

On Generative Agents in Recommendation

An Zhang*
National University of Singapore
Singapore
anzhang@u.nus.edu

Yuxin Chen*
National University of Singapore
Singapore
e1143404@u.nus.edu

Leheng Sheng*
Tsinghua University
Beijing, China
chenglh22@mails.tsinghua.edu.cn

Xiang Wang†
University of Science and Technology
of China
Hefei, China
xiangwang1223@gmail.com

Tat-Seng Chua
National University of Singapore
Singapore
dcscs@nus.edu.s

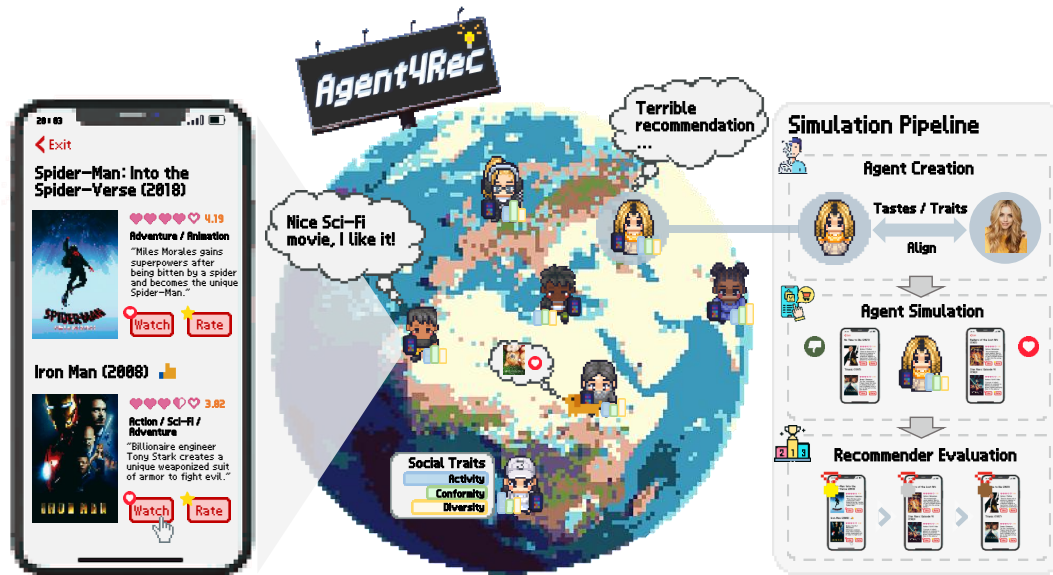


Figure 1: Illustration of Agent4Rec, a user simulator with 1,000 LLM-empowered generative agents in the movie recommendation scenario. These agents are initialized from the MovieLens-1M [15] dataset, embodying varied social traits and preferences. Each agent interacts with personalized movie recommendations in a page-by-page manner and undertakes various actions such as watching, rating, evaluating, exiting, and interviewing. With Agent4Rec, we would like to explore the potential of LLM-empowered generative agents in simulating the behavior of genuine, independent humans in recommendation environments.

ABSTRACT

Recommender systems are the cornerstone of today’s information dissemination, yet a disconnect between offline metrics and online performance greatly hinders their development. Addressing this challenge, we envision a recommendation simulator, capitalizing on recent breakthroughs in human-level intelligence exhibited by

Large Language Models (LLMs). We propose **Agent4Rec**, a user simulator in recommendation, leveraging LLM-empowered generative agents equipped with user profile, memory, and actions modules specifically tailored for the recommender system. In particular, these agents’ profile modules are initialized using real-world datasets (e.g., MovieLens, Steam, Amazon-Book), capturing users’ unique tastes and social traits; memory modules log both factual and emotional memories and are integrated with an emotion-driven reflection mechanism; action modules support a wide variety of behaviors, spanning both taste-driven and emotion-driven actions. Each agent interacts with personalized recommender models in a page-by-page manner, relying on a pre-implemented collaborative filtering-based recommendation algorithm. We delve into both the capabilities and limitations of Agent4Rec, aiming to explore an essential research question: “To what extent can LLM-empowered generative agents faithfully simulate the behavior of real, autonomous

* equal contribution.

† Xiang Wang is the corresponding author, also affiliated with the Institute of Dataspace, Hefei Comprehensive National Science Center.



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

humans in recommender systems?” Extensive and multi-faceted evaluations of Agent4Rec highlight both the alignment and deviation between agents and user-personalized preferences. Beyond mere performance comparison, we explore insightful experiments, such as emulating the filter bubble effect and discovering the underlying causal relationships in recommendation tasks. Our codes are available at <https://github.com/LehengTHU/Agent4Rec>.

CCS CONCEPTS

• **Information systems** → **Recommender systems**.

KEYWORDS

Recommender System, Large Language Model, Generative Agents

ACM Reference Format:

An Zhang*, Yuxin Chen*, Leheng Sheng*, Xiang Wang†, and Tat-Seng Chua. 2024. On Generative Agents in Recommendation. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '24)*, July 14–18, 2024, Washington, DC, USA. ACM, New York, NY, USA, 11 pages. <https://doi.org/10.1145/3626772.3657844>

1 INTRODUCTION

Recommender systems play a pivotal role in contemporary information dissemination, actively shaping individual preferences and cognitive processes [29, 63]. Despite their great success and widespread adoption, the conventional supervised recommendation approach falls short, as evidenced by the significant gap between offline metrics and online performance [64, 67]. This disconnect hinders the integration of academic research into real-world recommendation deployments, acting as a bottleneck for the field’s future advancements [13, 23]. Imagine a paradigm where a configurable simulation platform for recommender systems exists — one that faithfully captures user intent and encodes human cognitive mechanisms. Such a simulator undoubtedly has the potential to revolutionize traditional research paradigms in recommendations, offering an innovative pathway for data collection, recommender evaluation, and algorithmic development [6, 20, 50].

Recent strides in Large Language Models (LLMs), with their impressive capability [40, 47, 58] and profound comprehension of time and space [14, 53, 56], underscore the promise of the recommendation simulator paradigm. Specifically, LLMs serve as the foundational architecture in the development of generative agents [9, 42, 44]. These agents are then integrated into a recommendation environment, taking on the virtual users in simulators. Yet, in this promising research direction, developing a reliable simulator that faithfully mirrors personalized user preferences is non-trivial [54, 69]. Consequently, harnessing the potential of LLM-empowered generative agents by designing modules tailored for recommendation to emulate human behavior has become a key research focus.

To bridge this gap, we introduce **Agent4Rec** — a general user simulator in recommendation scenarios, which consists of two core facets: LLM-empowered generative agents and recommendation environment (*cf.* Figure 2). From the user’s perspective, we simulate 1,000 LLM-empowered generative agents per recommendation scenario, where each agent is initialized based on real-world datasets and composed of three essential modules: the user profile, memory, and action modules. The profile module functions as a repository for

personalized social traits and historical preferences [38], facilitating the alignment of user portraits with genuine human characteristics. The memory module records past viewing behaviors, system interactions, and emotional memories (*i.e.*, user feelings and fatigue levels) in natural language, enabling information retrieval, preference accumulation, and emotion-driven reflection in a coherent manner. The action module empowers these agents to interact directly with the recommendation environment, including taste-driven actions (*i.e.*, viewing or ignoring recommended movies, rating, generating post-viewing feelings), and emotion-driven actions (*i.e.*, exiting the system, evaluating recommendation lists, and expressing human-understandable comments). From the perspective of the recommender system simulation, items are recommended by a predetermined recommendation algorithm, sequenced in a page-by-page format to emulate real-world recommendation platforms. On the one hand, our simulator predominantly integrates collaborative filtering-based recommendation strategies, encompassing random, most popular, Matrix Factorization (MF) [24], LightGCN [16], and MultVAE [28]. On the other hand, we architect the simulator with extensibility as a core principle. By providing open interfaces, we empower researchers and practitioners to effortlessly deploy any recommendation algorithm of their choice.

