Quantifying the Potential to Escape Filter Bubbles: A Behavior-Aware Measure via Contrastive Simulation

Difu Feng[1,2], Qianqian Xu[1*], Zitai Wang[1], Cong Hua[1,2], Zhiyong Yang[2], Qingming Huang[2,3,1*]

[1] Institute of Computing Technology, Chinese Academy of Sciences
[2] School of Computer Science and Technology, University of Chinese Academy of Sciences
[3] Key Laboratory of Big Data Mining and Knowledge Management, University of Chinese Academy of Sciences
{fengdifu24, yangzhiyong21, qmhuang}@ucas.ac.cn, {xuqiangian, wangzitai, huacong23z}@ict.ac.cn

Abstract

Nowadays, recommendation systems have become crucial to online platforms, shaping user exposure by accurate preference modeling. However, such an exposure strategy can also reinforce users' existing preferences, leading to a notorious phenomenon named filter bubbles. Given its negative effects, such as group polarization, increasing attention has been paid to exploring reasonable measures to filter bubbles. However, most existing evaluation metrics simply measure the diversity of user exposure, failing to distinguish between algorithmic preference modeling and actual information confinement. In view of this, we introduce Bubble Escape Potential (BEP), a behavior-aware measure that quantifies how easily users can escape from filter bubbles. Specifically, BEP leverages a contrastive simulation framework that assigns different behavioral tendencies (e.g., positive vs. negative) to synthetic users and compares the induced exposure patterns. This design enables decoupling the effect of filter bubbles and preference modeling, allowing for more precise diagnosis of bubble severity. We conduct extensive experiments across multiple recommendation models to examine the relationship between predictive accuracy and bubble escape potential across different groups. To the best of our knowledge, our empirical results are the first to quantitatively validate the dilemma between preference modeling and filter bubbles. What's more, we observe a counter-intuitive phenomenon that mild random recommendations are ineffective in alleviating filter bubbles, which can offer a principled foundation for further work in this direction.

Code — https://github.com/fengdifu24/bepmetric

1 Introduction

Recommendation systems have become an integral part of online platforms, shaping how users access information in domains such as e-commerce, social media, and news. By modeling user preferences and tailoring content accordingly, these systems help users cope with overwhelming choices. However, this personalization comes at a cost: it often reinforces users' existing interests and behaviors, leading to a well-known phenomenon called the *filter bubble* (Pariser 2011). Within a filter bubble, users are repeatedly exposed to similar content, which may gradually restrict their worldview, amplify bias, and contribute to societal polarization (Bakshy, Messing, and Adamic 2015; Ledwich and Zaitsev 2020; Han et al. 2025).

Given the potential harm of filter bubbles, there has been growing interest in methods to detect and mitigate them. A common approach is to evaluate the diversity of content a user is exposed to, using metrics such as category count, coverage, or entropy (Gao et al. 2023; Piao et al. 2023; Gu et al. 2024). However, most existing methods focus solely on item-side properties and overlook a crucial aspect: user behavior. Recommendation systems are fundamentally interactive—the outcome is shaped not only by the algorithm but also by the user's own actions and preferences. Therefore, solely relying on content diversity fails to distinguish between algorithmic bias and natural user preference.

To incorporate user behavior into the measurement of filter bubble severity, we propose a new insight: **a filter bubble is more severe when a user actively tries to escape it but still fails to see diverse content.** In this light, we propose a novel, behavior-aware metric named **Bubble Escape Potential** (BEP). BEP quantifies how possible it is for users to escape from a filter bubble by comparing user behaviors with contrastive tendencies. Specifically, we design a contrastive simulation framework in which synthetic users exhibit either positive behavior—actively exploring new content—or negative behavior—reinforcing prior preferences. By comparing the exposure patterns generated by the same recommendation model for these two user types, we can decouple the effects of system bias from behavioral tendencies and provide a clearer diagnosis of bubble severity.

To implement this idea, we utilize large language model (LLM) agents (Yao et al. 2023; Wang et al. 2024), which have recently emerged as powerful tools for simulating user interaction. These agents are capable of controlled planning, decision-making, and consistent behavior generation (Park et al. 2023; Xie et al. 2024). Compared to traditional simulators or real-world datasets, LLM agents offer a distinct advantage: they allow us to precisely manipulate user goals and observe system responses under tightly controlled conditions. This makes them ideal for our framework, where accurate control of user behavior is key.

We evaluate BEP across multiple representative recom-

*Corresponding author.

