

# Large Language Model Powered Agents for Information Retrieval

Tutorial at SIGIR 2024 in Washington D.C.

An Zhang<sup>1</sup>, Yang Deng<sup>2</sup>, Yankai Lin<sup>3</sup>, Xu Chen<sup>3</sup>, Ji-Rong Wen<sup>3</sup>, Tat-Seng Chua<sup>1</sup>

<sup>1</sup> NExT++ Research Centre, National University of Singapore

<sup>2</sup> School of Computing and Information Systems, Singapore Management University

<sup>3</sup> Gaoling School of Artificial Intelligence, Renmin University of China

[an\\_zhang@nus.edu.sg](mailto:an_zhang@nus.edu.sg), [dengyang17dydy@gmail.com](mailto:dengyang17dydy@gmail.com), [yankailin@ruc.edu.cn](mailto:yankailin@ruc.edu.cn)  
[xu.chen@ruc.edu.cn](mailto:xu.chen@ruc.edu.cn), [jrwen@ruc.edu.cn](mailto:jrwen@ruc.edu.cn), [chuats@comp.nus.edu.sg](mailto:chuats@comp.nus.edu.sg)

July 14, 2024, Washington D.C., USA



## Personal Information

**Zhang An 张岸**

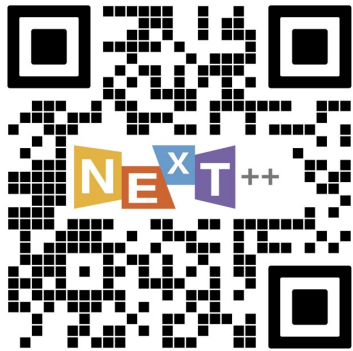
### ➤ **Education Background**

- 2021 – present: **Post-Doc**, NUS, School of Computing, NExT++ Research Centre
- 2016 – 2021: **Ph.D**, NUS, Department of Statistics and Data Science
- 2012 – 2016: **B.S.**, Southeast University, School of Mathematics

➤ **Research Interests:** LLM-empowered Agents, Robust and Trustable AI, Recommender System

➤ **Homepage:** <https://anzhang314.github.io/>

➤ **Email:** [an\\_zhang@nus.edu.sg](mailto:an_zhang@nus.edu.sg)



Homepage

- Part 1: Introduction of LLM-powered Agents
- Part 2: LLM-powered Agents with **Tool Learning**
- Part 3: LLM-powered Agents in **Social Network**
- **Part 4: LLM-powered Agents in Recommendation**
- Part 5: LLM-powered **Conversational Agents**
- Part 6: Open Challenges and Beyond



# Significant Gap Between LLMs and Recommender Systems (RecSys)

- Significant **gap** between large language models (LLMs) and recommender systems (RecSys).

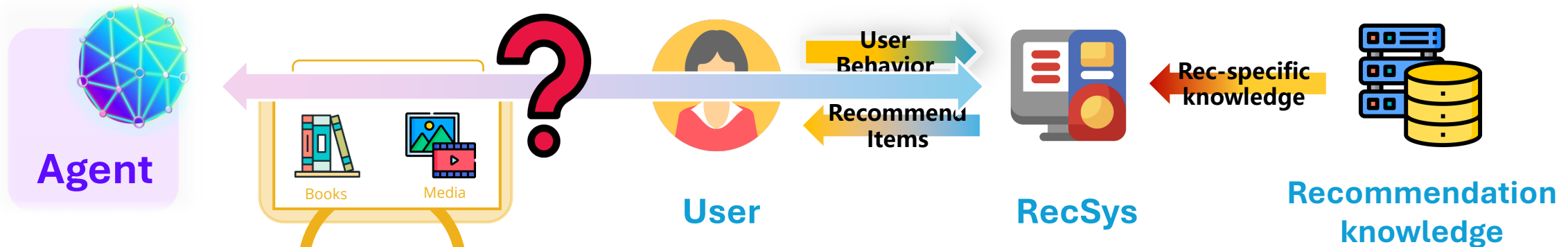
## How to bridge this gap?

	LLMs	RecSys
Scope	Language modelling	User behaviour modelling
Data	Rich <b>world</b> text-based sources	Sparse user-item interactions
Tokens	A chunk of text ( <b>Ten thousand</b> level)	Items ( <b>Billion</b> level)
Characteristics	<b>General</b> model; Open-world knowledge; <b>High complexity</b> and long inference time;	Leveraging <b>collaborative</b> signals; Lack of <b>cross-domain</b> adaptability; Struggle with <b>cold-start</b> problem; Limited <b>intention</b> understanding;

# Significant Gap Between LLMs and Recommender Systems (RecSys)

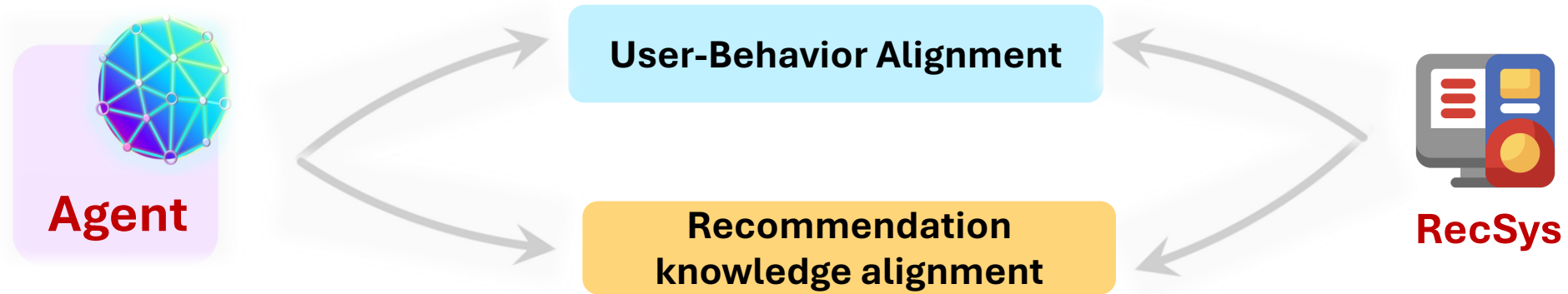
- Significant **gap** between large language models (LLMs) and recommender systems (RecSys).

How to bridge this gap?



- **Align recommendation space with language space.**
  - User behavior alignment
  - Recommendation knowledge alignment
- **Two critical components in RecSys:**
  - Understanding user's behavior/preference
  - Acquiring recommendation-specific knowledge





- LLM-powered Agents have potentials to solve long-standing problems in recommendation
  - Can an LLM-powered Agent faithfully simulate **users**?
  - Can an LLM-powered Agent be a better **recommender** with recommendation-specific knowledge?

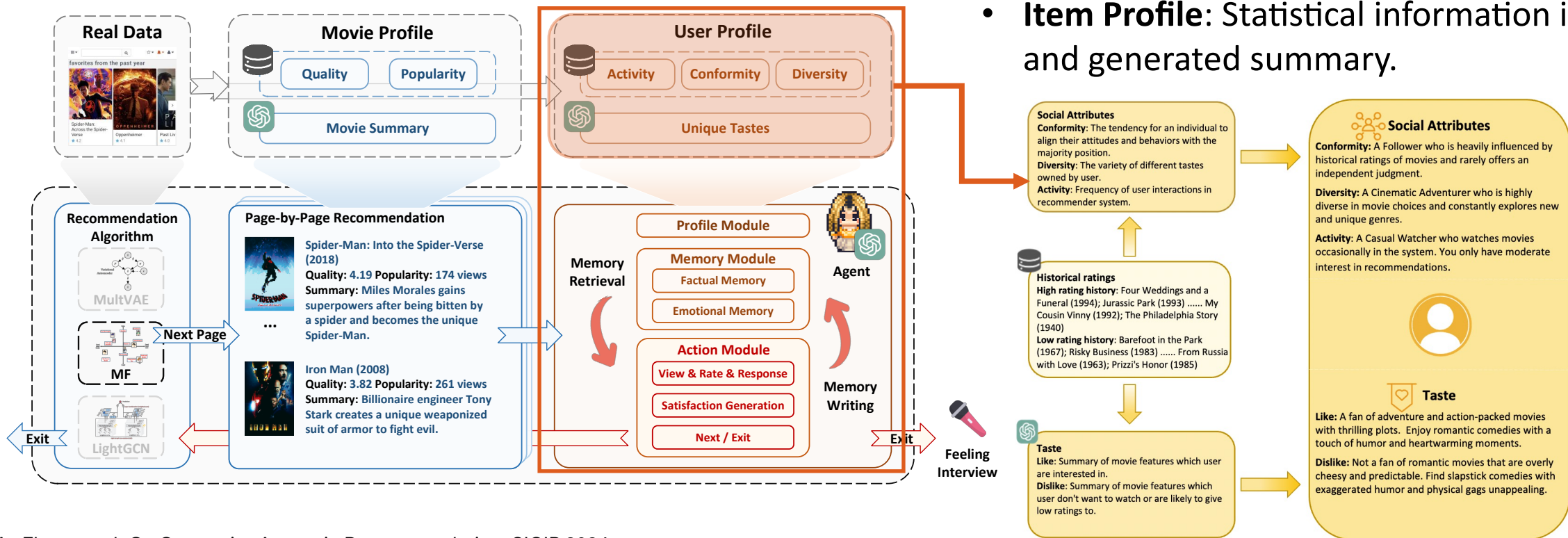
### Agents as Users

### Agent4Rec: Agent-driven user behavior simulation

#### Key Points:

- Can LLM-powered Agent generate faithful user behaviors?