To systematically evaluate the effectiveness and limitations of our proposed Agent4Rec, we conduct comprehensive experiments from both the user’s and recommender system’s perspectives. From the user’s standpoint, our primary focus lies in assessing the degree of agent alignment. Specifically, we evaluate to what extent the agent can ensure the coherence of the true user’s social traits, personality, and preferences using a variety of metrics and statistical tests. On the recommender system simulation front, we evaluate various recommenders configured with different algorithms. Our evaluation metrics span multiple dimensions, including the average number of items recommended and watched by users, average user ratings, user engagement time, and overall user satisfaction. In parallel, the agent feedback collected from the simulator serves as augmented data, enabling iterative training and refinement of the recommendation strategies. Subsequently, we assess the feedback-driven recommender enhancement using standardized offline metrics. This dual approach, combining simulation feedback with traditional offline evaluation, ensures a comprehensive assessment of recommendation algorithms.

To explore the potential of the simulator in investigating unresolved challenges in recommendation tasks, we undertake two experiments. In the first experiment, we emulate the filter bubble effect within the simulator — a scenario where users are consistently exposed to similar or reinforcing content, resulting in a reduction in item attribute diversity [10, 27]. This investigation aims to understand the extent to which feedback loops can amplify such centralized recommendation phenomena. Additionally, we utilize the simulator as a data collection tool to pioneer a data-oriented causal discovery [17, 62]. This approach yields a robust causal graph, enabling us to unveil the intricate latent causal relationships that infer the data generation progress in recommender systems [11].

Our main contributions are summarized as follows:

- We develop Agent4Rec, a general recommendation simulator utilizing LLM-empowered agents to emulate and deduce user-personalized preferences and behavior patterns. These agents, with their carefully designed modules tailored to recommendation, enable the emulation of human cognitive mechanisms.
- We delve into both the capabilities and limitations of Agent4Rec by conducting extensive evaluations for generative agent-based simulation in recommender systems. We employ statistical metrics and tests for user alignment evaluation and propose a dual parallel evaluation considering both offline performance and simulation feedback.
- Using Agent4Rec as a data collection tool, we replicate a pervasive issue – the filter bubble effect—and unveil the underlying causal relationships embedded within recommender system scenarios. These observations showcase the potential of Agent4Rec to inspire new research directions.

We believe that Agent4Rec stands at the intersection of cutting-edge technology and the challenges of recommender systems, offering an experimental platform for insights that will inspire more work in this research direction.

2 AGENT4REC

Agent4Rec, as a user simulation in recommendation scenarios, is expected to accurately mirror user behaviors, effectively forecast long-term user preferences, and systematically evaluate recommendation algorithms by leveraging the human-like capabilities of LLM-empowered generative agents. To achieve this goal, two core facets are considered: (1) designing agent architectures that faithfully mimic user personalized preferences and human cognitive reasoning, and (2) constructing a recommendation environment that ensures its reliability, extensibility, and adaptability. Figure 2 demonstrated the overview of Agent4Rec’s framework, which is developed by modifying LangChain, with all agents being powered by the gpt-3.5-turbo version of ChatGPT.

Task Formulation. Given a user $u \in \mathcal{U}$ and an item $i \in \mathcal{I}$, let $y_{ui} = 1$ denote that user u has interacted with item i , and subsequently rated it with $r_{ui} \in \{1, 2, 3, 4, 5\}$. Conversely, $y_{ui} = 0$ indicates that the user has not adopted the item. The quality of each item i can be represented by $R_i \doteq \frac{1}{\sum_{u \in \mathcal{U}} y_{ui}} \sum_{u \in \mathcal{U}} y_{ui} \cdot r_{ui}$, while its popularity is denoted by P_i . Additionally, the genre set of the item is given by $G_i \subset \mathcal{G}$. The simulator’s overarching goal is to faithfully distill the human genuine preferences such as \hat{y}_{ui} and \hat{r}_{ui} of user u for an unseen recommended item i .

2.1 Agent Architecture

Generative agents in Agent4Rec, utilizing LLM as its foundational architecture, refine their capabilities with three specialized modules tailored for recommendation scenarios: a profile module, a memory module, and an action module. Specifically, to emulate personalized genuine human behaviors, each agent integrates a user profile module to reflect individualized social traits and preferences. Furthermore, drawing inspiration from human cognitive processes, agents are equipped with memory and action modules, enabling them to store, retrieve, and apply past interactions and emotions to generate behaviors in a coherent manner.

2.1.1 Profile Module. In the domain of personalized recommendation simulation, the user profile module stands as a cornerstone, playing a crucial role in the efficacy of agents’ alignment with genuine human behaviors. To lay a reliable foundation for the generative agent’s subsequent simulations and evaluations, the benchmark dataset (e.g., MovieLens-1M [15], Steam [22], Amazon-Book [37]) is used for initialization. Each agent’s profile contains two components: social traits and unique tastes [33, 68].

Social traits encompass three key traits capturing the element of an individual’s personality and characteristics in recommendation scenarios, that is activity, conformity, and diversity. Activity quantifies the frequency and breadth of a user’s interactions with recommended items, delineating between users who extensively watch and rate a number of items and those who confine themselves to a minimal set [38]. The activity trait for user u can be mathematically articulated as: $T_{act}^u \doteq \sum_{i \in \mathcal{I}} y_{ui}$. Conformity delves into how closely a user’s ratings align with average item ratings, drawing a distinction between users with unique perspectives and those whose opinions closely mirror popular sentiments [3, 70]. For user u , the conformity trait is defined as: $T_{conf}^u \doteq \frac{1}{\sum_{i \in \mathcal{I}} y_{ui}} \sum_{i \in \mathcal{I}} y_{ui} \cdot |r_{ui} - R_i|^2$. Diversity reflects the user’s proclivity toward a diverse range of item genres or their inclination toward specific genres [2]. The diversity trait for user u is formulated as: $T_{div}^u \doteq |\cup_{i \in \{y_{ui}=1\}} G_i|$. Users typically exhibit specific distributions among these social traits (e.g., the long-tail distribution of user activity [66]). Accordingly, we segment them into three uneven tiers based on each respective trait.

To encode users’ personalized preferences in natural language, we randomly select 25 items for each user from their viewing history. Items rated 3 or above are categorized as ‘like’ by the user, while those rated below 3 are deemed ‘dislike’. Leveraging ChatGPT, we then distill and summarize the unique tastes and rating patterns the user exhibited. These personalized item tastes are incorporated as the second component into user profiles.

We underscore that in Agent4Rec, certain personal identifiers, such as name, gender, age, and occupation, are intentionally obscured to guarantee widespread applicability and address privacy concerns [7]. Although these attributes may be instrumental in shaping other types of agents, within the realm of recommendation, they do not dominate users’ item preferences. Such preferences can be adequately deduced from historical viewing records, rating patterns, and insights embedded in user interactions [25].

2.1.2 Memory Module. Humans retain diverse memories, bifurcating mainly into factual and emotional categories. Of these, emotional memories constitute the core of personal history and exert a stronger influence on decision-making [26]. Although pioneering studies on agents have detailed the architecture of memory, providing foundational blueprints for subsequent explorations, the emotional memories have been largely overlooked [42, 54].

In our Agent4Rec, we embed a specialized memory module within each generative agent, logging both factual and emotional memories. Tailored for the recommendation task, factual memories encapsulate interactive behaviors within the recommender system, while emotional memories capture the psychological feeling stemming from these interactions. Specifically, factual memory mainly contains the list of recommended items, along with user feedback.