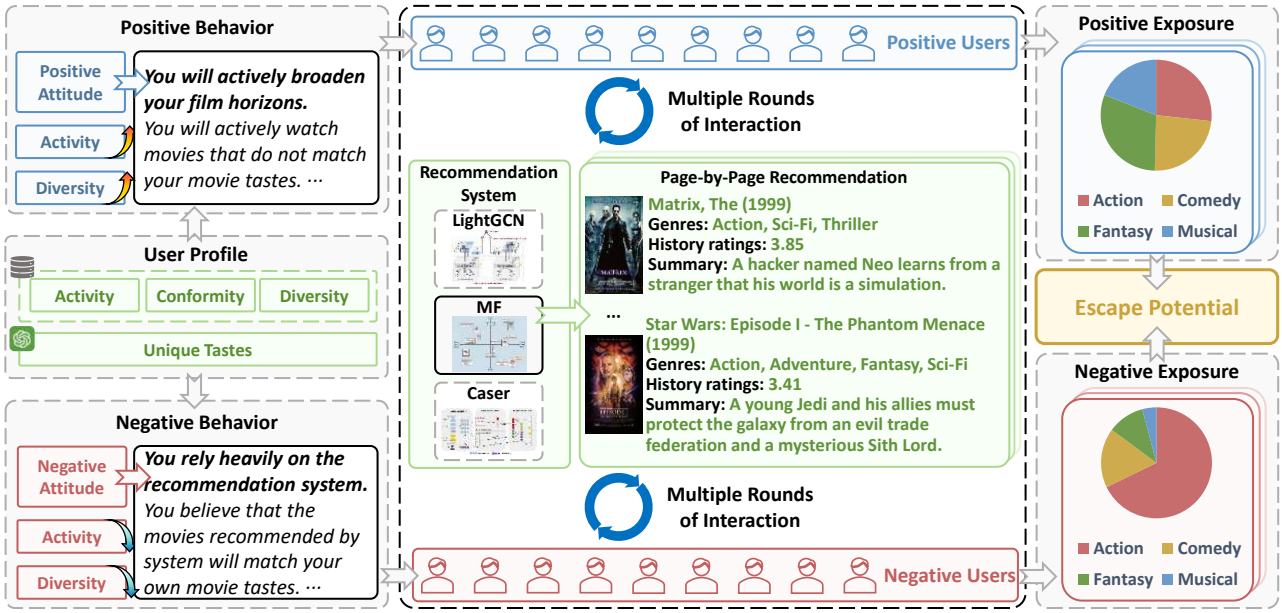


Figure 1: **Overview of the Bubble Escape Potential** (BEP) **evaluation framework.** We simulate two groups of users—**positive** and **negative**—who share user profiles (conformity and tastes). By modifying the user's activity, diversity and different attitudes in prompts, we generate positive users and negative users. Both groups interact with a selected recommendation system across multiple rounds, receiving page-by-page recommendations and making choices. By comparing their exposure, we quantify the system's **Bubble Escape Potential** (BEP).

mendation models and analyze the relationship between predictive accuracy and bubble severity. Our results not only validate BEP's ability to capture the trade-off between accurate preference modeling and information confinement, but also reveal a counter-intuitive finding: mild randomization in recommendation lists does *not* effectively reduce filter bubbles. This insight highlights the need for more principled strategies in mitigating algorithmic confinement.

In summary, our contributions are as follows:

- We introduce **Bubble Escape Potential** (BEP), a novel behavior-aware metric for measuring filter bubble severity by contrastive user behavioral intent.
- We present a contrastive simulation framework using LLM agents to systematically evaluate the influence of user behavior in recommendation scenarios.
- We conduct extensive empirical validation and uncover new insights into the complex trade-offs between personalization, diversity, and user freedom.

2 Related Work

2.1 Filter Bubble

The concept of the *filter bubble* was first introduced and widely spread by Eli Pariser in 2011 (Pariser 2011). Later studies have focused on understanding its causes and finding ways to reduce its impact.

There are three main methods used to study the filter bubble: (1) static datasets (Sukiennik, Gao, and Li 2024), (2) simulating interactions between users and recommendation systems (Anwar, Schoenebeck, and Dhillon 2024), and (3) mathematical modeling (Piao et al. 2023). (Sukiennik, Gao, and Li 2024) finds that the filter bubble becomes stronger as item classification becomes more detailed. (Anwar, Schoenebeck, and Dhillon 2024) distinguishes the filter bubble from homogeneity, arguing that it can involve both high and low inter-user diversity. (Piao et al. 2023) models the formation of the filter bubble using stochastic differential equations. What's more, (Gu et al. 2024) develops an adaptive imitation process to further explore its causes and potential solutions. (Gao et al. 2023) proposes a counterfactual interactive recommendation system that reduces the filter bubble by inferring information overexposure. (Zhang et al. 2024b) introduces a category-based retrieval method using a next-category prediction model to ease the filter bubble effect.

2.2 LLM Agents for User Simulation in RS

With the growing use of large language models (LLMs) in recommendation systems, researchers have started using LLM Agents as simulated users to enrich training data and explore system behaviors. RecAgent (Wang et al. 2023a) is the earliest framework to simulate users using LLM Agents in recommendation systems. Agent4Rec (Zhang et al. 2024a) focuses on simulating real user behavior and modeling feedback from interactions. Recently, more user simulation frameworks have emerged, expanding the user characteristics and improving the alignment with real users (Zhang et al. 2025; Cai et al. 2025; Liu et al. 2025). Some of these frameworks (Wang et al. 2023a; Zhang et al. 2024a) have attempted to simulate the filter bubble effect. However,

these efforts are still limited in depth.

2.3 Recommendation Systems

A recommendation system is an information filtering tool that delivers the most relevant content to a specific user, helping reduce information overload on modern internet platforms. Traditional collaborative filtering methods predict user preferences based on historical data (Sarwar et al. 2001; Koren, Bell, and Volinsky 2009; He et al. 2017; Wang et al. 2019, 2021a). Sequential recommendation focuses on leveraging the temporal order of user-item interactions (Hidasi et al. 2016; Kang and McAuley 2018). More recently, with the rapid development of deep learning (Han et al. 2024), researchers have explored applying LLMs to recommendation systems (Sun et al. 2019; Li et al. 2023).

2.4 Diversified Recommendation

Diversified recommendation has been a critical area of research in the field of recommender systems. It aims to balance relevance and diversity in the recommended items. MMR (Ziegler et al. 2005) is first to optimize both relevance and diversity by iteratively selecting items that maximize diversity. DPP (Kulesza and Taskar 2012) offers a probabilistic approach to model diversity by determinants. So far, diversified recommendations are extensively studied (Steck 2018; Zheng et al. 2021; Liu et al. 2023; Yang et al. 2023; Li et al. 2024; Coppolillo, Manco, and Gionis 2024).

Although our work is based on diversity, there are fundamental differences. First, we consider the role of user behavior within the filter bubble. Second, BEP and diversity metrics differ in key aspects. We well explain it in Sec. 3.3 and Sec. 4.3.

3 Preliminary

3.1 Problem Definition

In a recommendation system R, there are N items denoted as $I = \{i_1, i_2, ..., i_N\}$, categorized into M classes. Each item i_k belongs to a category c_k. For a user group U in R, their interactions are observed over T time periods, indexed by $t \in \{1, 2, ..., T\}$. At each time t, user $u \in U$ receives a list of recommended items from the system, represented as $L_{u,t} \subset I$. The user selects some of these items to interact with, resulting in the interaction set $S_{u,t} \subseteq L_{u,t}$. As t increases, the diversity of information in $L_{u,t}$ is generally expected to decrease.