- User Profile:** 1,000 LLM-empowered generative agents initialized with **real data** in various dataset and augmented by ChatGPT.
- Item Profile:** Statistical information in dataset and generated summary.

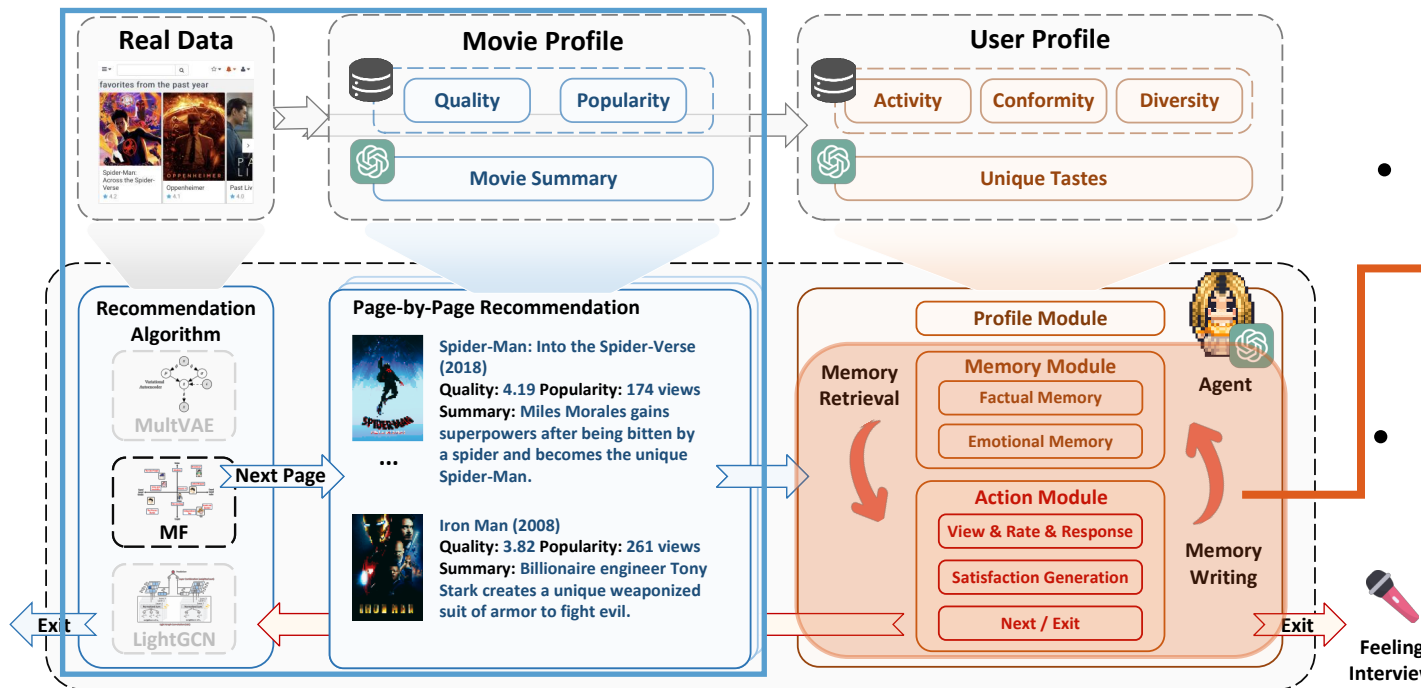


### Agents as Users

### Agent4Rec: Agent-driven user behavior simulation

#### Key Points:

- Can LLM-powered Agent generate faithful user behaviors?



- Agents as users: **1,000** LLM-empowered generative agents initialized from the real dataset.
- Memory** and **action** modules enable agents to recall past interests and plan future actions (**watch, rate, evaluate, exit, and interview**).
- Recommendation environment: Agent4Rec conducts personalized recommendations in a **page-by-page manner** and **pre-implements various recommendation algorithms**.



### Key Observations:

- Agents are capable of **preserving the user's social attributes and preference.**
- Incorporating agents' rating as augmented data can **enhance the recommender's performance.**

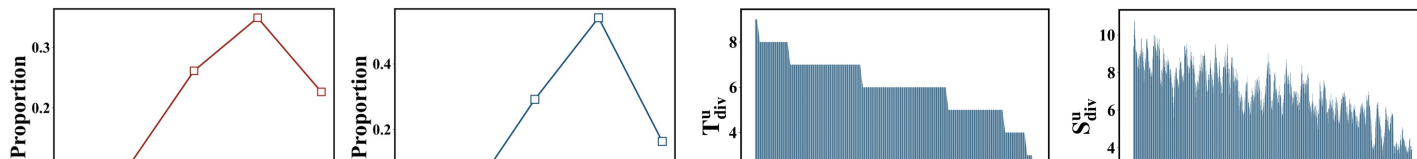


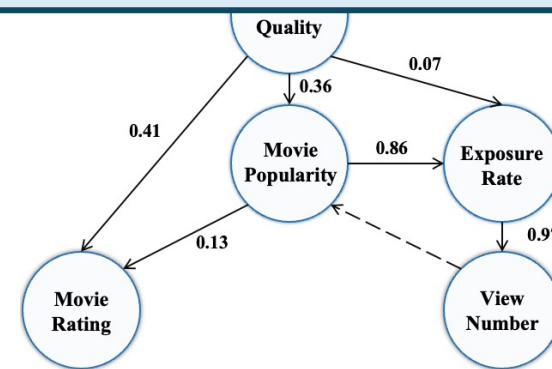
Table 3: Page-by-page recommendation enhancement results over various algorithms.

Offline	MF		MultVAE		LightGCN	
	Recall	NDCG	Recall	NDCG	Recall	NDCG
Origin	0.1506	0.3561	0.1609	0.3512	0.1757	0.3937
+ Viewed	<b>0.1570*</b>	<b>0.3694*</b>	<b>0.1612*</b>	<b>0.3549*</b>	<b>0.1765*</b>	<b>0.3942*</b>

**LLM-powered agents are able to generate faithful behaviors.**

By utilizing LLM-based LLM4Rec to analyse the results, we are able to **discover Causal Relations** among movie quality, movie rating, movie popularity, exposure rate, and view number.