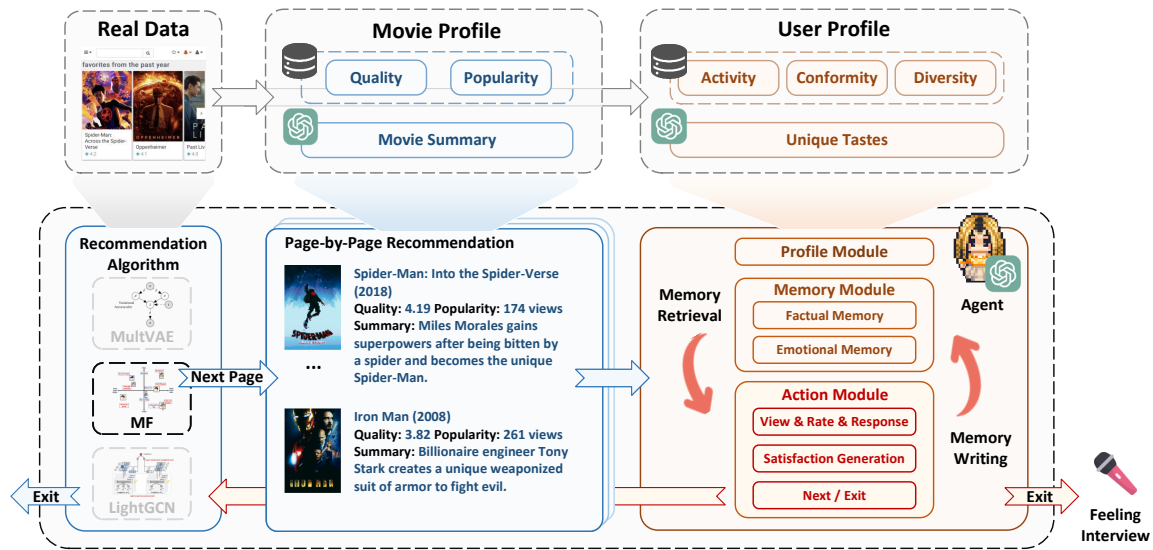


Figure 2: The overall framework of Agent4Rec. Our simulator consists of two core facets: LLM-empowered Generative Agents (Red Section) and Recommendation Environment (Blue Section). Both user and item profiles are initialized using real-world datasets (e.g., MovieLens-1M, Steam, Amazon-Book). The recommendation algorithm, an adaptable component of the system, generates item recommendations presented to the agents in a page-by-page manner. Agents, enhanced with specialized memory and action modules tailored for recommendation scenarios, simulate a wide range of behaviors including viewing, rating items, providing feedback, navigating to the next page, exiting the system, and participating in post-exit interviews.

The feedback covers aspects like whether the user watches the item, his corresponding ratings, and potential exit behaviors. The emotional memory, on the other hand, records user feelings during system interactions, such as levels of fatigue and overall satisfaction. We aim to ensure the generative agent does not merely react based on past factual interactions but also takes into account the feelings, thereby mirroring genuine human behaviors more closely.

We store memories in two formats: natural language descriptions and vector representations. The former is designed for easy understanding by humans [45]. While vector representations are primed for efficient memory retrieval and extraction [57].

To help agents interact with the recommendation environment, we introduce three memory operations: memory retrieval, memory writing, and memory reflection.

- **Memory Retrieval:** Grounded in insights from studies [31, 59], this operation assists the agent in distilling the most relevant information from its memory module.
- **Memory Writing:** This operation enables the recording of agent-simulated interactions and emotions into the memory stream.
- **Memory Reflection:** Recognizing the influence of emotions over user behaviors in recommendations, we incorporate an emotion-driven self-reflection mechanism. This stands in contrast to conventional agent memory designs, such as self-summarization [36], self-verification [48], and self-correction [49], which predominantly condense or deduce advanced factual knowledge, often sidelining emotional feelings. Here, once the agent’s actions surpass a pre-defined count, it triggers a reflection process. Armed with the prowess of the LLM, the agent introspects its satisfaction with the recommendations and assesses its fatigue levels, offering a deeper understanding of its cognitive state.

2.1.3 Action Module. Equipping agents with user profiles and memory modules enables them to exhibit diverse actions akin to humans based on current observations [58]. In Agent4Rec, we design an action module specifically tailored for recommendation domain, which encompasses two broad categories of actions:

- **Taste-driven Actions:** view, rate, and generate post-viewing feelings for items. In Agent4Rec, the recommended items are first generated by recommendation algorithms and then presented to agents in a page-by-page manner (further details are available in Section 2.2). Guided by their taste, agents assess each item on the page for consistency with their preferences. They may choose to watch certain items that pique their interest while bypassing others and subsequently providing ratings and feelings for each item they watch.
- **Emotion-driven Actions:** exit and rate recommender systems, and do post-exit interviews. Emotions in the recommendation environment can shape an agent’s experience significantly, which are often overlooked in simulations [46, 54]. An agent’s satisfaction with previously recommended items and its current fatigue level influences its decision to continue exploring further recommendation pages or to exit the recommender system. To better simulate this multifaceted decision-making, we enhance the agent’s ability for emotional reasoning via Chain-of-Thought [56]. Initially, the agent discerns the current recommendation page and retrieves its satisfaction level with preceding recommendations from its emotional memory. Following this, the agent autonomously expresses its satisfaction and fatigue level for the current recommendation page. Drawing upon these insights, combined with its personalized activity trait, the agent decides whether to exit the system. Post-exit, we conduct an interview with each agent, aiming at

capturing agents' ratings and overall impressions of the recommender system, offering explicit explanations of their behaviors while navigating the system. This interview-style feedback provides a richer and more human-understandable evaluation of the system, enhancing the insights from traditional metrics. An in-depth exploration of the interview can be found in Section 3.5.

2.2 Recommendation Environment

Agent4Rec simulates the interactions between agents and the recommendation environment. We discuss three aspects of environment construction that resonate with real-world scenarios, including item profile generation, page-by-page recommendation scenarios, and recommendation algorithm designs.

- **Item Profile Generation:** We construct item profiles to capture key item features, including quality, popularity, genre, and summary. Quality is deduced from historical ratings, popularity is based on the number of reviews, while genre and summary are generated by LLM. Our goal goes beyond encapsulating the uniqueness of the item in a profile to simulate the recommendation scene of real users. We also aim to test whether LLM has potential hallucinations regarding the item. Our approach utilizes a few-shot learning approach, tasking the LLM with classifying the item into one of 18 genres and generating a summary using only the item title. If the LLM's genre classification aligns with the data, it signifies its knowledge of the item. To maintain reliability, items causing genre misclassification by the LLM are pruned, reducing hallucination risks. This approach ensures the agent's trustworthiness in simulating user behavior.
- **Page-by-Page Recommendation Scenario:** Our simulator mirrors the operation of real-world recommendation platforms like Netflix, YouTube, and Douban, functioning in a page-by-page manner. Users are initially presented with a list of item recommendations on each page. Based on interactions, preferences, and feedback, subsequent pages could be set to tailor the recommendations further, aiming for a more refined user experience. For further details and experimental results, please refer to Section 3.4.
- **Recommendation Algorithm Designs:** In Agent4Rec, the recommendation algorithm is structured as a standalone module, with a core focus on extensibility. This design encompasses pre-implemented collaborative filtering-based strategies, including random, most popular, Matrix Factorization (MF) [24], LightGCN [16], and MultVAE [28]. Moreover, it incorporates an open interface, enabling researchers and practitioners to effortlessly integrate external recommendation algorithms. This adaptability ensures that Agent4Rec could be a versatile platform for comprehensive evaluations and the collection of valuable user feedback in the future.

3 AGENT ALIGNMENT EVALUATION.

With the specialized simulator, Agent4Rec, tailored for recommendations, we would like to explore an essential research question:

- **RQ1:** To what extent can LLM-empowered generative agents truly simulate the behaviors of genuine, independent humans in recommender systems?

In this section, we will delve into the capability and limitations of generative agents from various perspectives, including the alignment of user behavior (such as user taste, rating distribution, and social traits) and the evaluation of the recommendation environment (including recommendation strategies evaluation, page-by-page recommendation enhancements, and the case study of interview).

Motivation. To appropriately respond to recommended items, generative agents need to have a clear understanding of their own preferences. We conjecture that an independent, personalized agent, initialized from real users in MovieLens-1M, should maintain long-term preference coherence. In practice, this implies that the agent should be adept at distinguishing the items that real users favor.

Setting. To validate how well generative agents align with the preferences encoded in their user profiles, we task agents with distinguishing between the items that the corresponding real users have interacted with and those they have not. We conduct experiments on three real-world datasets (*i.e.*, MovieLens-1M [15], Steam[22], Amazon-Book[37]). Specifically, a total of 1,000 agents will each be randomly assigned 20 items. Among these, the ratio between items the user has interacted with (*i.e.*, $y_{ui} = 1$) but was not utilized for profile initialization and items the user has not interacted with (*i.e.*, $y_{ui} = 0$) is set as $1 : m$, with $m \in \{1, 2, 3, 9\}$. Under this setting, agent responses (*i.e.*, \hat{y}_{ui}) to recommended items are considered binary discrimination, taking values between 0 and 1.