3.2 Existing Metrics

A common perspective in recent studies (Piao et al. 2023; Wang et al. 2023a; Sukiennik, Gao, and Li 2024; Zhang et al. 2024a; Gao et al. 2023) is that the reduction in diversity of recommended items over time, denoted as $diversity(L_{u,t})$, serves as an indicator of the filter bubble effect. Several metrics have been proposed to quantify $diversity(L_{u,t})$:

- **Standardized information entropy** (Piao et al. 2023): $\tilde{s}_{u,t} = \frac{s_{u,t}}{s_u^*}$, where s_u^* is a user-specific normalization term, and $s_{u,t} = -\sum_{c=1}^{M} f_{u,t}^c log f_{u,t}^c$. Here, $f_{u,t}^c$ denotes the proportion of category c in $L_{u,t}$.

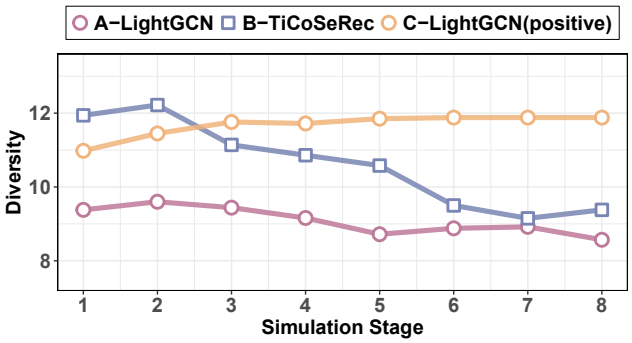


Figure 2: Change in diversity under different settings: A-LightGCN, B-TiCoSeRec, and C-LightGCN(positive) in *ml-1m*. Users' positive behaviors in LightGCN surpasses the gap between different recommendation systems.

- **Category coverage rate** $C_{u,t}$ (Sukiennik, Gao, and Li 2024): $C_{u,t} = \frac{cate(L_{u,t})}{cate(I)}$, where $cate(L)$ is the number of distinct categories in the item set L.
- **Top-1 genre percentage** P_{top-1} (Zhang et al. 2024a): the average proportion of the most frequent genre among the recommended movies.

These metrics are generally consistent in how they capture diversity. Most studies rely on simulation or statistical analysis to track how $diversity(L_{u,t})$ evolves over time (Piao et al. 2023; Wang et al. 2023a; Sukiennik, Gao, and Li 2024; Zhang et al. 2024a,b). By comparing the trend of diversity decline across different recommendation algorithms or user groups, researchers aim to uncover the underlying causes of filter bubbles.

3.3 Limitations of Existing Metrics

Although these methods provide some insights into the severity of filter bubbles, they generally have a substantial limitation. Specifically, the decline in diversity may result from various factors beyond the filter bubble, including the users' own preferences and behavior. **However, they do not clearly distinguish between the influence of filter bubbles and the natural outcome of preference modeling, which may create a misleading correlation between accuracy and filter bubble severity.**

To better understand these concerns, we conduct a case study using simulated users. We present three line graphs in Figure 2. It shows the decline in diversity under different settings: Line A, B, and C. Line A uses LightGCN (He et al. 2020), line B replaces LightGCN with TiCoSeRec (Dang et al. 2023), and line C retains LightGCN but modifies user behaviors to be more positive (details discussed later). We find that the diversity in Group C decreases more slowly than in Group B, suggesting that user behavior significantly affects diversity trends—sometimes even more than the choice of recommendation algorithm.

Hence, it is necessary to explore a more reasonable measure for filter bubbles, which can better decouple filter bubbles from the other factors, which we will elaborate on later.

4 Method

The overall process of our approach is illustrated in Figure 1. In this section, we describe the workflow in three parts. First, we explain how user behaviors are modeled. Second, we describe how interaction data is collected through simulation. Last, we illustrate how our new metric, **Bubble Escape Potential** (BEP), is calculated.

4.1 Behaviors of Users

We conduct two separate user simulations to study the impact of filter bubbles: one with users assigned positive behavior, and the other with users assigned negative behavior. In both simulations, users are modeled using agents powered by large language models (LLMs). The general format of the user simulation prompts follows the settings introduced in (Zhang et al. 2024a).

Each simulated user is defined by two types of characteristics: social traits and unique tastes. Unique tastes are extracted and summarized by LLMs from real users' browsing histories. Social traits are derived from real-world data and include three aspects: a) *Activity*: the frequency and range of a user's interactions. b) *Conformity*: the extent to which a user's ratings align with the average item ratings. c) *Diversity*: a user's tendency to engage with different types or categories of items.

To simulate positive behavior, we assign users a prompt that encourages them to actively escape filter bubbles. We also set their activity and diversity levels to the highest values. The prompt used in this simulation is shown in the blue box. In the prompt box, the specific type of *[item]* varies depending on the type of the item.

To simulate negative behavior, we assign users a prompt that encourages reliance on the recommendation system. We also set their activity and diversity levels to the lowest values. The corresponding prompt is shown in the red box.

4.2 User Simulation

In the user simulation, a group of users U interacts with the recommendation system R for T rounds. During the simulation, the system continuously collects users' interaction data $\{S_{u,t}\}$ and updates its model through re-training.

Prompt of Positive Behavior

You will actively broaden your [item] horizons. You will actively watch [item] that do not match your [item] tastes. Specifically, when you see the recommended list, while you are watching [item] that align with your tastes, you will also watch a few [item] that do not match your tastes.

At the start of the simulation, we perform a cold-start initialization using an interaction set extracted from an official dataset, denoted as A_0. In each round t, the recommendation model is trained on A_{t-1} to produce a new model R'_t. Then, each user u receives a recommendation list $L_{t,u}$ and selects a set of items $S_{t,u}$ to interact with based on predefined behaviors. All interactions in round t are then merged into A_{t-1} to form a new interaction set A_t. After completing all T rounds, we collect all recommendation lists $\{L_{u,t}\}$ and use the previously described method to compute the **Bubble Escape Potential**.

