 1726–1732.

[43] Reza Jafari Ziarani and Reza Ravanmehr. 2021. Serendipity in Recommender Systems: A Systematic Literature Review. *Journal of Computer Science and Technology* 36 (2021), 375–396.