Results. Table 1 reports the empirical discrimination results across various metrics. The best performance for each metric is highlighted in bold and marked with an asterisk. We observe that:

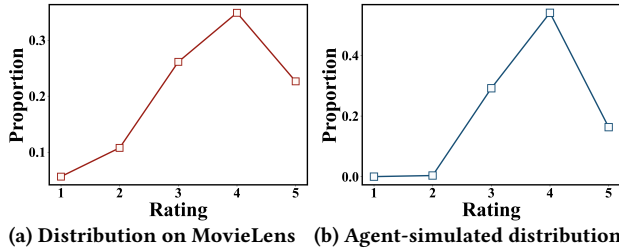
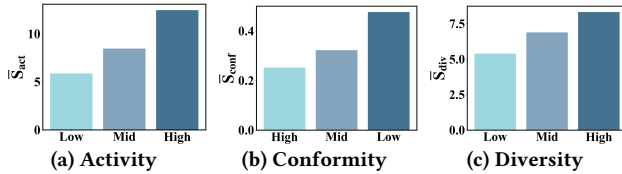
- **Generative agents consistently and impressively identify items aligned with user preferences.** Specifically, regardless of the number of distracting items introduced, agents maintain a high accuracy of around 65% and recall of about 75%. This high fidelity can be attributed to the personalized profile faithfully mirroring the user's genuine interests. It further reflects that the agents effectively encapsulate a substantial portion of the real preferences, signifying the feasibility of LLM-powered generative agents in recommendation simulation.
- **Agents tend to maintain a relatively consistent count of preferred items, a phenomenon potentially stemming from inherent hallucinations in the LLM.** Notably, while Accuracy and Recall show satisfactory results, metrics like Precision and F1 Score, which emphasize the cost of false positives, experience a sharp decline from nearly 70% to about one quarter as the proportion of user-liked items decreases. We attribute this failure to LLM's inherent hallucinations that agents tend to consistently pick a set number of items. Such behavior highlights challenges in designing a more reliable recommendation simulator using LLM-empowered generative agents. However, we emphasize that in the subsequent simulation results with recommendation algorithms, a substantial proportion of recommended items align with user preferences, thereby endorsing high trustworthiness in those simulation outcomes.

3.1 Rating Distribution Alignment

Motivation. Beyond ensuring user-aligned behaviors at a micro-level for each agent, a comprehensive evaluation of the simulator requires that the generative agents accurately mirror real-world

Table 1: User Taste Alignment across MovieLens, Amazon-Book, and Steam datasets

1:m	MovieLens				Amazon-Book				Steam			
	Accuracy	Recall	Precision	F1 Score	Accuracy	Recall	Precision	F1 Score	Accuracy	Recall	Precision	F1 Score
1:1	0.6912*	0.7460	0.6914*	0.6982*	0.7190*	0.7276*	0.7335*	0.7002*	0.6892*	0.7059	0.7031*	0.6786*
1:2	0.6466	0.7602	0.5058	0.5874	0.6842	0.6888	0.5763	0.5850	0.6755	0.7316	0.5371	0.5950
1:3	0.6675	0.7623	0.4562	0.5433	0.6707	0.6909	0.4423	0.5098	0.6505	0.7381*	0.4446	0.5194
1:9	0.6175	0.7753*	0.2139	0.3232	0.6617	0.6939	0.2369	0.3183	0.6021	0.7213	0.1901	0.2822

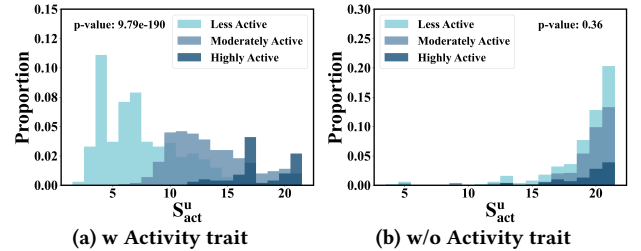
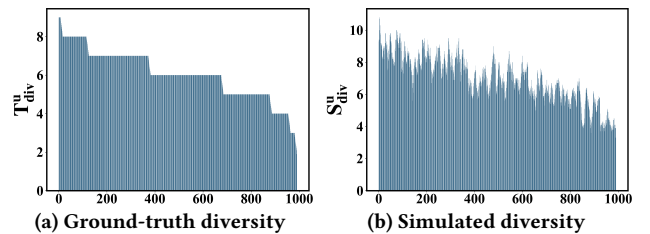
**Figure 3: Comparison between ground-truth and agent-simulated rating distributions.****Figure 4: Averaged scores of activity, conformity, and diversity among agent groups with varying degrees of social traits.**

user behavioral patterns at a macro scale. Specifically, in terms of rating distribution, the goal is to maintain consistency with MovieLens-1M data distributions.

Results. Figures (3a) and (3b) illustrate the ground-truth rating distribution in the MovieLens-1M dataset and the distribution of ratings generated by agents, respectively. The figures demonstrate a strong alignment between the simulated rating distribution and the actual distribution. Specifically, ratings at 4 dominate the overall distribution, while low ratings (1-2) constitute only a small portion. It is worth mentioning that agents tend to give very few 1-2 ratings, which differs from genuine human behavior. This is attributed to the LLM’s extensive prior knowledge of movies, as agents tend to avoid watching low-quality films in advance. Consequently, simulating the user action of giving low ratings after choosing to watch a movie becomes challenging. Moreover, Figure (3b) represents the simulation results under the MF algorithm, and similar trends are observed under other recommendation algorithms. This consistency across different algorithms validates the potential of Agent4Rec. We demonstrate through ablation that the social traits component of our agents contribute critically to the believability of agent behavior.

3.2 Social Traits Alignment

Motivation. In real-world recommendation scenarios, user behavior patterns are influenced by various factors, not solely determined by personal preferences but also impacted by social attributes such as activity, conformity, and diversity [38]. Accordingly, we specially designed a user profile for Agent4Rec that incorporates these

**Figure 5: Distribution of interaction numbers among agents with varying levels of user activity.****Figure 6: Comparison between the individual level distributions of ground-truth and agent-simulated diversity scores.**

social traits. Following the definitions in Section 2.1.1 and based on statistics from MovieLens, agents are categorized into three different levels (high, medium, and low) for each trait, giving a consistent prompt for agents within each tier. We conjecture that agents exhibiting similar preferences might still display differentiated behavior patterns based on their unique social traits.

Results. To validate the significance of the user profile module’s design, it is essential to probe from multiple perspectives, including (1) whether agents categorized into different tiers exhibit distinct behaviors (Figure 4); (2) whether ablation studies can detect differences between agents with and without social traits in their profiles (Figure 5); and (3) whether the simulated traits distribution of agents aligns with that of actual users (Figure 6).

Extensive experimental results from distribution analysis, ablation studies, and statistical tests demonstrate the pivotal role of social traits in user profile construction, significantly influencing agent behavior. This further illustrates the universal applicability and referential value of Agent4Rec’s setup, which is specifically designed for recommendation tasks. It’s worth noting that the diversity trait might exhibit minimal distinction in agent behavior, possibly due to the strong overlap of movie categories in the MovieLens dataset, suggesting further research and validation with alternative datasets.

Table 2: Recommendation strategies evaluation.

	\bar{P}_{view}	\bar{N}_{like}	\bar{P}_{like}	\bar{N}_{exit}	\bar{S}_{sat}
Random	0.312	3.3	0.269	2.99	2.93
Pop	0.398	4.45	0.360	3.01	3.42
MF	0.488	6.07*	0.462	3.17*	3.80
MultVAE	0.495	5.69	0.452	3.10	3.75
LightGCN	0.502*	5.73	0.465*	3.02	3.85*

3.3 Recommendation Strategy Evaluation

Motivation. Human users display varying levels of satisfaction with different recommendation Algorithms. For example, users generally gain higher satisfaction with advanced strategies, compared to random recommendations. Assuming generative agents can accurately simulate genuine human behaviors, it’s plausible that these agents would exhibit similar satisfaction trends to those observations in humans.