Prompt of Negative Behavior

You rely heavily on the recommendation system. You believe that [item] recommended by the recommendation system will match your own [item] tastes. Specifically, when you see the recommended list, you will watch a few top-ranked [item] first, and then watch other [item] based solely on your own taste.

4.3 Bubble Escape Potential

Given a recommendation system R, we define its corresponding **Bubble Escape Potential** as $\mathsf{BEP}(R)$, which quantifies the probability of users to escape from the filter bubble induced by R. To estimate $\mathsf{BEP}(R)$, we compare the diversity of items recommended to two groups of simulated users: one exhibiting positive behaviors and the other exhibiting negative behaviors.

We begin by assigning all agents in user group U with positive behaviors and simulating their interactions with the recommendation system over T consecutive rounds. The full recommendation history for user u is:

$$L_u = \{L_{u,1}, L_{u,2}, ..., L_{u,T}\} \tag{1}$$

Taking a further step, we define the diversity of recommendations for user u at round t as the number of distinct categories in the list:

$$D_{u,t} = |\{\, c(i) \mid i \in L_{u,t} \,\}| \tag{2}$$

where $c : \mathcal{I} \rightarrow \mathcal{C}$ is a mapping from items to their categories and $\mathcal{C}$ is the set of all possible categories. This gives a sequence of diversity values for each user:

$$D_u = \{D_{u,1}, D_{u,2}, ..., D_{u,T}\} \tag{3}$$

We then reassign the same users with negative behaviors and repeat the simulation for another T rounds, collecting their corresponding diversity values D'_u in the same way.

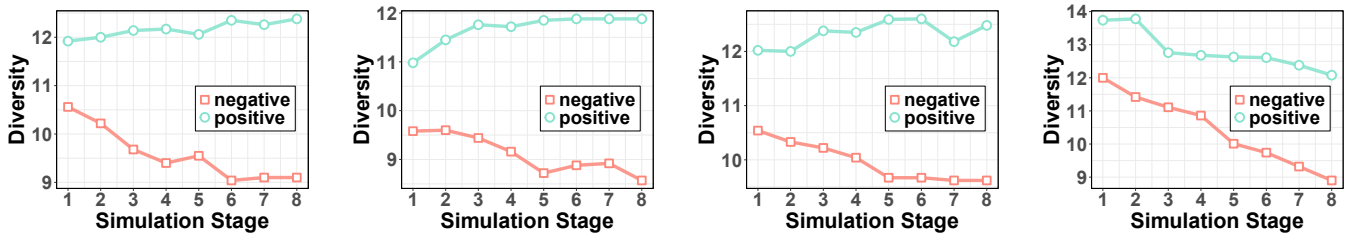
For each round t, the estimated escape potential $\widehat{\mathsf{BEP}}_{R,t}$ is defined as the ratio between the total diversity of the positive-behavior users and that of the negative-behavior users:

$$\widehat{\mathsf{BEP}}_t(R) = \frac{\sum_{u \in U} D_{u,t}}{\sum_{u \in U} D'_{u,t}} \tag{4}$$

Finally, the overall escape potential for system R is calculated by averaging over all rounds:

$$\widehat{\mathsf{BEP}}(R) = \frac{1}{T} \sum_{t=1}^{T} \widehat{\mathsf{BEP}}_{R,t}. \tag{5}$$

Based on the previous description of user behaviors, it is expected to have $\mathsf{BEP}(R) \geq 1$. Moreover, it can be known



(a) (b) (c) (d)

Figure 3: The variation in the diversity of recommendations received by users under different behavioral patterns changes over time as the simulation rounds increase in different recommendations in *ml-1m*. Blue lines represent positive users, while red lines represent negative users. Each of these pictures represents a different recommendation system: (a) MF. (b) LightGCN. (c) DiffRec. (d) TiCoSeRec.

that an important property exists that **the smaller the value of** $\mathsf{BEP}(R)$**, the more severe the filter bubble in recommendation system** R.

In a nutshell, the advantages of our metric can be listed as follows:

- By contrasting users with different behavioral intents, BEP decouples the influence of user preference modeling from system-induced confinement.
- It enables precise diagnosis of filter bubble without relying on assumptions about user intent or model internals.
- Our experiments using BEP are the first to quantitatively validate the inherent tension between accurate preference modeling and the emergence of filter bubbles.
- When the diversity of information received by positive and negative users increases by the same proportion simultaneously, BEP remains unchanged. This property makes it robust to uniform diversification strategies.

5 Experiments

In this section, we present the experimental setup and results. The experiments aim to answer the following research questions:

- **(RQ1)** To what extent can the simulated users approximate the real users?
- **(RQ2)** Do the positive and negative behavior settings significantly influence user actions?
- **(RQ3)** Is it sufficient to set two types of behaviors?
- **(RQ4)** What is the relationship between accuracy and BEP in different recommendation systems?
- **(RQ5)** Can introducing randomness into recommendation strategies help balance accuracy and BEP?
- **(RQ6)** How do user groups with different characteristics vary in their potential to escape filter bubbles?

5.1 Datasets

We conduct our experiments on two real-world datasets:

- **MovieLens-1M** (Harper and Konstan 2015): a widely used benchmark dataset of movies in collaborative filtering. We select the 1M version as *ml-1m*.
- **Amazon-Books** (Ni, Li, and McAuley 2019): a large-scale real-world dataset collected from Amazon's book category, comprising user reviews and ratings.

5.2 Baselines & Metrics

To evaluate our proposed metric comprehensively, we select several representative and recent recommendation models: Random recommendation, BPR-MF (Rendle et al. 2012), LightGCN (He et al. 2020), Caser (Tang and Wang 2018), DiffRec (Wang et al. 2023b), and TiCoSeRec (Dang et al. 2023). We evaluate recommendation accuracy by HR@k, NDCG@k, and MAP. To measure the severity of filter bubbles, we use our proposed metric, Bubble Escape Potential (BEP).