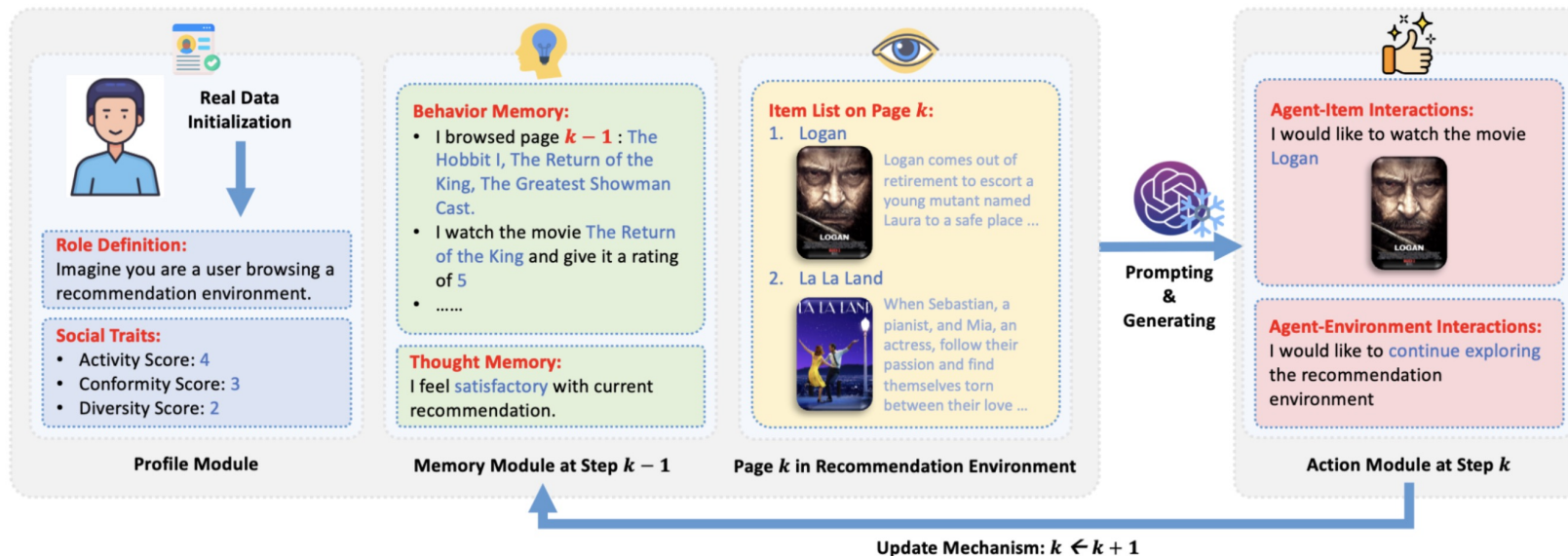
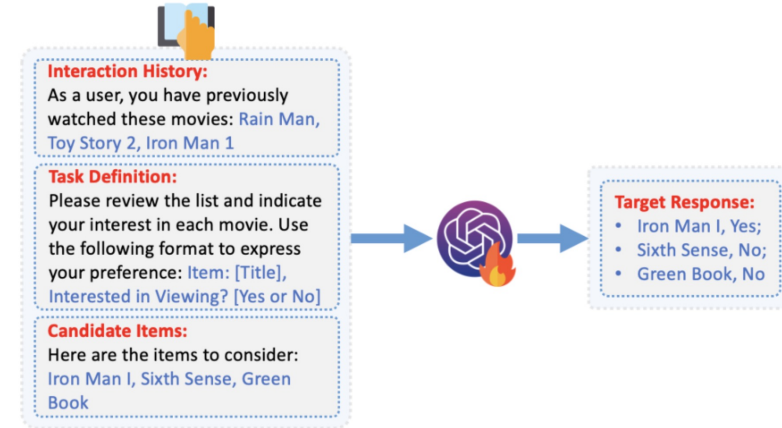
- Offer a simulation platform to test and fine-tune recommender models.**



### Agents as Users

#### Key Points :

- Can LLM-powered Agents generated behaviors benefit the recommender?
- Cooperating updated Agent4Rec framework with **finetuning GPT-3.5-turbo** as a warmup, agents can accurately select their interested items among candidate set.



- Agents have potentials to **replace discriminative learning with generative learning paradigms** for user modeling in recommendation.
- Conduct extensive experiments **on three dataset** from different domains (movie, book, game).

### Key Observations:

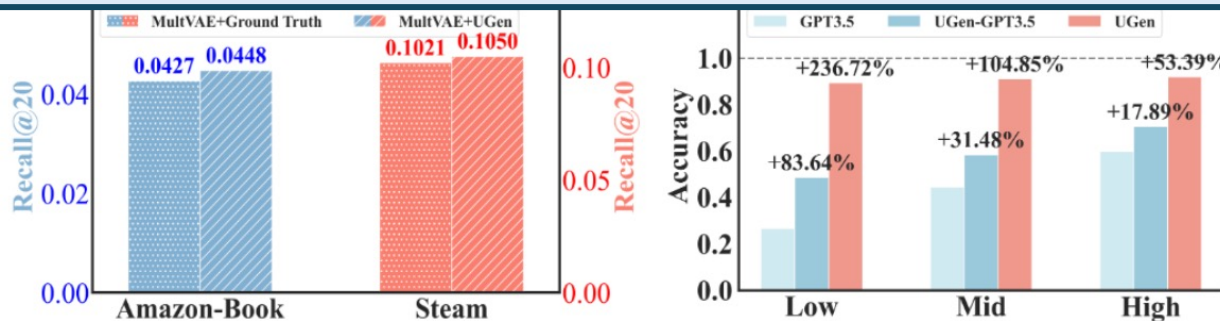
- Agents are capable of **providing effective behaviors**, especially in scenarios with sparse data.

Table 2: Faithfulness Evaluation of Agent’s Behavior Alignment with Real User Preferences. Average ground-truth positives are **7.14 (MovieLens)**, **6.57 (Amazon-Book)**, and **5.80 (Steam)**. UGen shows significant improvement with  $p$ -value  $\ll 0.05$ .

	MovieLens				Amazon-Book				Steam			
	Acc	Pre	Rec	#Select	Acc	Pre	Rec	#Select	Acc	Pre	Rec	#Select
GPT3.5	0.5295	0.4307	0.7369	11.63	0.4202	0.3855	<b>0.9072</b>	17.10	0.4350	0.3430	0.9164	16.59
GPT4	0.6930	0.5743	0.6577	7.00	0.7947	0.6500	0.6003	5.16	0.7844	0.5103	0.7072	6.22
RecAgent	0.6168	0.4519	<b>0.8921</b>	13.95	0.5411	0.3714	0.8150	14.65	0.4916	0.3485	<b>0.9389</b>	15.55
RAH	0.5758	0.4096	0.6383	9.44	0.7253	0.3355	0.3950	7.45	0.6118	0.3874	0.6262	10.37

	MovieLens		Amazon-Book		Steam	
	Recall@20	NDCG@20	Recall@20	NDCG@20	Recall@20	NDCG@20
MF	0.1529	0.3186	0.0257	0.0480	0.0694	0.0567
+ Random	0.1365	0.2913	0.0199	0.0225	0.0526	0.0432
+ GPT3.5	0.1448	0.3089	0.0253	0.0330	<u>0.0732</u>	<u>0.0608</u>
+ RecAgent	0.1400	0.2990	0.0254	0.0317	0.0696	0.0567
+ RAH	0.1363	0.2917	0.0257	0.0370	0.0731	0.0604
+ UGen	<b>0.1667</b>	<b>0.3396</b>	<b>0.0413</b>	<b>0.0573</b>	<b>0.0807</b>	<b>0.0659</b>
Imp.% over MF	9.03%	6.59%	60.70%	19.38%	16.28%	16.23%

Behaviors generated by LLM-powered agents **can benefit recommenders.**



+ Random	0.1650	0.3358	0.0257	0.0354	0.0762	0.0604
+ GPT3.5	0.1693	0.3462	0.0408	0.0536	0.0817	0.0694
+ RecAgent	0.1650	0.3393	0.0386	0.0518	0.0802	0.0668
+ RAH	0.1597	0.3340	0.0391	0.0542	0.0867	0.0719
+ UGen	<b>0.1899</b>	<b>0.3722</b>	<b>0.0555</b>	<b>0.0752</b>	<b>0.1140</b>	<b>0.0952</b>
Imp.% over LightGCN	2.82%	2.59%	32.14%	12.24%	28.67%	25.76%