Setting. For a fair comparison, we employ mainly collaborative filtering-based recommendation strategies (*i.e.*, random, most popular, MF [24], LightGCN [16], and MultVAE [28]) in our simulator and evaluate their satisfaction trends on MovieLens. In Agent4Rec, four movies are displayed on each recommendation page, and agents will take actions such as viewing and rating based on their personal preferences. Simultaneously, agents decide whether to proceed to the next recommended page or exit the recommendation system based on their satisfaction. The system enforces agents’ exit after browsing a maximum of five pages. Once an agent exits, we request him to give a satisfaction score for the recommendation system, ranging from 1 to 10. Regarding rating higher than 3 as a signal of like, we collect the following multi-facets metrics after the whole simulation completes: average viewing ratio (*i.e.*, \bar{P}_{view}), the average number of likes (*i.e.*, \bar{N}_{like}), the average ratio of likes (*i.e.*, \bar{P}_{like}), the average number of exit page (*i.e.*, \bar{N}_{exit}), and the average user satisfaction score (*i.e.*, \bar{S}_{sat}).

Results. Table 2 reports the multi-facet satisfaction metrics for various recommendation strategies. It is clear that agents exhibit higher satisfaction with algorithm-based recommendations compared to random and popularity-based recommendations. This phenomenon aligns with observations in the real world, where well-designed recommendation algorithms can effectively address the issue of information overload in modern society, thereby enhancing users’ online experience [5]. Furthermore, in line with prevailing insights in the research community, LightGCN outperforms both MF and MultVAE across diverse evaluation criteria. Such findings underscore the fine-grained evaluation capabilities of LLM-empowered agents for recommendation strategies. This further shed light on the potential of an agent-driven recommendation simulator for A/B testing, offering a more cost-efficient alternative to conventional online A/B testing.

3.4 Page-by-Page Recommendation Enhancement

Motivation. In real-world settings, recommendation platforms frequently collect immediate user behaviors to iteratively refine the recommender, aiming to accurately capture users’ latest preferences.

Table 3: Page-by-page recommendation enhancement results over various algorithms.

Offline	MF		MultVAE		LightGCN	
	Recall	NDCG	Recall	NDCG	Recall	NDCG
Origin	0.1506	0.3561	0.1609	0.3512	0.1757	0.3937
+ Unviewed	0.1523	0.3557	0.1598	0.3487	0.1729	0.3849
+ Viewed	0.1570*	0.3604*	0.1613*	0.3540*	0.1765*	0.3943*
Simulation	\bar{N}_{exit}	\bar{S}_{sat}	\bar{N}_{exit}	\bar{S}_{sat}	\bar{N}_{exit}	\bar{S}_{sat}
Origin	3.17	3.80	3.10	3.75	3.02	3.85
+ Unviewed	3.03	3.77	3.01	3.77	3.06	3.81
+ Viewed	3.27*	3.83*	3.18*	3.87*	3.10*	3.92*

Our aim in designing the page-by-page recommendation setting is to emulate this feedback-driven recommendation enhancement.

Setting. After a complete recommendation simulation, we collect both viewed and unviewed movies for each agent. These two types of movies are then added as positive signals to the training set of each user to re-train the recommendation algorithms. We evaluate the performance of these retrained recommenders using standard offline metrics, such as Recall@20 and NDCG@20, as well as satisfaction simulation evaluations, namely \bar{N}_{exit} and \bar{S}_{sat} .

Results. As depicted in Table 3, by leveraging movies viewed by the agent as augmented data, all recommendation algorithms exhibit improvements in both offline evaluation metrics and simulated satisfaction evaluations. However, when the training dataset is augmented with unviewed movies, the overall user experience typically deteriorates. Successfully emulating this feedback-driven recommendation augmentation indicates that movie choices by the agent can serve as a consistent indicator of a user’s unique preferences.

3.5 Case Study of Feeling Interview

Motivation. Compared to conventional recommendation simulators [20, 50], the unique strength of LLM-empowered agent simulation lies in its ability to provide human-comprehensible explanations [19]. Eliciting explanations from agents provides insights into the reliability of simulations, enabling us to further refine the recommender system.

Results. Figure 7 presents a case from post-exit interviewing on MovieLens. The agent rates the recommended movies based on their personal tastes, social traits, and emotional memory. Specifically, the agent recognized that the recommender system did suggest movies in line with their preferences. However, there were aspects leading to less satisfaction. For instance, even though the agent has diverse interests, the system tended to recommend popular movies.

4 INSIGHTS AND EXPLORATION

Given Agent4Rec’s promising simulation capabilities, we pose another profound research question:

- **RQ2:** Can Agent4Rec provide some insights on unresolved problems in the recommendation domain?

In this section, we discuss two insights drawn from our simulation results: replicating the filter bubble phenomenon and exploring causal discovery in movie recommendation tasks. We acknowledge

Case Study of Feeling Interview

Question: Do you feel satisfied with the recommender system? Rate it from 1-10 and give an explanation.

Rating: 6

Reason: While the recommender system did provide me with some movies that aligned with my taste, there were also a few recommendations that I disliked. The system took into account my personal preferences and historical ratings, which I appreciate as a Balanced Evaluator. However, as an Occasional Viewer, I am not easily swayed by movie recommendations and only seek out movies that strictly align with my taste. The system failed to consistently meet my preferences, resulting in a few unsatisfied memories. Additionally, as a Cinematic Trailblazer, I enjoy exploring unique and obscure movies, but the recommendations seemed to focus more on popular classics. Overall, the system had its strengths but fell short of fully satisfying my movie-watching preferences.

Figure 7: One case study of post-exit feeling interview.

the difficulty of RQ2, and confine our discussion here to potential paths for future exploration.

4.1 Filter Bubble Effect

Motivation. The filter bubble effect is a pervasive challenge in recommender systems [12, 39]. This issue emerges when algorithm-based recommenders predict which movies users might prefer based on user feedback loops [35], resulting in increasingly homogeneous recommended contents. Our primary goal is to assess Agent4Rec’s ability to replicate the filter bubble phenomenon.

Setting. To ensure fairness, we divide the MovieLens’ movie pool into four equal parts, allowing the MF-based recommender to undergo four complete simulation rounds. In each round, recommended movies span up to 5 pages, with the MF-based recommender retrained after each simulation round. We evaluate the filter bubble effect based on content diversity at the individual user level. Two metrics are employed: $\bar{P}_{\text{top1-genre}}$, which represents the average percentage of top-1 genres among recommended movies, and \bar{N}_{genres} , indicating the average number of genres recommended in each simulation.

Results. Figure 8 reveals that as the number of iterations increases, movie recommendations tend to be more centralized. Specifically, the genre diversity, represented by \bar{N}_{genres} , decreases, while the dominance of the primary genre, denoted by $\bar{P}_{\text{top1-genre}}$, intensifies. This result further validates Agent4Rec’s capability to reflect the filter bubble effect, an issue commonly observed in real-world recommender systems.

4.2 Discovering Causal Relationships

Motivation. Causal discovery aims to infer a causal structure, often represented as a causal graph, from observational data. This technique is crucial to comprehending the underlying mechanisms of specific fields. A recommendation simulator can aid researchers with data collection and in addressing latent confounding issues.

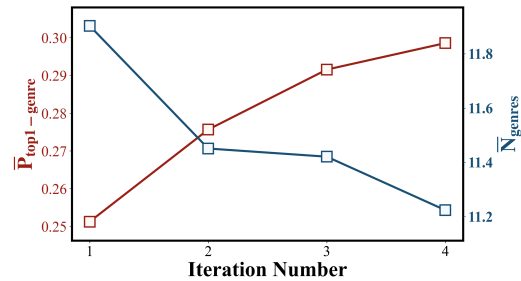


Figure 8: The simulation performance of Agent4Rec to emulate the filter bubble effect.

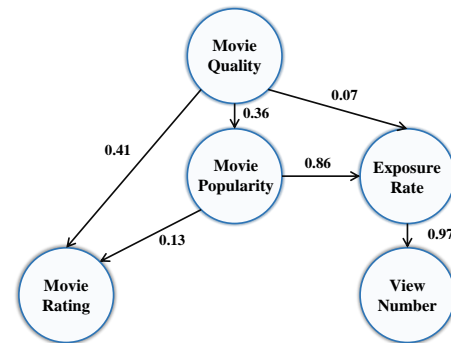


Figure 9: Learned causal graph among movie quality, popularity, exposure rate, view number, and movie rating.