5.3 Implementation Details

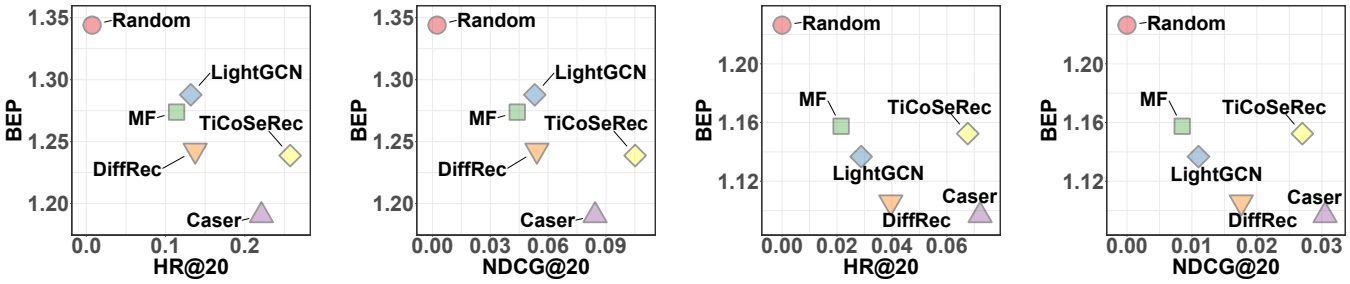
We adopt the *leave-one-out strategy* (Wang et al. 2021b; Liu et al. 2021; Dang et al. 2023) to prepare the test data. For each user's behavior sequence, the last interacted item is used as the test set, while the remaining interactions form the training set. For each user-item pair (u, i_u) in the test set, we record the position of item i_u in the recommendation list L_u generated by the system as p_u.

For user simulation, following Agent4Rec (Zhang et al. 2024a), we first select 1000 users with frequent interactions to form the cold-start dataset. Then, we randomly select 200 users from this group and infer their social characteristics and personal preferences based on their historical behaviors. The global parameters used in the experiment are set as $N = 200, T = 8$. To guarantee reproduction, we use Qwen2.5-14B-Instruct-1M (Yang et al. 2025) as the LLM supporting the user group. For the recommendation models, we use the official implementations and retain their default settings.

5.4 Results

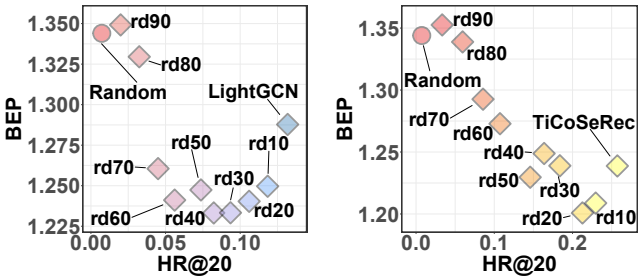
Reality of simulated users (RQ1). In Table 1 and Table 2, we adopt the same method as (Zhang et al. 2024a) to test the alignment between the simulated users and the real users, including the interaction accuracy and the distribution of ratings. Based on these tables, we observe the following:

- The alignment accuracy of the simulated users is around 70%. This validates the reality of users.



(a) (b) (c) (d)

Figure 4: The accuracy and the Bubble Escape Potential (BEP) corresponding to different recommendation systems under *ml-1m* and *Amazon-Books*: (a) HR@20 vs. BEP in *ml-1m*. (b) NDCG@20 vs. BEP in *ml-1m*. (a) HR@20 vs. BEP in *Amazon-Books*. (b) NDCG@20 vs. BEP in *Amazon-Books*. This set of results reveals the *Accuracy-Bubble Dilemma*.



(a) (b)

Figure 5: The accuracy and the Bubble Escape Potential (BEP) corresponding to different recommendation systems after adding randomization under *ml-1m*: (a) LightGCN. (b) TiCoSeRec.

Behavior	Prediction	Recall	Accuracy	F1 Score
positive	0.67	0.76	0.68	0.70
negative	0.72	0.58	0.68	0.63

Table 1: The degree of alignment between the preferences of simulated users and real users for different behaviors.

- The distribution of simulated users' ratings is similar to that of real users ($D_{KL}(P||Q) \approx 0.125$). This indicates that simulated user ratings are similar to real users'.

Distinction of behaviors (RQ2). Figure 3 illustrates how the diversity of recommended items evolves across simulation stages for both positive and negative users under four recommendation systems in *ml-1m* dataset. We will present the precise results of these figures in the supplementary material. From these results, we observe the following:

- Under the influence of negative behaviors, the diversity of information received by users significantly decreases.
- Under the influence of positive behaviors, the diversity of information received by users actually increases (Figure 3(a-c)), and in some cases it is suppressed (Figure 3(d)).
- Regardless of user type, the diversity stabilizes in later stages with only minor fluctuations.

Rating	Ratio of Agent (P)	Ratio of Real Users (Q)
1	0.001	0.056
2	0.054	0.108
3	0.164	0.261
4	0.436	0.349
5	0.345	0.226

Table 2: Rating distribution comparison between agent-simulated users (P) and real users in the *ml-1m* dataset (Q).

Models	HR@20	BEP	BEP-**weak**
Random	0.001	1.35	1.20
TiCoSeRec	0.257	1.24	1.05

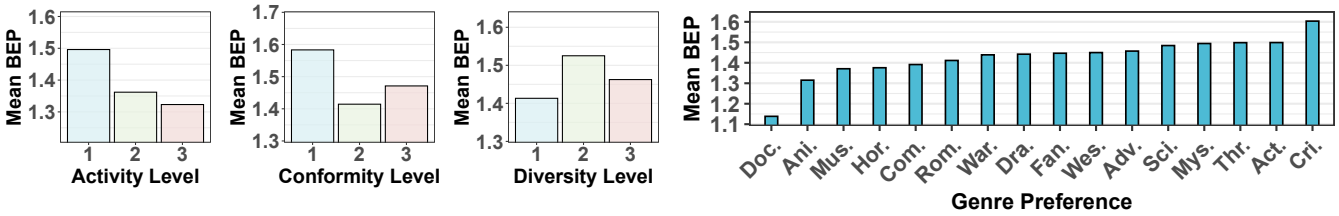
Table 3: Comparison of recommendation performance and Bubble Escape Potential (BEP) using original and weakened behavior (denoted as BEP-weak) on the *ml-1m* dataset.