Table 4: Human Evaluation on Steam

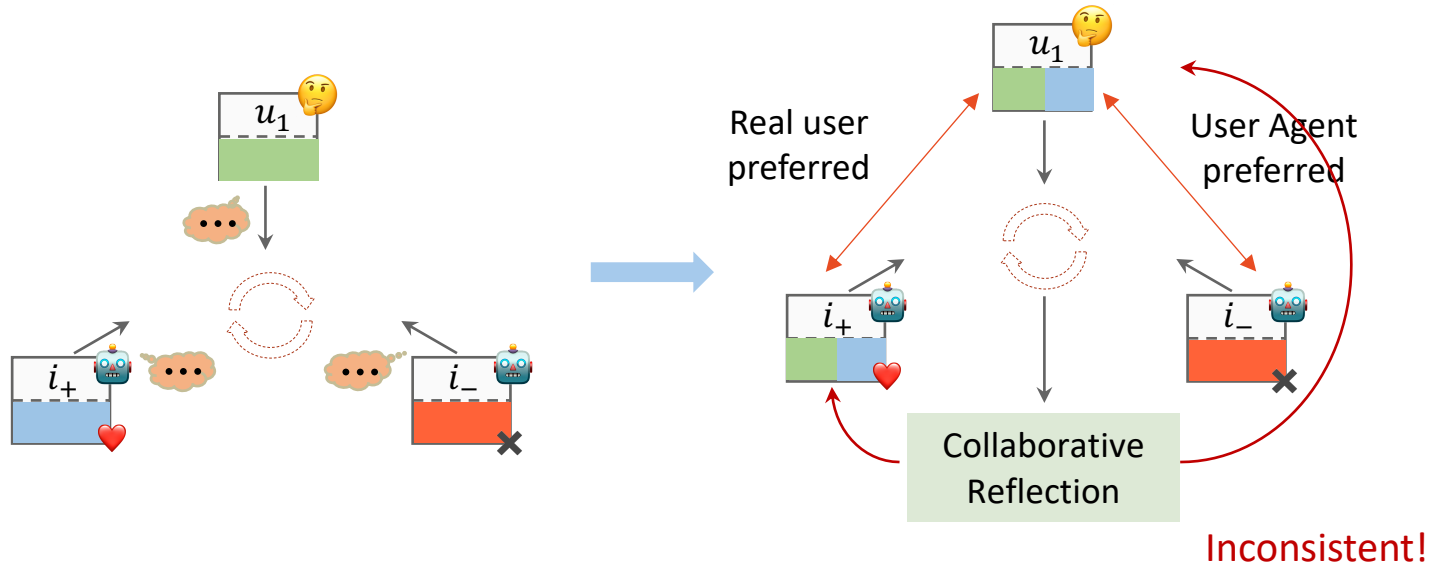
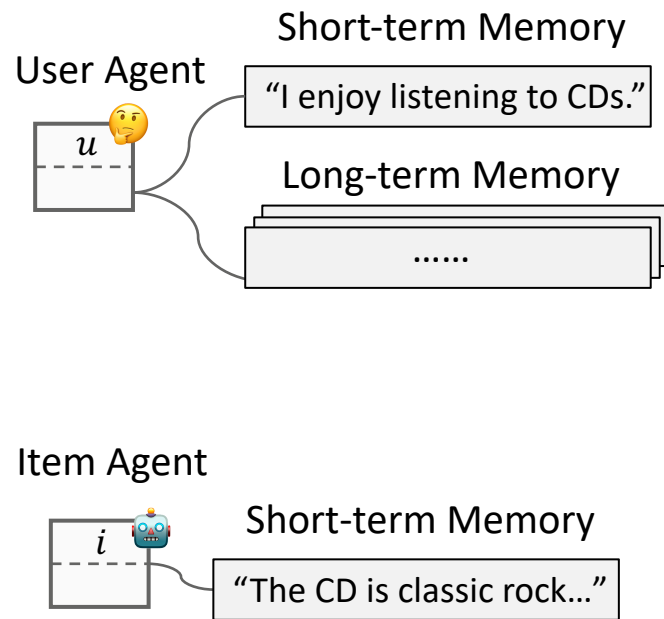
	Random	Pop	MF	MF+Full	MF+Human
Average Rank	4.72	3.22	2.61	2.50	<b>1.94</b>

### Agents as Users & Items

AgentCF: text-based collaborative learning

Key Points:

- Can LLM-powered Agent simulate collaborative signals/user-item interactions?



### Agents as Users & Items

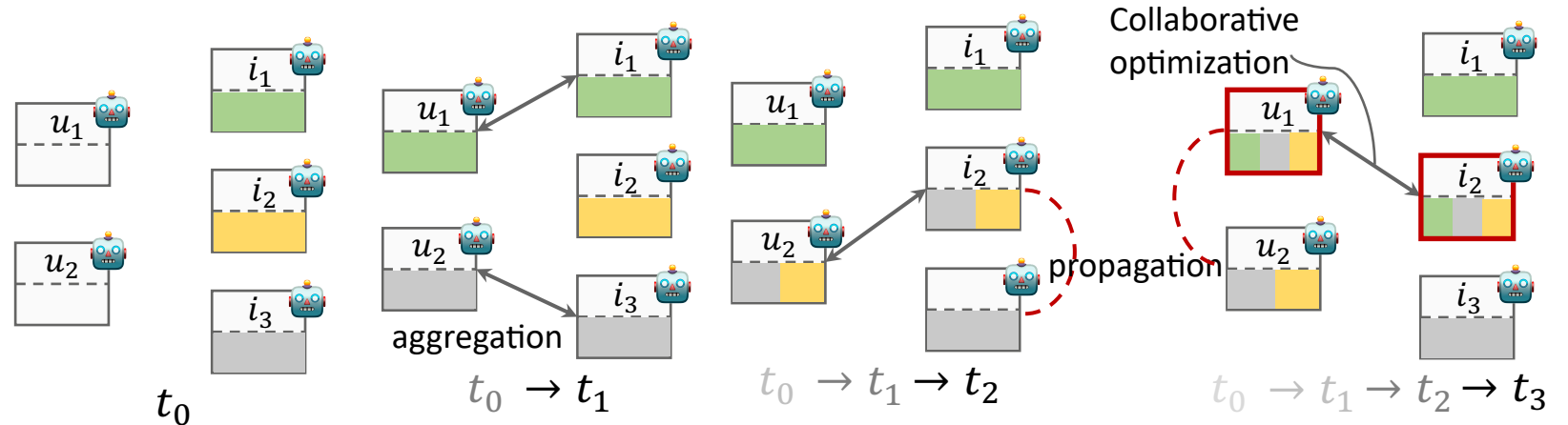
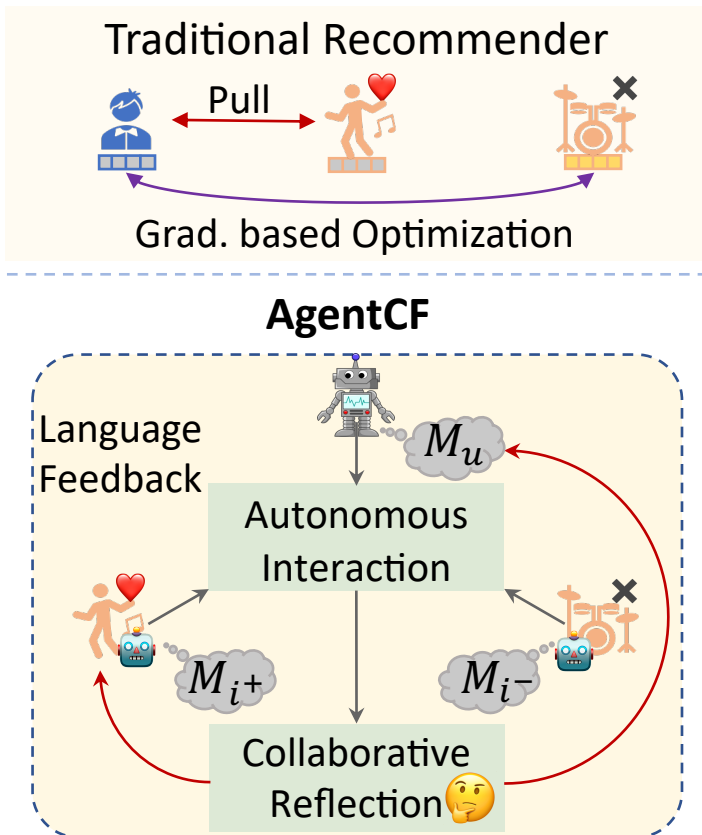
#### AgentCF: text-based collaborative learning

#### Key Points:

- Can LLM-powered Agent simulate collaborative signals/user-item interactions?

Real World: Bought

- Key idea:** Parameter-free text-based collaborative optimization.



### Key Observations:

- Agents are capable of simulating user-item interactions.

Method	CDs <sub>sparse</sub>			CDs <sub>dense</sub>			Office <sub>sparse</sub>			Office <sub>dense</sub>		
	N@1	N@5	N@10	N@1	N@5	N@10	N@1	N@5	N@10	N@1	N@5	N@10
BPR <sub>full</sub>	0.1900	0.4902	0.5619	0.3900	0.6784	0.7089	0.1600	0.3548	0.4983	0.5600	0.7218	0.7625
SASRec <sub>full</sub>	0.3300	0.5680	0.6381	0.5800	0.7618	0.7925	0.2500	0.4106	0.5467	0.4700	0.6226	0.6959
BPR <sub>sample</sub>	0.1300	0.3597	0.4907	0.1300	0.3485	0.4812	0.0100	0.2709	0.4118	0.1200	0.2705	0.4576
SASRec <sub>sample</sub>	<u>0.1900</u>	0.3948	<u>0.5308</u>	0.1300	0.3151	0.4676	0.0700	0.2775	0.4437	<b>0.3600</b>	<b>0.5027</b>	<b>0.6137</b>

Agents can faithfully **simulate user-item interactions**.

