

1 A Appendix

2 A.1 Broader Impacts

3 The benchmark and toolkit presented in this work are expected to advance the development of agentic
4 recommender systems and bring significant positive societal implications. The improved contextual
5 understanding of agentic recommender systems enables higher-quality personalized recommendations
6 that can benefit diverse domains, including e-commerce, short-video platforms, etc. Furthermore,
7 the standardized evaluation framework facilitates the development of more robust and adaptable
8 recommendation systems that can better serve evolving user needs. Our platform can also provide a
9 flexible foundation for incorporating ethical AI principles, such as fairness and privacy preservation,
10 into future recommendation systems. These advancements collectively contribute to building more
11 user-centric, explainable, and socially beneficial recommendation technologies.

12 A.2 Limitations

13 While AgentRecBench provides a comprehensive evaluation framework for agentic recommender
14 systems, several limitations point to valuable directions for future work. First, the current benchmark
15 primarily focuses on evaluating emerging agentic recommender systems, and we plan to incorporate
16 more traditional and deep learning-based baselines for more thorough comparative analysis. Second,
17 our environment currently operates on textual information, and we aim to extend it to incorporate
18 multimodal data (e.g., images, videos) to better reflect real-world recommendation scenarios where
19 agents need to process diverse content types. Finally, while the current framework evaluates single-
20 agent systems, we plan to extend it to support the evaluation of multi-agent recommendation systems
21 where collaborative or competitive agents interact to improve recommendation outcomes.

22 A.3 Supplemented experimental results of the cold-start recommendation tasks and 23 evolving-interest recommendation tasks

24 Tables 1 and 2 show the complementary experimental results of our task on DeepSeek V3 and
25 GPT-4o-mini. Overall, our evaluation results on other mainstream models, such as DeepSeek V3 and
26 GPT-4o-mini, demonstrate that the agent’s performance is consistent with the findings presented in
27 the main text. Baseline666, DummyAgent, and RecHackers all exhibit strong performance, further
28 validating the robustness and reliability of our evaluation method in accurately reflecting agent
29 capabilities.

30 A.4 Case Study

31 We present the core workflows of the two agents, DummyAgent, and RecHackers in Figure 1, and
32 2, respectively. Overall, these agents rely on similar types of information for the ranking process,
33 although their specific implementations differ. The ranking decisions are primarily based on three
34 key components: (1) historical user reviews, which reflect past preferences; (2) a list of candidate
35 items to be ranked; and (3) detailed item information, which helps evaluate the relevance of each
36 item to the user’s preferences.

37 A major innovation among these agents lies in item-side feature engineering. Notably, the Baseline666
38 team employed platform-specific feature extraction methods, enabling a robust and adaptable ranking
39 strategy across different data sources. For instance, on the Amazon platform, features such as item ID,
40 name, star rating, number of reviews, and item description were extracted. On Yelp, they focused on
41 core attributes like item ID, name, star rating, and review count. In contrast, the Goodreads platform
42 required more diverse features, including author, publication year, and similar books.

43 Review-side feature engineering represents another critical component, aimed at identifying and
44 extracting the most informative reviews to enrich the understanding of both user preferences and item
45 characteristics. For example, the DummyAgent team implemented a platform-tailored strategy: on
46 Yelp, they extracted not only the review text but also interactive attributes such as “useful,” “cool,”
47 and “funny”; on Amazon, they incorporated publication dates and purchase verification indicators;
48 and on Goodreads, in addition to the review text and ratings, they utilized metadata such as review
49 date, number of votes, number of comments, and reading status.

Table 1: Performance comparison on cold-start recommendation tasks with the average HR@N metric (N=1,3,5).

DeepSeek-V3							
Category	Method	Amazon		Goodreads		Yelp	
		User	Item	User	Item	User	Item
Agentic RS	BaseAgent	16.3	14.0	22.3	16.2	4.0	4.0
	CoTAgent	19.3	9.0	20.3	14.5	4.3	4.3
	MemoryAgent	15.3	11.0	16.7	15.5	4.3	4.3
	CoTMemAgent	16.7	12.0	16.7	15.5	3.7	4.3
	Baseline666	50.3	48.7	38.7	49.5	1.3	2.7
	DummyAgent	59.0	45.0	37.7	49.5	1.3	2.7
	RecHackers	59.7	47.0	49.3	46.1	3.0	3.3
	Agent4Rec	45.6	28.0	37.3	11.1	2.7	0.7
GPT-4o-mini							
Category	Method	Amazon		Goodreads		Yelp	
		User	Item	User	Item	User	Item
Agentic RS	BaseAgent	15.3	9.0	18.7	15.2	3.0	4.0
	CoTAgent	17.0	9.3	20.0	15.5	2.7	4.3
	MemoryAgent	15.0	7.3	16.0	18.2	3.0	4.3
	CoTMemAgent	14.7	7.7	15.3	16.0	3.7	4.0
	Baseline666	46.0	32.0	33.0	21.5	0.3	2.3
	DummyAgent	36.7	30.7	32.7	24.6	0.3	3.3
	RecHackers	39.0	36.3	37.7	33.3	1.3	3.3
	Agent4Rec	36.0	20.3	39.3	7.1	0.7	1.3

Table 2: Performance comparison on evolving-interest recommendation tasks with the average HR@N metric (N=1,3,5).