- Across all methods, positive users consistently receive more diverse recommendations than negative users.

Test of weakened behavior (RQ3). For calculating BEP, just two behaviors are sufficient. Nevertheless, we still design two weakened behaviors (weakly positive and weakly negative), and calculate the BEP in the same form on *ml-1m*. The results of random and TiCoSeRec are in Table 3. We find that their BEP decreases, and BEP of Random is still greater than that of TiCoSeRec. This is in line with our expectations.

Correlations between BEP and accuracy (RQ4). Figure 4 demonstrates the trade-off between recommendation accuracy and the severity of filter bubbles, as quantified by Bubble Escape Potential, in *ml-1m* and *Amazon-Books*. Notably, a lower BEP indicates a more severe filtering effect, meaning that users are more deeply trapped in their personalized content loops. From the results, we observe the following:

- **There is a dilemma between accuracy and filter bubble.** As shown in Figure 4, we observe negative correlation between accuracy and BEP: models that achieve higher HR@20 or NDCG@20 like Caser and TiCoSeRec



(a) (b) (c) (d)

Figure 6: **Comparison of Mean Bubble Escape Potential** (BEP) **across user characteristics.** (a-c): Three intrinsic user profiles—(a) *Activity Level*, (b) *Conformity Level*, and (c) *Diversity Level*. Users with lower conformity and higher diversity tend to have higher BEP, suggesting that if they choose positive actions, it will lead to a much broader expansion of the range of information they can accept. (d): Mean BEP grouped by *Genre Preference*.

tend to produce lower BEP, implying stronger filter bubble effects. In contrast, less accurate models like Random or DiffRec yield higher BEP, suggesting weaker behavioral reinforcement and broader exposure.

- Moreover, a clear structural distinction is observed between different model types. Non-sequential models (MF, DiffRec, LightGCN) are clustered in the upper-left regions of the plots, while sequential models (Caser, TiCoSeRec) are in the lower-right. This indicates that sequential models generally provide higher accuracy but at the cost of more severe filter bubble formation. One possible explanation is that sequential models place greater emphasis on recent user behavior sequences, potentially narrowing the diversity of exposed content and overlooking long-term or global user preferences.

Impact of introducing randomness on recommendation systems (RQ5). To investigate the effect of controlled noise on filter bubble severity, we modify the output of two representative models, LightGCN and TiCoSeRec, by randomly replacing a portion k ($k = 10\%, 20\%, 30\%, ..., 90\%$) of their recommendation lists with items sampled randomly. Figure 5(a) & (b) track the models' trajectories on the accuracy versus BEP plots as the level of randomness increases. The results reveal the following observations:

- The impact of randomness is non-monotonic. As randomness k increases to 30%, BEP drops slightly. Surprisingly, the BEP reaches its *lowest* point around $k = 30\%$, implying that small-scale randomization may inadvertently reinforce personalization biases. However, as randomness continues to increase, BEP begins to rise, returning to its original level near $k = 70\%$ and eventually surpassing it. At $k = 80\% \sim 90\%$, the models approach the performance of a fully random recommender. These suggest that introducing randomness does not help recommendation systems better balance the prevention of filter bubble and accuracy.

The Bubble Escape Potential of different user groups (RQ6). Based on BEP, we analyze how user characteristics affect filter bubbles. For unique tastes of users, we match the corresponding keywords of each genre to form *Genre Preferences*. Then, for each genre preference, we calculate the average value of the BEP of all users with it. Figure 6 shows BEP of users with different *Genre Preferences*. The results reveal the following observations:

- Specifically, active users (higher activity level) demonstrate low BEP values, indicating that systems tend to restrict exposure even when users frequently interact with the platform. Similarly, users with narrow interests (lower diversity level) also experience more difficulty escaping the filter bubble, as indicated by lower BEP scores. These results suggest that user effort alone is not sufficient to overcome algorithmic confinement.

- Regarding the conformity level, there is no obvious correlation with BEP. Considering the definition of conformity, this is in line with common sense.

- Furthermore, we observe notable differences in BEP across genre preferences. As shown in Figure 6 (d), users favoring niche or high-engagement genres such as thriller (Thr.), action (Act.), and crime (Cri.) exhibit significantly higher BEP than those favoring documentaries or animations, resembling a long-tail trend (Wang et al. 2023c; Yang et al. 2024; Li et al. 2025; Wang et al. 2025). This implies that the underlying content ecosystem also plays a role in how filter bubbles form and persist, reinforcing the importance of modeling both user behavior and item characteristics in filter bubble analysis.

6 Conclusion

This paper offers a novel perspective on understanding and mitigating filter bubbles in recommendation systems by introducing the metric of bubble escape potential (BEP). Unlike traditional metrics that are entangled with user preference modeling, our metric provides a behavior-independent, quantitative approach to assess the severity of filter bubble. Through empirical analysis, we demonstrate how different recommendation systems vary in their tendency to create filter bubbles and explore the potential of random recommendation strategies to alleviate this issue. These findings advance understanding of the accuracy–bubble dilemma and provide a foundation for developing more inclusive, socially responsible recommendation systems.

Acknowledgements

This work was supported in part by National Natural Science Foundation of China: 62525212, 62236008, 62441232, U21B2038, U23B2051, 92370102, and 62502500, in part by Youth Innovation Promotion Association CAS, in part by the Strategic Priority Research Program of the Chinese Academy of Sciences, Grant No. XDB0680201, in part by the China National Postdoctoral Program for Innovative Talents under Grant BX20240384, in part by Beijing Natural Science Foundation under Grant No. L252144, in part by General Program of the Chinese Postdoctoral Science Foundation under Grant No. 2025M771558, and in part by the Fundamental Research Funds for the Central Universities.