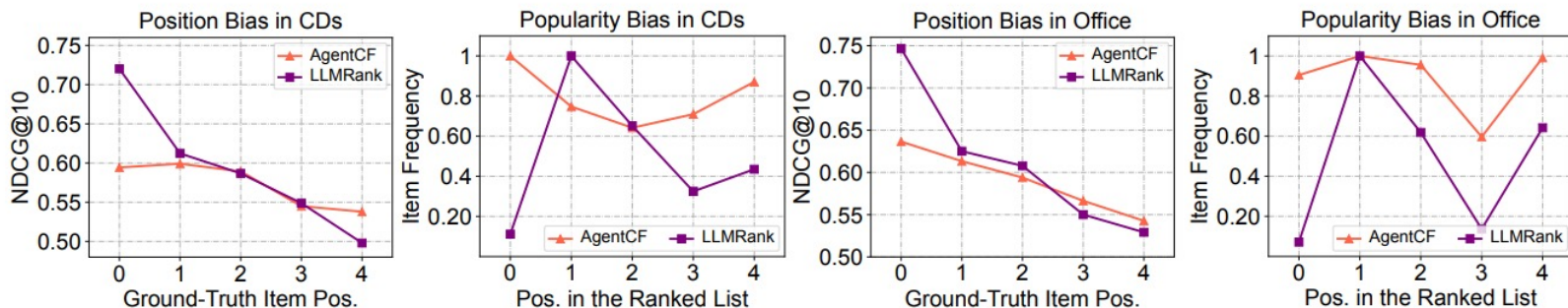
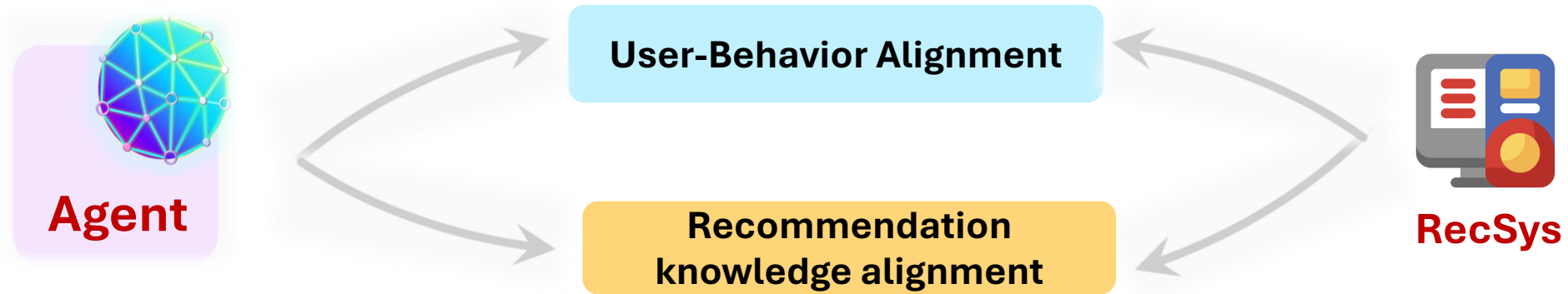


Figure 2: Analysis of whether our approach can simulate personalized agents to mitigate position bias and popularity bias.



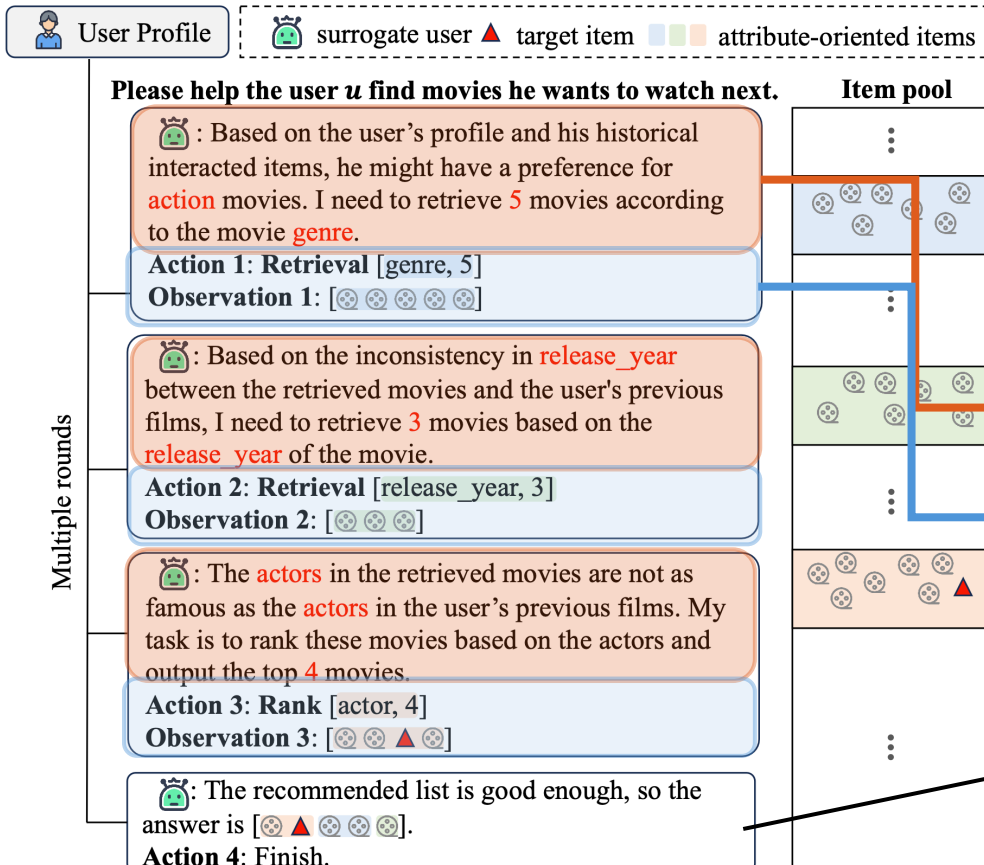
- LLM-empowered have potentials to solve long-standing problems in recommendation
  - Can an LLM-powered Agent faithfully simulate **users**?
    - **Agent4Rec, UGen, AgentCF, RecAgent**
  - Can an LLM-powered Agent be a better **recommender** with recommendation-specific knowledge?

### Agent as Recommender

### ToolRec: Tool-enhanced LLM-based recommender

#### Key Points:

- Can Agents **Utilize External Tools** to Enhance Recommendations?



#### Key Idea:

- Use **LLMs** to understand current contexts and preferences, and apply **attribute-oriented tools** to find suitable items.

#### Two stages:

- Learning Preferences:** LLM-based surrogate user learns user preferences and makes decisions
- Exploration of Items:** uses attribute-oriented tools to explore a wide range of items

❖ Process finishes when the LLM-based surrogate user is satisfied with the item list

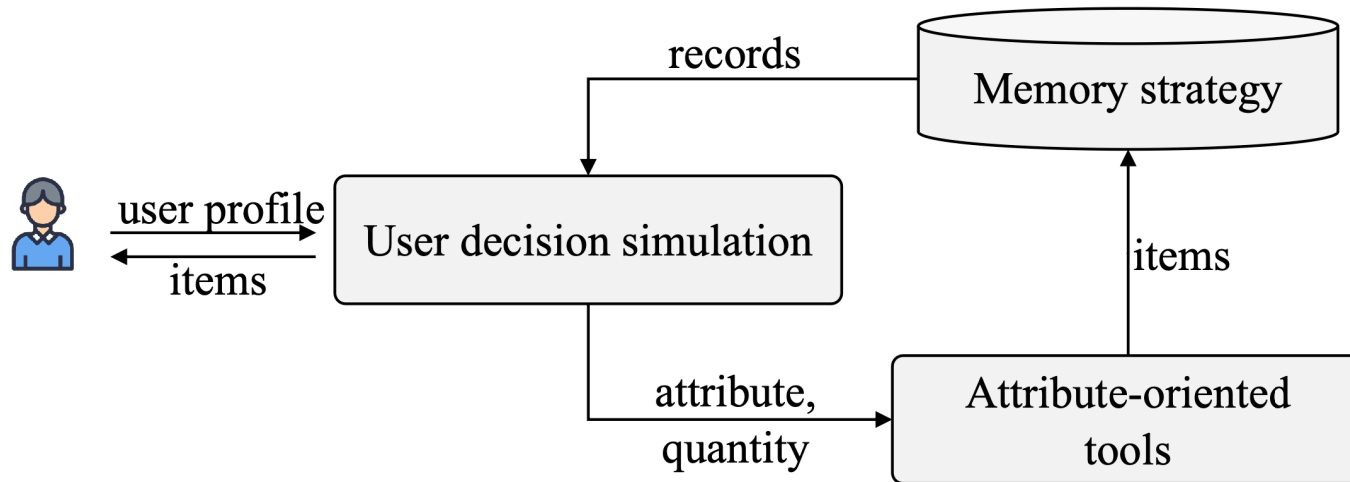


### Agent as Recommender

#### ❑ ToolRec: Tool-enhanced LLM-based recommender

#### ▪ Key Points:

- Can Agents **Utilize External Tools** to Enhance Recommendations?