DeepSeek-V3							
Category	Method	Amazon		Goodreads		Yelp	
		Long Term	Short Term	Long Term	Short Term	Long Term	Short Term
Agentic RS	BaseAgent	17.3	19.7	32.0	29.7	5.7	4.3
	CoTAgent	17.7	16.0	24.0	16.3	3.3	3.0
	MemoryAgent	18.3	17.7	29.3	29.7	5.0	4.0
	CoTMemAgent	13.3	18.7	23.3	21.3	4.3	4.0
	Baseline666	50.7	71.3	66.0	63.3	0.0	0.0
	DummyAgent	65.0	65.7	66.7	60.7	10.3	6.3
	RecHackers	64.3	68.0	68.7	66.3	9.3	7.7
	Agent4Rec	34.0	46.3	41.3	42.7	10.0	6.0
GPT-4o-mini							
Category	Method	Amazon		Goodreads		Yelp	
		Long Term	Short Term	Long Term	Short Term	Long Term	Short Term
Agentic RS	BaseAgent	13.3	12.0	12.7	14.0	5.3	4.7
	CoTAgent	13.3	11.7	12.7	13.0	5.3	4.7
	MemoryAgent	12.0	10.7	16.3	15.7	5.3	4.3
	CoTMemAgent	11.3	12.3	17.7	16.3	5.3	4.7
	Baseline666	35.0	49.0	35.0	38.7	0.0	0.0
	DummyAgent	34.7	48.0	42.7	31.3	4.0	2.3
	RecHackers	39.7	46.3	53.7	44.3	5.7	2.0
	Agent4Rec	24.0	34.7	32.7	35.3	7.0	4.3

50 In summary, the key design elements across these agents can be distilled into three core principles:
 51 (1) effective workflows are built on a combination of user history, candidate items, item details,
 52 and platform-specific features, all integrated through large language models (LLMs) to produce
 53 rankings; (2) extracting representative, platform-specific item attributes is essential for enhancing
 54 model performance; and (3) prioritizing reviews that are both highly relevant and information-rich is
 55 crucial for improving ranking quality.

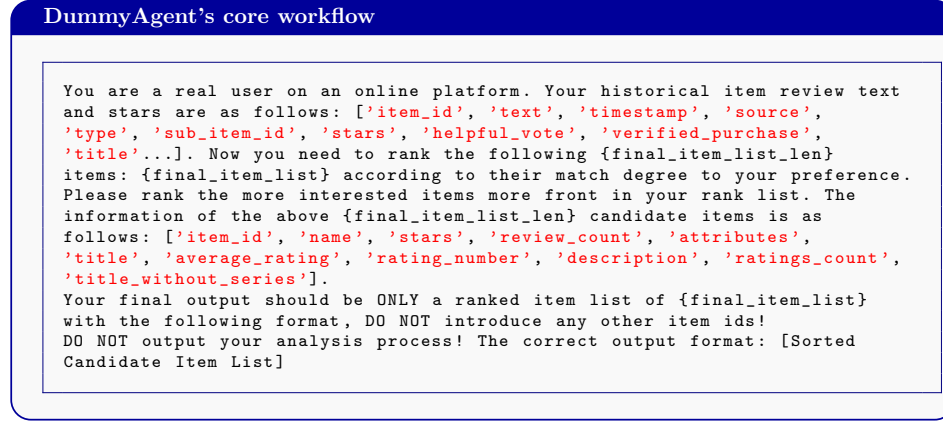


Figure 1: Illustration of the agentic workflow of a superior recommendation agent (DummyAgent), which conducts domain-adaptive item-side feature engineering to enhance personalization.

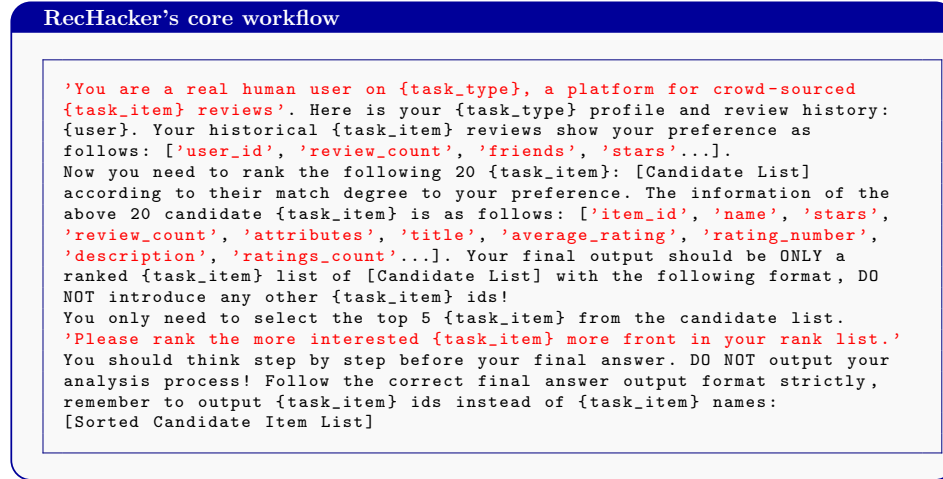


Figure 2: Illustration of the agentic workflow of a superior recommendation agent (RecHacker), which conducts domain-adaptive item-side feature engineering to enhance personalization.