In light of this, a question arises: can Agent4Rec be instrumental in uncovering causal relationships in recommender systems?

Setting. To understand the factors influencing movie ratings on MovieLens, for each movie, we collect data on four principal variables in addition to its rating simulated by agents: movie quality and popularity (sourced from the movie profile), exposure rate, and the number of times the movie is watched (sourced from the simulator). To probe the potential causal relationships within this simulated data, we employ the DirectLiNGAM algorithm [51]. This algorithm discovers a causal graph, *i.e.*, a weighted directed acyclic graph (DAG) in a linear system.

Results. Based on the learned causal graph in Figure 9, we can observe that:

- **Movie quality and movie popularity are the causes of movie ratings [43].** Although movie quality contributes most to agent ratings, agent ratings are also slightly influenced by the popularity of movies. This aligns with the inclination of humans to give high ratings to popular movies in real-world scenarios.
- **The feedback loop of amplifying the popularity bias is observed.** Highly popular movies receive increased exposure, resulting in more views by agents. When these popular items are introduced into the new training dataset, recommender systems are inclined to expose them more in the subsequent iterations, giving rise to what is known as popularity bias [11, 17, 65].

5 RELATED WORKS

In this section, we review two research lines of related works: LLM-empowered generative agents and recommender simulator.

5.1 LLM-empowered Generative Agents

AI agents are artificial entities that sense their environment, make decisions, and take actions [58]. Humans have long been pursuing intelligent agents to handle various tasks [8, 21, 41]. Recently, with the remarkable capability of Large Language Model (LLM) demonstrated, a substantial body of research has emerged [53]. Generative Agent [42] is a pioneer work that designs general agents equipped with memory, planning, and reflection abilities to simulate the human’s daily life. Building upon this universal framework, the following agent architecture can be bifurcated into task-oriented agents and simulation-oriented agents [58].

The core objective of task-oriented agents is to execute predefined tasks established by humans [19, 34, 44, 52, 55, 57]. Within this category, Voyager [52] integrates LLM into the Minecraft universe, endowing the in-game character with the capacity to effectively explore the virtual world. ChatDev [44] introduces a novel agent collaboration paradigm in software development, harnessing group intelligence to streamline the development pipeline. AutoGen [57] further extends this paradigm by defining more complex roles and facilitating multi-agent communication to tackle a diverse range of challenges. In the realm of recommendation systems, RecMind [55] designs an LLM-powered autonomous recommender agent that possesses self-inspiring planning abilities and the capability to leverage external tools. InteRecAgent [19] takes a step beyond by enhancing LLM-based agents with integrated components, yielding recommenders that are both explainable and conversational.

Simulation-oriented agents are geared towards replicating human behaviors in specific scenarios, thereby enabling the acquisition of valuable data and the exploration of social issues [9, 30, 31, 42, 54, 60]. Among these agents, S^3 [9] leverages autonomous agents to simulate the dynamic evolution of public opinion on social platforms in response to trending social events. SANDBOX [31] facilitates communication among a group of agents to address social issues thus providing ethically sound data for LLM fine-tuning. Differing from task-oriented agents in the field of recommendation, simulation-oriented agents seek to emulate user behaviors within recommender systems rather than focusing on the recommender. RecAgent [54] attempts to integrate diverse user behaviors in recommendation environments, taking into account external social relationships. Focusing on simulating and evaluating user interactions with the recommenders, our framework Agent4Rec also falls into the category of simulation-oriented agents.

5.2 Recommendation Simulator

Recommendation simulator is a cornerstone in the field of recommendation systems [1, 32, 35, 61]. It diverges from conventional methods that rely on real-world user data, as it excels in replicating or simulating user interactions within recommender systems. This approach offers a cost-effective alternative to online environments and also presents a potential avenue for addressing prevalent challenges such as casual discovery and filter bubbles [4, 18]. Early forays into recommendation simulators primarily serve as rich data sources for subsequent utilization, particularly in the realm of reinforcement learning [20, 46, 50]. For example, Virtual Taobao [50] proposes a virtual environment based on real user interactions to simulate user behaviors on e-commerce platforms. RecSim [20]

provides comprehensive toolkits for effectively simulating user behaviors in the setting of sequential recommendation. Also, RecoGym [46] integrates both traditional recommendation algorithms and reinforcement learning framework, followed by evaluation with online and offline metrics. MINDSim [32] simulates user behaviors in a news website. But these traditional simulators all fall short in relatively simple rules, thus lacking flexibility and validity. Recently, LLM-empowered agents have demonstrated significant promise in approximating human-like intelligence, which showcases the considerable potential for recommendation simulators [53, 58]. RecAgent [54] makes the primary attempt to construct a recommendation platform to integrate diverse user behaviors involving movie-watching, chatting, posting, and searching. Instead of behavior integration, we set our sights on in-depth exploration to emulate and evaluate user interactions with both rule-based and algorithm-based recommenders in recommender systems.

6 LIMITATIONS AND FUTURE WORK

Even though Agent4Rec offers a promising research direction in recommender system simulation, we recognize its potential limitations, risks, and challenges that require further exploration and in-depth investigation.

- **Datasource Constraints.** Agent4Rec is implemented exclusively utilizing offline datasets and is primarily constrained by two key factors. First, LLMs necessitate prior knowledge regarding the recommended items, rendering most offline datasets — with only IDs or lacking detailed item descriptions — ill-suited for this task. Furthermore, while online data undoubtedly aligns more naturally with simulators, providing an unbiased perspective to evaluate their effectiveness, acquiring such data poses a considerable challenge.
- **Limited Action Space.** The action space of Agent4Rec is currently limited, omitting critical factors that influence user decisions, such as social networks, advertising, and word-of-mouth marketing. While this simplification facilitates a reliable evaluation of LLM-empowered agents under simple scenarios, it also introduces a gap in real-world user decision-making processes. A key direction for our future work is to encompass a wider spectrum of influential factors to better capture the multifaceted nature of user behaviors, ensuring the simulations are more universally representative of recommendation scenarios.
- **Hallucination in LLM.** Occasional hallucinations have been observed in simulations, such as the LLM failing to accurately simulate human users providing unfavorable ratings to adopted items, fabricating non-existent items and rating them, or not adhering to the required output format. Such inconsistencies can lead to inaccurate simulation outcomes. In light of these observations, our future goal is to fine-tune an LLM specifically for simulating user behavior in recommendation scenarios to enhance the simulator’s stability and precision.

ACKNOWLEDGMENTS

This research is supported by the National Science and Technology Major Project (2023ZD0121102), National Natural Science Foundation of China (92270114), the National Research Foundation, Singapore under its Industry Alignment Fund – Pre-positioning (IAF-PP)

Funding Initiative, and the NExT Research Center. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the author(s) and do not reflect the views of National Research Foundation, Singapore.