- **LLMs** as the central controller, simulating the user decision.
- **Attribute-oriented Tools**: rank tools & retrieval tools.
- **Memory strategy** can ensure the correctness of generated items and cataloging candidate items.

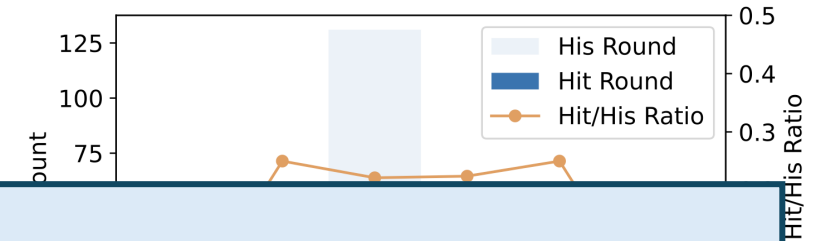
# Agent as Recommender

## ToolRec

### Key Observations:

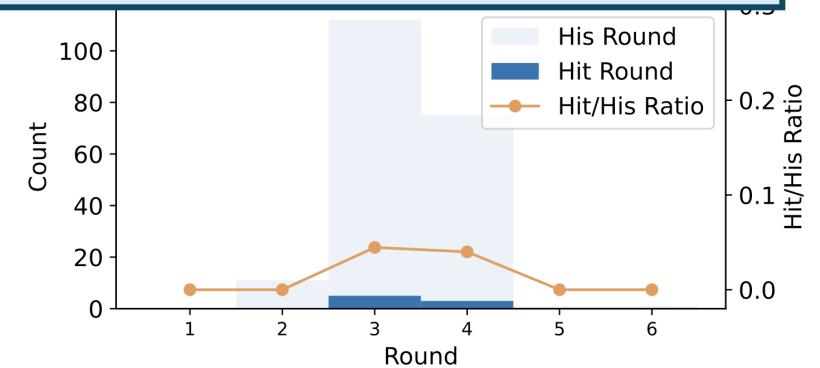
- Benefiting from rank tools and retrieval tools, ToolRec **excels** on the ML-1M and Amazon-Book datasets compared to baseline recommenders, demonstrating that it can **better align with the users' intent**.

	ML-1M		Amazon-Book		Yelp2018	
	Recall	NDCG	Recall	NDCG	Recall	NDCG
SASRec	0.203±0.047	0.1017±0.016	0.047±0.015	0.0205±0.006	0.030±0.005	0.0165±0.006



**Agents Utilizing External Tools can Enhance Recommendations.**

ToolRec	0.215±0.044	0.1171±0.018	0.053±0.013	0.0259±0.005	0.028±0.003	0.0159±0.001
ToolRec <sub>B</sub>	0.185±0.018	0.0895±0.002	0.043±0.013	0.0223±0.008	0.025±0.005	0.0136±0.009
Improvement	3.36%	15.10%	14.28%	5.14%	-29.16%	-27.32%



(b) Amazon-Book.

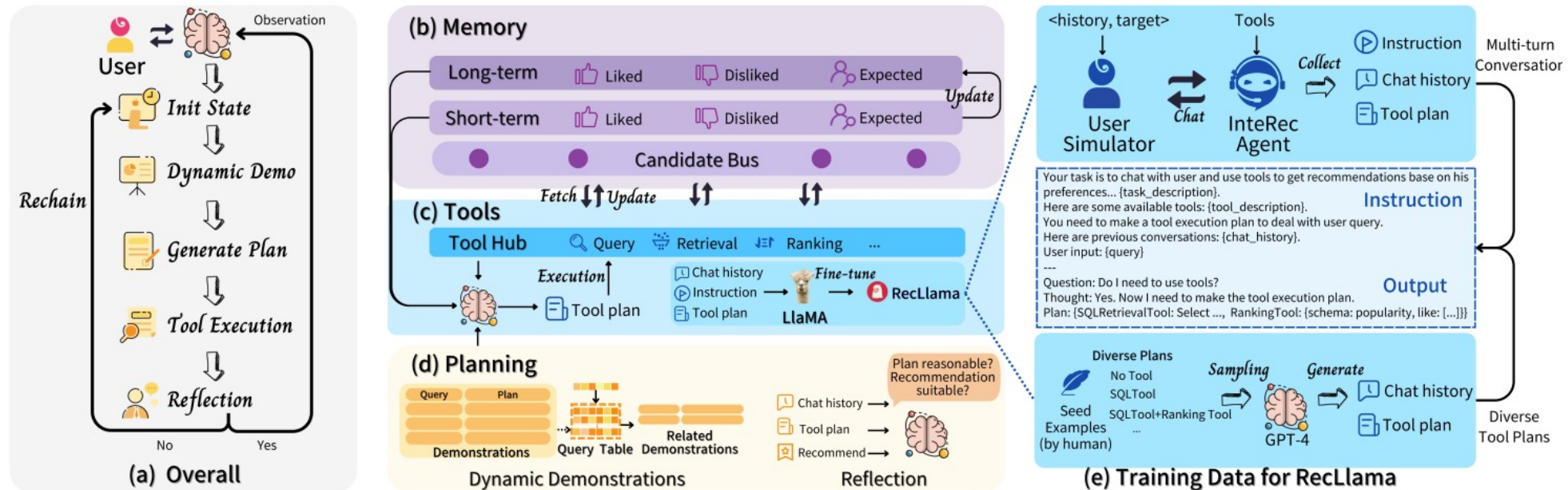
- ToolRec shows subpar performance on the Yelp2018 dataset - **local (niche) businesses**.
- Most processes **conclude** in three or four rounds, indicating that the LLM can understand user preferences **after a few iterations**.

### Agent as Recommender

#### InteRecAgent: Interactive Recommender.

#### Key Points:

- Agents can create a **versatile** and **interactive** recommender system.



- InteRecAgent** enables traditional recommender systems, such as those ID-based matrix factorization models, to become interactive systems with a natural language interface.

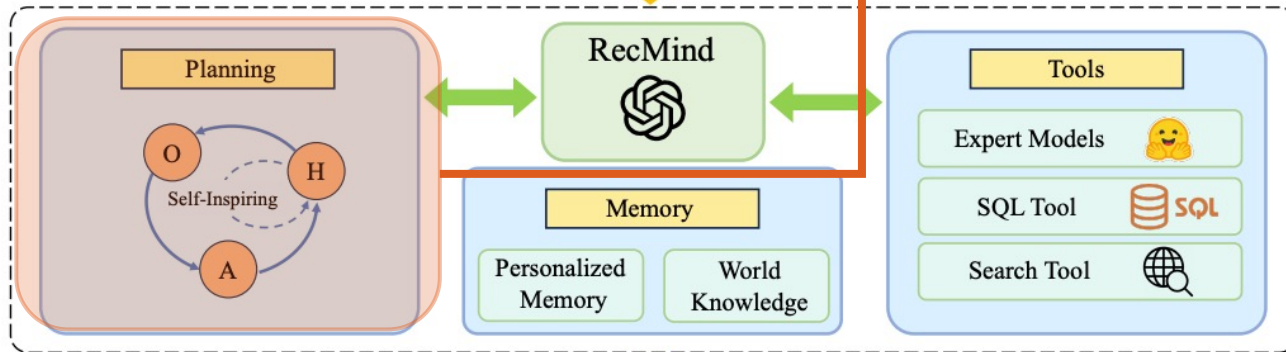
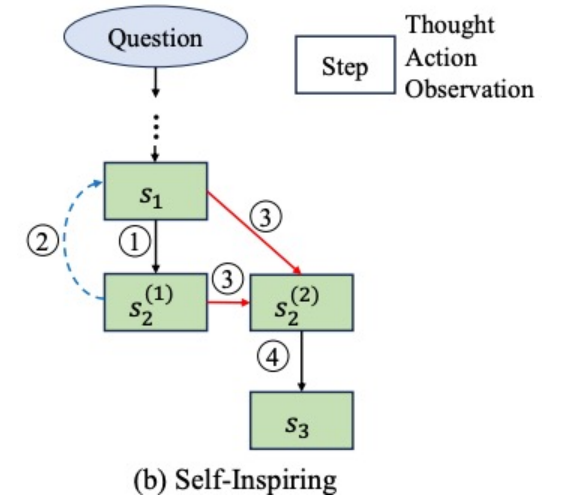
### Agent as Recommender

RecMind: Recommender agent with Self-Inspiring planning ability

Key Points:

- Can Agents with self-inspiring planning Enhance Recommendations?