References

Anwar, M. S.; Schoenebeck, G.; and Dhillon, P. S. 2024. Filter bubble or homogenization? disentangling the long-term effects of recommendations on user consumption patterns. In *International World Wide Web Conference*, 123–134.

Bakshy, E.; Messing, S.; and Adamic, L. A. 2015. Exposure to ideologically diverse news and opinion on Facebook. *Science*, 1130–1132.

Cai, S.; Zhang, J.; Bao, K.; Gao, C.; Wang, Q.; Feng, F.; and He, X. 2025. Agentic feedback loop modeling improves recommendation and user simulation. In *International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2235–2244.

Coppolillo, E.; Manco, G.; and Gionis, A. 2024. Relevance Meets Diversity: A User-Centric Framework for Knowledge Exploration Through Recommendations. In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 490–501.

Dang, Y.; Yang, E.; Guo, G.; Jiang, L.; Wang, X.; Xu, X.; Sun, Q.; and Liu, H. 2023. TiCoSeRec: Augmenting data to uniform sequences by time intervals for effective recommendation. *IEEE Transactions on Knowledge and Data Engineering*, 2686–2700.

Gao, C.; Wang, S.; Li, S.; Chen, J.; He, X.; Lei, W.; Li, B.; Zhang, Y.; and Jiang, P. 2023. CIRS: Bursting filter bubbles by counterfactual interactive recommender system. *ACM Transactions on Information Systems*, 1–27.

Gu, M.; Zhao, T.-F.; Yang, L.; Wu, X.; and Chen, W.-N. 2024. Modeling Information Cocoons in Networked Populations: Insights From Backgrounds and Preferences. *IEEE Transactions on Computational Social Systems*, 4497–4510.

Han, B.; Xu, Q.; Bao, S.; Yang, Z.; Zi, K.; and Huang, Q. 2025. LightFair: Towards an Efficient Alternative for Fair T2I Diffusion via Debiasing Pre-trained Text Encoders. In *Conference on Neural Information Processing Systems*.

Han, B.; Xu, Q.; Yang, Z.; Bao, S.; Wen, P.; Jiang, Y.; and Huang, Q. 2024. AUCSeg: AUC-oriented Pixel-level Long-tail Semantic Segmentation. In *Conference on Neural Information Processing Systems*, 126863–126907.

Harper, F. M.; and Konstan, J. A. 2015. The MovieLens Datasets: History and Context. *ACM Transactions on Interactive Intelligent Systems*, 1–19.

He, X.; Deng, K.; Wang, X.; Li, Y.; Zhang, Y.; and Wang, M. 2020. Lightgcn: Simplifying and powering graph convolution network for recommendation. In *International ACM SIGIR Conference on Research and Development in Information Retrieval*, 639–648.

He, X.; Liao, L.; Zhang, H.; Nie, L.; Hu, X.; and Chua, T.-S. 2017. Neural collaborative filtering. In *International World Wide Web Conference*, 173–182.

Hidasi, B.; Karatzoglou, A.; Baltrunas, L.; and Tikk, D. 2016. Session-based recommendations with recurrent neural networks. In *International Conference on Learning Representations*.

Kang, W.-C.; and McAuley, J. 2018. Self-attentive sequential recommendation. In *IEEE International Conference on Data Mining*, 197–206.

Koren, Y.; Bell, R.; and Volinsky, C. 2009. Matrix factorization techniques for recommender systems. *Computer*, 30–37.

Kulesza, A.; and Taskar, B. 2012. Determinantal Point Processes for Machine Learning. *Foundations and Trends in Machine Learning*, 123–286.

Ledwich, M.; and Zaitsev, A. 2020. Algorithmic extremism: Examining YouTube's rabbit hole of radicalization. *First Monday*.

Li, F.; Si, X.; Tang, S.; Wang, D.; Han, K.; Han, B.; Zhou, G.; Song, Y.; and Chen, H. 2024. Contextual Distillation Model for Diversified Recommendation. In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 5307–5316.

Li, J.; Zhang, W.; Wang, T.; Xiong, G.; Lu, A.; and Medioni, G. 2023. GPT4Rec: A generative framework for personalized recommendation and user interests interpretation. In *International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1774–1784.

Li, S.; Xu, Q.; Yang, Z.; Wang, Z.; Zhang, L.; Cao, X.; and Huang, Q. 2025. Focal-SAM: Focal Sharpness-Aware Minimization for Long-Tailed Classification. In *International Conference on Machine Learning*.

Liu, J.; Gu, S.; Li, D.; Zhang, G.; Han, M.; Gu, H.; Zhang, P.; Lu, T.; Shang, L.; and Gu, N. 2025. AgentCF++: Memory-enhanced LLM-based Agents for Popularity-aware Cross-domain Recommendations. *International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2566–2571.

Liu, W.; Xi, Y.; Qin, J.; Dai, X.; Tang, R.; Li, S.; Zhang, W.; and Zhang, R. 2023. Personalized Diversification for Neural Re-ranking in Recommendation. In *IEEE International Conference on Data Engineering*, 802–815.

Liu, Z.; Chen, Y.; Li, J.; Yu, P. S.; McAuley, J.; and Xiong, C. 2021. Contrastive self-supervised sequential recommendation with robust augmentation. *CoRR*, abs/2108.06479.

Ni, J.; Li, J.; and McAuley, J. 2019. Justifying Recommendations using Distantly-Labeled Reviews and Fine-Grained Aspects. In *Conference on Empirical Methodsin Natural Language Processing*, 188–197.

Pariser, E. 2011. *The filter bubble: What the Internet is hiding from you*. penguin UK.

Park, J. S.; O'Brien, J.; Cai, C. J.; Morris, M. R.; Liang, P.; and Bernstein, M. S. 2023. Generative agents: Interactive simulacra of human behavior. In *ACM Symposium on User Interface Software and Technology*, 1–22.