REFERENCES

- [1] Mohammad Mehdi Afsar, Trafford Crump, and Behrouz H. Far. 2023. Reinforcement Learning based Recommender Systems: A Survey. *ACM Comput. Surv.* 55, 7 (2023), 145:1–145:38.
- [2] Oscar Celma and Perfecto Herrera. 2008. A new approach to evaluating novel recommendations. In *RecSys*.
- [3] Jiawei Chen, Hande Dong, Xiang Wang, Fuli Feng, Meng Wang, and Xiangnan He. 2023. Bias and Debias in Recommender System: A Survey and Future Directions. *ACM Trans. Inf. Syst.* 41, 3 (2023), 67:1–67:39.
- [4] Kinshi Chen, Shuang Li, Hui Li, Shaohua Jiang, Yuan Qi, and Le Song. 2019. Generative Adversarial User Model for Reinforcement Learning Based Recommendation System. In *ICML*. PMLR.
- [5] Aminu Da'u and Naomie Salim. 2020. Recommendation system based on deep learning methods: a systematic review and new directions. *Artificial Intelligence Review* 53, 4 (2020), 2709–2748.
- [6] Gabriel Dulac-Arnold, Daniel J. Mankowitz, and Todd Hester. 2019. Challenges of Real-World Reinforcement Learning. *CoRR abs/1904.12901* (2019).
- [7] Michael D. Ekstrand, Mucun Tian, Ion Madrazo Azpiaz, Jennifer D. Ekstrand, Oghenemaro Anuyah, David McNeill, and Maria Soledad Pera. 2018. All The Cool Kids, How Do They Fit In?: Popularity and Demographic Biases in Recommender Evaluation and Effectiveness. In *FAT*.
- [8] Stan Franklin and Arthur C. Graesser. 1996. Is it an Agent, or Just a Program?: A Taxonomy for Autonomous Agents. 1193 (1996), 21–35.
- [9] Chen Gao, Xiaochong Lan, Zhihong Lu, Jinzhu Mao, Jinghua Piao, Huandong Wang, Depeng Jin, and Yong Li. 2023. S³: Social-network Simulation System with Large Language Model-Empowered Agents. *CoRR abs/2307.14984* (2023).
- [10] Chongming Gao, Wenqiang Lei, Jiawei Chen, Shiqi Wang, Xiangnan He, Shijun Li, Biao Li, Yuan Zhang, and Peng Jiang. 2023. CIRS: Bursting Filter Bubbles by Counterfactual Interactive Recommender System. *ACM Trans. Inf. Syst.* 42, 1 (2023), 1 – 27.
- [11] Chen Gao, Yu Zheng, Wenjie Wang, Fuli Feng, Xiangnan He, and Yong Li. 2022. Causal inference in recommender systems: A survey and future directions. *arXiv preprint arXiv:2208.12397* (2022).
- [12] Zhaolin Gao, Tianshu Shen, Zheda Mai, Mohamed Reda, Bouadjenek, Isaac Waller, Ashton Anderson, Ron Bodkin, and Scott Sanner. 2022. Mitigating the Filter Bubble while Maintaining Relevance: Targeted Diversification with VAE-based Recommender Systems. In *SIGIR*.
- [13] Carlos Alberto Gomez-Urbe and Neil Hunt. 2016. The Netflix Recommender System: Algorithms, Business Value, and Innovation. *ACM Trans. Manag. Inf. Syst.* 6, 4 (2016), 13:1–13:19.
- [14] Wes Gurnee and Max Tegmark. 2023. Language Models Represent Space and Time. *CoRR abs/2310.02207* (2023).
- [15] F. Maxwell Harper and Joseph A. Konstan. 2016. The MovieLens Datasets: History and Context. *ACM Trans. Interact. Intell. Syst.* 5, 4 (2016), 19:1–19:19.
- [16] Xiangnan He, Kuan Deng, Xiang Wang, Yan Li, Yong-Dong Zhang, and Meng Wang. 2020. LightGCN: Simplifying and Powering Graph Convolution Network for Recommendation. In *SIGIR*.
- [17] Yue He, Zimu Wang, Peng Cui, Hao Zou, Yafeng Zhang, Qiang Cui, and Yong Jiang. 2022. CausPref: Causal Preference Learning for Out-of-Distribution Recommendation. In *WWW*.
- [18] Jin Huang, Harrie Oosterhuis, Maarten de Rijke, and Herke van Hoof. 2020. Keeping Dataset Biases out of the Simulation: A Debaised Simulator for Reinforcement Learning based Recommender Systems. In *RecSys*.
- [19] Xu Huang, Jianxun Lian, Yuxuan Lei, Jing Yao, Defu Lian, and Xing Xie. 2023. Recommender AI Agent: Integrating Large Language Models for Interactive Recommendations. *CoRR abs/2308.16505* (2023).
- [20] Eugene Ie, Chih-Wei Hsu, Martin Mladenov, Vihan Jain, Sanmit Narvekar, Jing Wang, Rui Wu, and Craig Boutilier. 2019. RecSim: A Configurable Simulation Platform for Recommender Systems. *CoRR abs/1909.04847* (2019).
- [21] Nicholas R. Jennings, Katia P. Sycara, and Michael J. Wooldridge. 1998. A Roadmap of Agent Research and Development. *Auton. Agents Multi Agent Syst.* 1, 1 (1998), 7–38.
- [22] Wang-Cheng Kang and Julian J. McAuley. 2018. Self-Attentive Sequential Recommendation. In *ICDM*.
- [23] Ron Kohavi, Alex Deng, Brian Frasca, Roger Longbotham, Toby Walker, and Yu Xu. 2012. Trustworthy online controlled experiments: five puzzling outcomes explained. In *KDD*.
- [24] Yehuda Koren, Robert M. Bell, and Chris Volinsky. 2009. Matrix Factorization Techniques for Recommender Systems. *Computer* 42, 8 (2009), 30–37.
- [25] Michal Kosinski, David Stillwell, and Thore Graepel. 2013. Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the national academy of sciences* 110, 15 (2013), 5802–5805.
- [26] Kevin S LaBar and Roberto Cabeza. 2006. Cognitive neuroscience of emotional memory. *Nature Reviews Neuroscience* 7, 1 (2006), 54–64.
- [27] Zhenyang Li, Yancheng Dong, Chen Gao, Yizhou Zhao, Dong Li, Jianye Hao, Kai Zhang, Yong Li, and Zhi Wang. 2023. Breaking Filter Bubble: A Reinforcement Learning Framework of Controllable Recommender System. In *WWW*.
- [28] Dawen Liang, Rahul G. Krishnan, Matthew D. Hoffman, and Tony Jebara. 2018. Variational Autoencoders for Collaborative Filtering. In *WWW*.
- [29] Jiayi Liao, Sihang Li, Zhengyi Yang, Jiancan Wu, Yancheng Yuan, Xiang Wang, and Xiangnan He. 2024. Large Language-Recommendation Assistant. In *SIGIR*.
- [30] Jiaju Lin, Haoran Zhao, Aochi Zhang, Yiting Wu, Huqiyue Ping, and Qin Chen. 2023. AgentSims: An Open-Source Sandbox for Large Language Model Evaluation. *CoRR abs/2308.04026* (2023).
- [31] Ruibo Liu, Ruixin Yang, Chenyan Jia, Ge Zhang, Denny Zhou, Andrew M. Dai, Diyi Yang, and Soroush Vosoughi. 2023. Training Socially Aligned Language Models in Simulated Human Society. *CoRR abs/2305.16960* (2023).
- [32] Xufang Luo, Zheng Liu, Shitao Xiao, Xing Xie, and Dongsheng Li. 2022. MINDSims: User Simulator for News Recommenders. In *WWW*.
- [33] Hao Ma, Irwin King, and Michael R. Lyu. 2009. Learning to recommend with social trust ensemble. In *SIGIR*.
- [34] Zhao Mandi, Shreeya Jain, and Shuran Song. 2023. RoCo: Dialectic Multi-Robot Collaboration with Large Language Models. *CoRR abs/2307.04738* (2023).
- [35] Masoud Mansoury, Himan Abdollahpouri, Mykola Pechenizkiy, Bamshad Mobasher, and Robin Burke. 2020. Feedback loop and bias amplification in recommender systems. In *CIKM*.
- [36] Jordan K. Matelsky, Felipe Parodi, Tony Liu, Richard D. Lange, and Konrad P. Kornding. 2023. A large language model-assisted education tool to provide feedback on open-ended responses. *CoRR abs/2308.02439* (2023).
- [37] Julian J. McAuley, Christopher Targett, Qinfeng Shi, and Anton van den Hengel. 2015. Image-Based Recommendations on Styles and Substitutes. In *SIGIR*.
- [38] Manel Mezghani, Corinne Amel Zayani, Ikram Amous, and Faiez Gargouri. 2012. A user profile modeling using social annotations: a survey. In *WWW*.
- [39] Tien T Nguyen, Pik-Mai Hui, F Maxwell Harper, Loren Terveen, and Joseph A Konstan. 2014. Exploring the filter bubble: the effect of using recommender systems on content diversity. In *WWW*.
- [40] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *NeurIPS*.
- [41] Liviu Panait and Sean Luke. 2005. Cooperative Multi-Agent Learning: The State of the Art. *Auton. Agents Multi Agent Syst.* 11, 3 (2005), 387–434.
- [42] Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative Agents: Interactive Simulacra of Human Behavior. *CoRR abs/2304.03442* (2023).
- [43] Judea Pearl, Madelyn Glymour, and Nicholas P Jewell. 2016. *Causal inference in statistics: A primer*. John Wiley & Sons.
- [44] Chen Qian, Xin Cong, Cheng Yang, Weize Chen, Yusheng Su, Juyuan Xu, Zhiyuan Liu, and Maosong Sun. 2023. Communicative Agents for Software Development. *CoRR abs/2307.07924* (2023).
- [45] Yujia Qin, Shengding Hu, Yankai Lin, Weize Chen, Ning Ding, Ganqu Cui, Zheni Zeng, Yufei Huang, Chaojun Xiao, Chi Han, Yi Ren, Fung, Yusheng Su, Huadong Wang, Cheng Qian, Runchu Tian, Kunlun Zhu, Shihao Liang, Xingyu Shen, Bokai Xu, Zhen Zhang, Yining Ye, Bowen Li, Ziwei Tang, Jing Yi, Yuzhang Zhu, Zhenning Dai, Lan Yan, Xin Cong, Yaxi Lu, Weilin Zhao, Yuxiang Huang, Junxi Yan, Xu Han, Xian Sun, Dahai Li, Jason Phang, Cheng Yang, Tongshuang Wu, Heng Ji, Zhiyuan Liu, and Maosong Sun. 2023. Tool Learning with Foundation Models. *CoRR abs/2304.08354* (2023).
- [46] David Rohde, Stephen Bonner, Travis Dunlop, Flavian Vasile, and Alexandros Karatzoglou. 2018. RecoGym: A Reinforcement Learning Environment for the problem of Product Recommendation in Online Advertising. *CoRR abs/1808.00720* (2018).
- [47] Abulhair Saparov, Richard Yuanzhe Pang, Vishakh Padmakumar, Nitish Joshi, Seyed Mehran Kazemi, Najoung Kim, and He He. 2023. Testing the General Deductive Reasoning Capacity of Large Language Models Using OOD Examples. In *NeurIPS*.
- [48] Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. 2023. Toolformer: Language Models Can Teach Themselves to Use Tools. *CoRR abs/2302.04761* (2023).
- [49] Bilgehan Sel, Ahmad Al-Tawaha, Vanshaj Khattar, Lu Wang, Ruoxi Jia, and Ming Jin. 2023. Algorithm of Thoughts: Enhancing Exploration of Ideas in Large Language Models. *CoRR abs/2308.10379* (2023).
- [50] Jing-Cheng Shi, Yang Yu, Qing Da, Shi-Yong Chen, and Anxiang Zeng. 2019. Virtual-Taobao: Virtualizing Real-World Online Retail Environment for Reinforcement Learning. In *AAAI*.