Rating Prediction	Direct Recommendation	Sequential Recommendation	Review Summarization	Explanation Generation
How will <b>user_X</b> rate the item "Kusco-Murphy Tart Hair"? The rating should be an integer between 1 to 5, with 1 being lowest and 5 being highest.	From the item candidates listed below, choose the top 10 items to recommend to <b>user_X</b> and rank them in order of priority from highest to lowest. Candidates: ["Rogaine Women Hair Regrowth Treatment", .....]	<b>user_X</b> has interacted with the following items in chronological order: ["Old Spice Body Wash Red Zone", .....] Please recommend the next item that the user might interact with. Choose the top 10 products to recommend in order of priority, from highest to lowest.	Write a review title to summarize the review from <b>user_X</b> to item "Chrome Razor and Shaving Brush Stand". The review is "The stand is more solid then I expected for the price. The shape of this stand allows me to hang the shaving brush over the soap bowl. I couldn't do that with stand I had gotten with the kit."	Help <b>user_X</b> to generate a 5-star explanation for item "FoliGrowth Hair Growth Supplement".



5	["Propidren by HairGenics", "Nutrafol Women's Balance Hair Growth Supplements, Ages 45 and Up", .....]	["Old Spice Hair Styling Pomade for Men", "Lume Whole Body Deodorant - Invisible Cream Stick - 72 Hour Odor Control", .....]	Great quality for good price.	This product is essential for growing and maintaining healthy hair! This is a product to be bought in bulk because you can never have enough of it.
---	--	--	-------------------------------	---

- Self-inspires:
- At each intermediate planning step, the agent “self-inspires” to consider all previously explored paths for the next planning, both generating alternative thoughts and backtracking.

### Agent as Rec Assistant

❑ RAH: Reflection-enhanced user alignment for Rec assistant

■ Key Points:

- Can Agents with **Learn-Act-Critic loop** comprehend a user's personality from their behaviors?

**Item:** Harry Potter and the Sorcerer's Stone (Movie)

**Description:** Harry Potter and the Sorcerer's Stone is the first film in the Harry Potter series based on the novels by J.K. Rowling. The story follows Harry Potter, a young wizard who discovers his magical heritage as .....

**Characteristic:** Fantasy, Adventure, Family-friendly, Magic, Wizardry, Coming-of-age, British film, .....

**Analyze User Comment:** In the user comment, the mention of the plot being "very mysterious" suggests the user appreciates the suspense and intrigue in the narrative. However, the user also points out some imprecise plots in .....

**Analyze User Action:** The user's action indicates liking.

(a) Perceive Agent

**Reflection:** If directly add newly learned personalities into the personality library, there will be some duplications in User Preference; there is no duplication in User Dispreference; there exit conflicts between User Preference and User Dispreference.

**Need Optimize Preference:** Yes

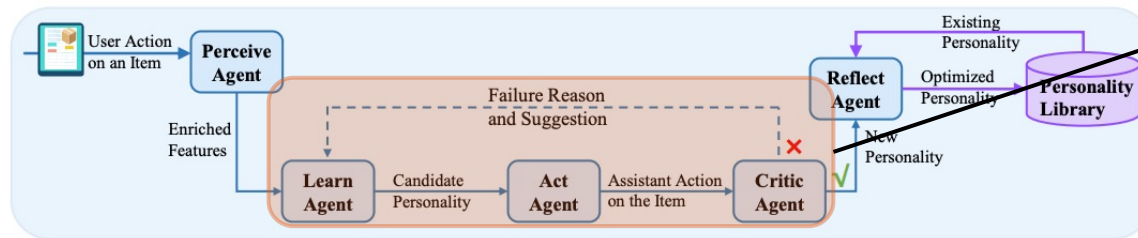
**Need Optimize Dispreference:** Yes

**How to Optimize Preference :** Merge similar preferences to avoid redundancy

**How to Optimize Dispreference :** Split the dispreference into more pieces to avoid conflicts.

**Results:**  
{Optimized Preference} & {Optimized Dispreference}

(e) Reflect Agent



(f) The process of the assistant to learn personalities from user actions.

**Analyze Why Like:** The movie offers an engaging storyline featuring magic, adventure, and coming-of-age themes, which could appeal to .....

**Analyze Why Dislike:** Some people might not like the movie if they are not fans of fantasy or magic-themed narratives. The movie's focus on a young protagonist and his friends might not be appealing to .....

**Learned Preference:** | Fantasy and Adventure themes | Mysterious and engaging plot | .....

**Learned Dispreference:** | Plot loophole | .....

(b) Learn Agent

**Guess Like:** The user may like the movie because it is a fantasy and adventure film based on a novel, with .....

**Guess Dislike:** The user may dislike the movie if they are not a fan of the specific style of British films or if they .....

**Analysis:** Based on the user's preferences for fantasy and adventure themes, the user may like the movie. However, since the user may also dislike the movie because .....

**User Comment (Predicted) :** The fantasy and adventure elements kept me engaged, while .....

**User Action:** { Like, Dislike or Neutral }

(c) Act Agent

✓ **Critic:** The predicted action is correct

-----

✗ **Critic:** The predicted action is wrong

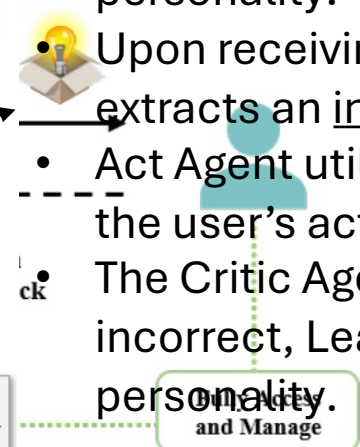
**Reasons:** The possible reason is that the user's preference is too general and thus can not provide an strong evidence regarding to the item. And the dispreference can be .....

**Suggestions:** Learn from the user interaction again, extract more specific preferences, and .....

(d) Critic Agent

❖ Learn-Act-Critic Loop:

- Learn Agent collaborates with the Act and Critic Agents in an **iterative process** to grasp the user's personality.
- Upon receiving user feedback, Learn Agent extracts an initial personality as a candidate.
- Act Agent utilizes this candidate as input to predict the user's actual action.
- The Critic Agent then assesses the accuracy. If incorrect, Learn Agent refines the candidate's personality.