Piao, J.; Liu, J.; Zhang, F.; Su, J.; and Li, Y. 2023. Human–AI adaptive dynamics drives the emergence of information cocoons. *Nature Machine Intelligence*, 1214–1224.

Rendle, S.; Freudenthaler, C.; Gantner, Z.; and Schmidt-Thieme, L. 2012. BPR: Bayesian personalized ranking from implicit feedback. *CoRR*, abs/1205.2618.

Sarwar, B.; Karypis, G.; Konstan, J.; and Riedl, J. 2001. Item-based collaborative filtering recommendation algorithms. In *International World Wide Web Conference*, 285–295.

Steck, H. 2018. Calibrated Recommendations. In *ACM Conference on Recommender Systems*, 154–162.

Sukiennik, N.; Gao, C.; and Li, N. 2024. Uncovering the deep filter bubble: narrow exposure in short-video recommendation. In *International World Wide Web Conference*, 4727–4735.

Sun, F.; Liu, J.; Wu, J.; Pei, C.; Lin, X.; Ou, W.; and Jiang, P. 2019. BERT4Rec: Sequential recommendation with bidirectional encoder representations from transformer. In *ACM International Conference on Information and Knowledge Management*, 1441–1450.

Tang, J.; and Wang, K. 2018. Personalized top-n sequential recommendation via convolutional sequence embedding. In *ACM International Conference on Web Search and Data Mining*, 565–573.

Wang, L.; Ma, C.; Feng, X.; Zhang, Z.; Yang, H.; Zhang, J.; Chen, Z.; Tang, J.; Chen, X.; Lin, Y.; et al. 2024. A survey on large language model based autonomous agents. *Frontiers of Computer Science*, 186345.

Wang, L.; Zhang, J.; Yang, H.; Chen, Z.; Tang, J.; Zhang, Z.; Chen, X.; Lin, Y.; Song, R.; Zhao, W. X.; et al. 2023a. User behavior simulation with large language model based agents. In *ACM Transactions on Information Systems*.

Wang, W.; Xu, Y.; Feng, F.; Lin, X.; He, X.; and Chua, T.-S. 2023b. Diffusion recommender model. In *International ACM SIGIR Conference on Research and Development in Information Retrieval*, 832–841.

Wang, Z.; Huang, Y.; Dou, Z.; and Wen, J.-R. 2019. Adversarial Preference Learning with Pairwise Comparisons. In *ACM International Conference on Multimedia*, 656–664.

Wang, Z.; Xu, Q.; Yang, Z.; Cao, X.; and Huang, Q. 2021a. Implicit Feedbacks are Not Always Favorable: Iterative Relabeled One-Class Collaborative Filtering against Noisy Interactions. In *ACM International Conference on Multimedia*, 3070–3078.

Wang, Z.; Xu, Q.; Yang, Z.; He, Y.; Cao, X.; and Huang, Q. 2023c. A Unified Generalization Analysis of Re-Weighting and Logit-Adjustment for Imbalanced Learning. In *Conference on Neural Information Processing Systems*, 48417–48430.

Wang, Z.; Xu, Q.; Yang, Z.; Xu, Z.; Zhang, L.; Cao, X.; and Huang, Q. 2025. A Unified Perspective for Loss-Oriented Imbalanced Learning via Localization. In *IEEE Conference on Computer Vision and Pattern Recognition*, 1–19.

Wang, Z.; Zhang, J.; Xu, H.; Chen, X.; Zhang, Y.; Zhao, W. X.; and Wen, J.-R. 2021b. Counterfactual data-augmented sequential recommendation. In *International ACM SIGIR Conference on Research and Development in Information Retrieval*, 347–356.

Xie, C.; Chen, C.; Jia, F.; Ye, Z.; Lai, S.; Shu, K.; Gu, J.; Bibi, A.; Hu, Z.; Jurgens, D.; et al. 2024. Can Large Language Model Agents Simulate Human Trust Behavior? In *Conference on Neural Information Processing Systems*.

Yang, A.; Yu, B.; Li, C.; Liu, D.; Huang, F.; Huang, H.; Jiang, J.; Tu, J.; Zhang, J.; Zhou, J.; et al. 2025. Qwen2.5-1M Technical Report. *CoRR*, abs/2501.15383.

Yang, L.; Wang, S.; Tao, Y.; Sun, J.; Liu, X.; Yu, P. S.; and Wang, T. 2023. DGRec: Graph Neural Network for Recommendation with Diversified Embedding Generation. In *ACM International Conference on Web Search and Data Mining*, 661–669.

Yang, Z.; Xu, Q.; Wang, Z.; Li, S.; Han, B.; Bao, S.; Cao, X.; and Huang, Q. 2024. Harnessing Hierarchical Label Distribution Variations in Test Agnostic Long-tail Recognition. In *International Conference on Machine Learning*, 56624–56664.

Yao, S.; Zhao, J.; Yu, D.; Du, N.; Shafran, I.; Narasimhan, K.; and Cao, Y. 2023. React: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations*.

Zhang, A.; Chen, Y.; Sheng, L.; Wang, X.; and Chua, T.-S. 2024a. On generative agents in recommendation. In *International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1807–1817.

Zhang, T.; Yang, L.; Xiao, Z.; Jiang, W.; and Ning, W. 2024b. On practical diversified recommendation with controllable category diversity framework. In *International World Wide Web Conference*, 255–263.

Zhang, Z.; Liu, S.; Liu, Z.; Zhong, R.; Cai, Q.; Zhao, X.; Zhang, C.; Liu, Q.; and Jiang, P. 2025. Llm-powered user simulator for recommender system. In *AAAI Conference on Artificial Intelligence*, 13339–13347.

Zheng, Y.; Gao, C.; Chen, L.; Jin, D.; and Li, Y. 2021. DGCN: Diversified Recommendation with Graph Convolutional Networks. In *Proceedings of the Web Conference*, 401–412.

Ziegler, C.-N.; McNee, S. M.; Konstan, J. A.; and Lausen, G. 2005. Improving Recommendation Lists through Topic Diversification. In *the International World Wide Web Conference*, 22–32.