- [51] Shohei Shimizu, Takanori Inazumi, Yasuhiro Sogawa, Aapo Hyvarinen, Yoshinobu Kawahara, Takashi Washio, Patrik O Hoyer, Kenneth Bollen, and Patrik Hoyer. 2011. DirectLiNGAM: A direct method for learning a linear non-Gaussian structural equation model. *Journal of Machine Learning Research-JMLR* 12, Apr (2011), 1225–1248.
- [52] Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandilekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. 2023. Voyager: An Open-Ended Embodied Agent with Large Language Models. *CoRR abs/2305.16291* (2023).
- [53] Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, Wayne Xin Zhao, Zhewei Wei, and Ji-Rong Wen. 2023. A Survey on Large Language Model based Autonomous Agents. *CoRR abs/2308.11432* (2023).
- [54] Lei Wang, Jingsen Zhang, Hao Yang, Zhiyuan Chen, Jiakai Tang, Zeyu Zhang, Xu Chen, Yankai Lin, Ruihua Song, Wayne Xin Zhao, Jun Xu, Zhicheng Dou, Jun Wang, and Ji-Rong Wen. 2023. When Large Language Model based Agent Meets User Behavior Analysis: A Novel User Simulation Paradigm. *CoRR abs/2306.02552* (2023).
- [55] Yancheng Wang, Ziyang Jiang, Zheng Chen, Fan Yang, Yingxue Zhou, Eunah Cho, Xing Fan, Xiaojiang Huang, Yanbin Lu, and Yingzhen Yang. 2023. RecMind: Large Language Model Powered Agent For Recommendation. *CoRR abs/2308.14296* (2023).
- [56] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. In *NeurIPS*.
- [57] Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Shaokun Zhang, Erkang Zhu, Beibin Li, Li Jiang, Xiaoyun Zhang, and Chi Wang. 2023. AutoGen: Enabling Next-Gen LLM Applications via Multi-Agent Conversation Framework. *CoRR abs/2308.08155* (2023).
- [58] Zhiheng Xi, Wenxiang Chen, Xin Guo, Wei He, Yiwen Ding, Boyang Hong, Ming Zhang, Junzhe Wang, Senjie Jin, Enyu Zhou, Rui Zheng, Xiaoran Fan, Xiao Wang, Limao Xiong, Yuhao Zhou, Weiran Wang, Changhao Jiang, Yicheng Zou, Xiangyang Liu, Zhangyue Yin, Shihan Dou, Rongxiang Weng, Wensen Cheng, Qi Zhang, Wenjuan Qin, Yongyan Zheng, Xipeng Qiu, Xuanjing Huan, and Tao Gui. 2023. The Rise and Potential of Large Language Model Based Agents: A Survey. *CoRR abs/2309.07864* (2023).
- [59] Yuchen Xia, Manthan Shenoy, Nasser Jazdi, and Michael Weyrich. 2023. Towards autonomous system: flexible modular production system enhanced with large language model agents. *CoRR abs/2304.14721* (2023).
- [60] Yuzhuang Xu, Shuo Wang, Peng Li, Fuwen Luo, Xiaolong Wang, Weidong Liu, and Yang Liu. 2023. Exploring Large Language Models for Communication Games: An Empirical Study on Werewolf. *CoRR abs/2309.04658* (2023).
- [61] Mengyue Yang, Quanyu Dai, Zhenhua Dong, Xu Chen, Xiuqiang He, and Jun Wang. 2021. Top-N Recommendation with Counterfactual User Preference Simulation. In *CIKM*.
- [62] An Zhang, Fangfu Liu, Wenchang Ma, Zhibo Cai, Xiang Wang, and Tat-Seng Chua. 2023. Boosting Differentiable Causal Discovery via Adaptive Sample Reweighting. *ICLR* (2023).
- [63] An Zhang, Wenchang Ma, Xiang Wang, and Tat seng Chua. 2022. Incorporating Bias-aware Margins into Contrastive Loss for Collaborative Filtering. In *NeurIPS*.
- [64] An Zhang, Leheng Sheng, Zhibo Cai, Xiang Wang, and Tat-Seng Chua. 2023. Empowering Collaborative Filtering with Principled Adversarial Contrastive Loss. In *NeurIPS*.
- [65] An Zhang, Jingnan Zheng, Xiang Wang, Yancheng Yuan, and Tat-Seng Chua. 2023. Invariant Collaborative Filtering to Popularity Distribution Shift. In *WWW*.
- [66] Shengyu Zhang, Ziqi Jiang, Jiangchao Yao, Fuli Feng, Kun Kuang, Zhou Zhao, Shuo Li, Hongxia Yang, Tat-seng Chua, and Fei Wu. 2023. Causal Distillation for Alleviating Performance Heterogeneity in Recommender Systems. *IEEE Trans. Knowl. Data Eng.* (2023).
- [67] Shuai Zhang, Lina Yao, Aixin Sun, and Yi Tay. 2019. Deep Learning Based Recommender System: A Survey and New Perspectives. *ACM Comput. Surv.* 52, 1 (2019), 5:1–5:38.
- [68] Wayne Xin Zhao, Jinpeng Wang, Yulan He, Jian-Yun Nie, Ji-Rong Wen, and Xiaoming Li. 2015. Incorporating Social Role Theory into Topic Models for Social Media Content Analysis. *IEEE Trans. Knowl. Data Eng.* 27, 4 (2015), 1032–1044.
- [69] Yuyue Zhao, Jiancan Wu, Xiang Wang, Wei Tang, Dingxian Wang, and Maarten de Rijke. 2024. Let Me Do It For You: Towards LLM Empowered Recommendation via Tool Learning. In *SIGIR*.
- [70] Yu Zheng, Chen Gao, Xiang Li, Xiangnan He, Yong Li, and Depeng Jin. 2021. Disentangling User Interest and Conformity for Recommendation with Causal Embedding. In *WWW*.