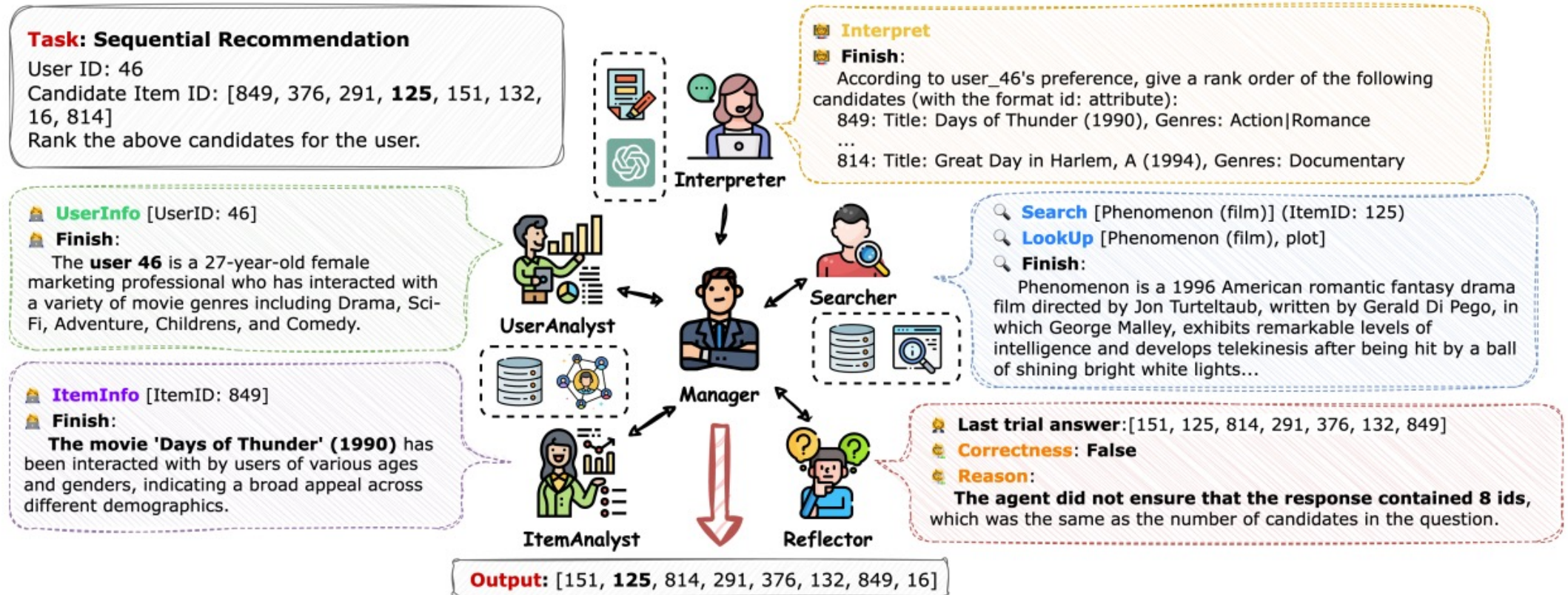


### Multi-Agent as Recommender

MACRec: enhance RecSys through multi-agent collaboration

Key Points:

- Multi-agents with different **roles** work collaboratively to tackle a specific recommendation task.

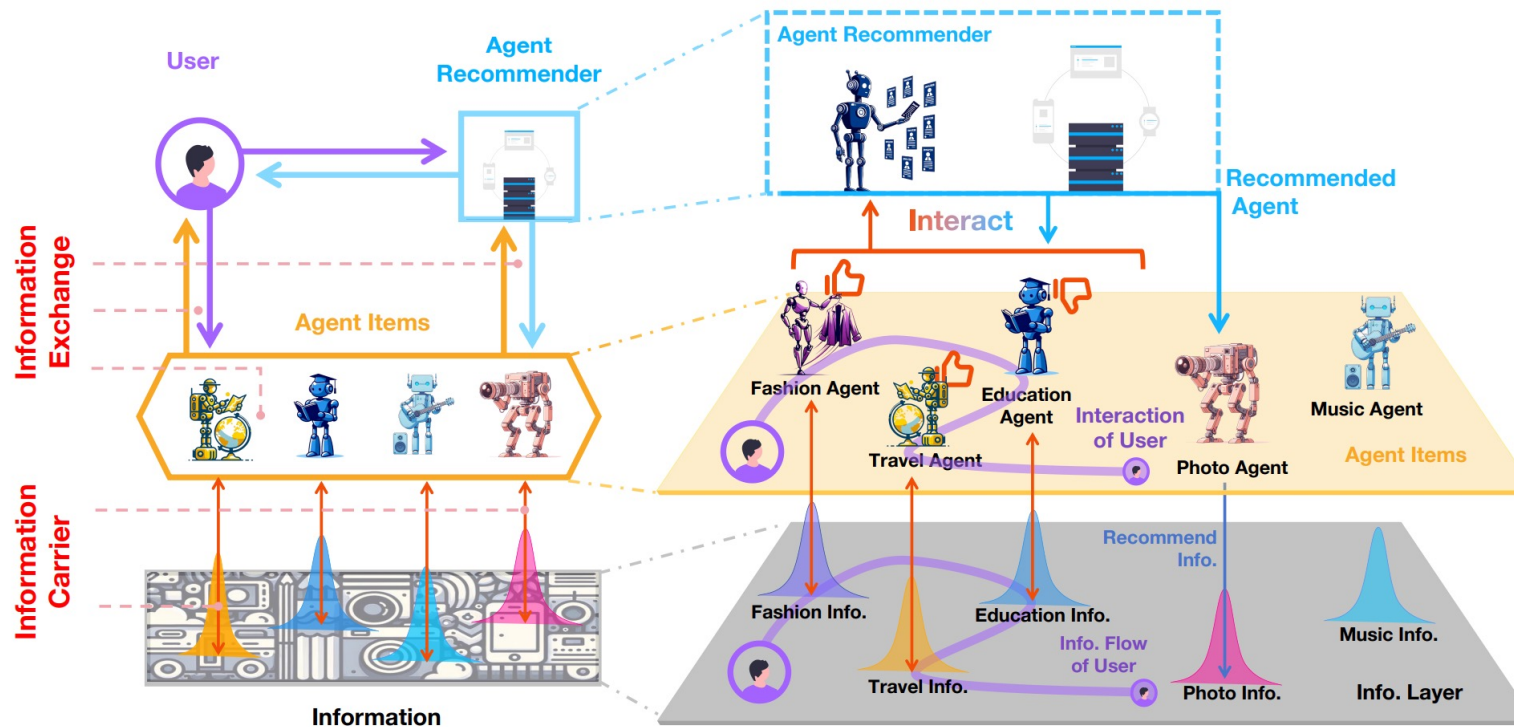


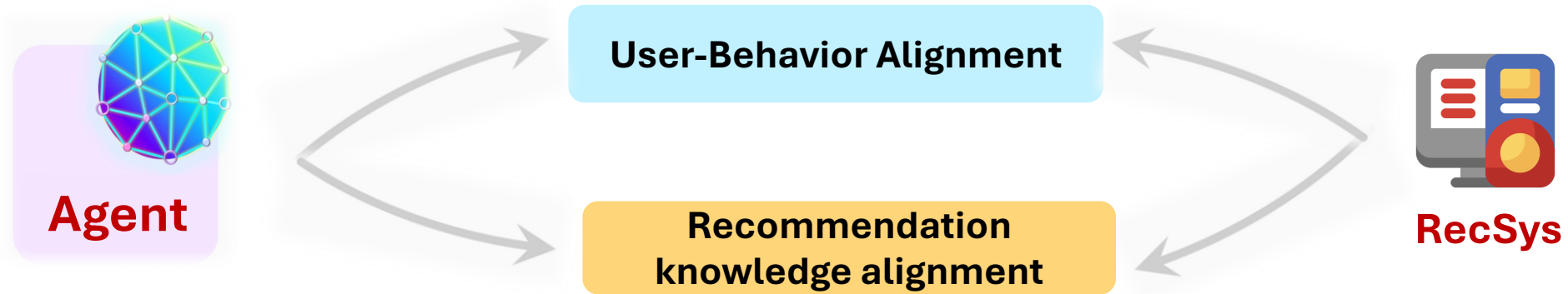
### Agent Recommender

□ Rec4Agentverse: Agent recommender for Agent platform

■ Key Points:

- Treating LLM-based Agents in Agent platform as items in the recommender system.
- **Agent Recommender** is employed to recommend personalized Agent Items for each user.



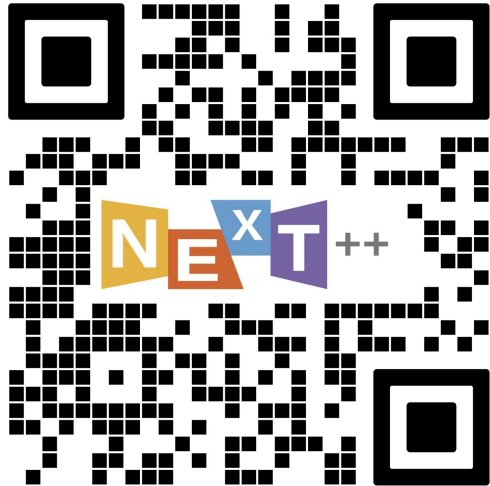


- LLM-empowered have potentials to solve long-standing problems in recommendation
  - Can an LLM-powered Agent faithfully simulate **users**?
    - **Agent4Rec, UGen, AgentCF, RecAgent**
  - Can an LLM-powered Agent be a better **recommender** with recommendation-specific knowledge?
    - **ToolRec, InteRecAgent, RecMind, RAH, MACRec, Rec4Agentverse**



# Thanks for listening!

**Email:** [an\\_zhang@nus.edu.sg](mailto:an_zhang@nus.edu.sg)



An Zhang's Homepage



Resources