NeurIPS 2025

R^2ec: Towards Large Recommender Models with Reasoning

Runyang You, Yongqi Li, Xinyu Lin, Xin Zhang, Wenjie Wang, Wenjie Li, Liqiang Nie

The Hong Kong Polytechnic University • National University of Singapore • University of Science and Technology of China • Harbin Institute of Technology (Shenzhen)

Research Motivation

Reasoning LLMs Bring New Frontiers to Recommendation

- DeepSeek-R1, GPT-o1, etc. achieve large gains via test-time compute on math & coding
- Can recommender models reap the same benefit?, i.e., **think to recommend?**

Bridging Recommendation and Reasoning Requires Novel Solutions

While existing approaches have begun exploring LLM reasoning for recommendations, they typically treat reasoning as an **external auxiliary module** that augments conventional recommendation pipelines, which suffers from:

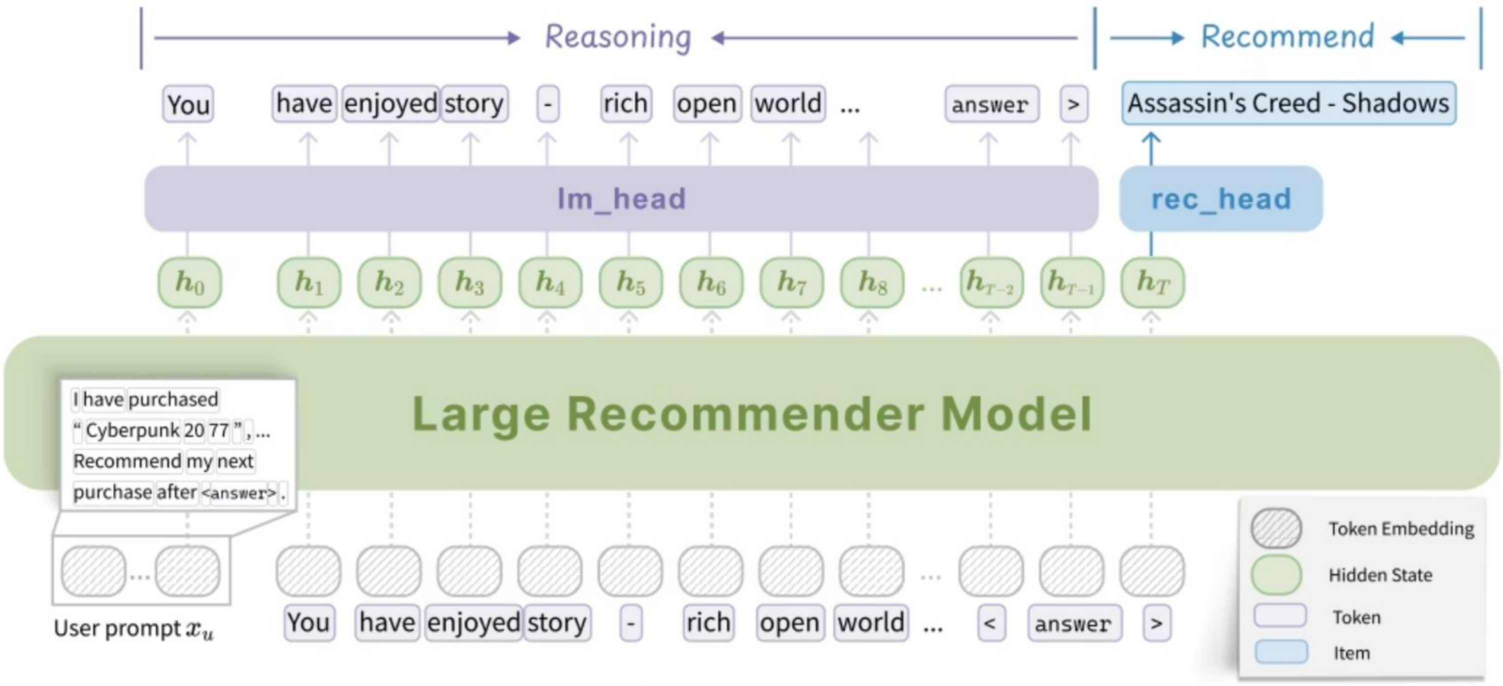
Disjointed Reasoning and Recommendation Processes

Inference Efficiency Bottlenecks

Data Scarcity for Recommendation Reasoning

→ We need a **unified large recommender model** that intrinsically incorporates reasoning capabilities within a single architecture, and an optimization strategy to **jointly optimize reasoning-then-recommendation** without reasoning annotations

R^2ec: Dual-Head Architecture



Language-Modeling Head

Generates reasoning tokens through autoregressive decoding.

Recommendation Head

Scores items efficiently with item embeddings encoded by the model itself.

Inference

1. **Encode user context and preferences**

   User prompt → shared hidden states

2. **Generate reasoning trajectory**

   Reasoning tokens reveal decision logic

3. **Score items efficiently**

   Final hidden state → semantic item matching

Recommendation Policy Optimization

Without labeled reasoning data, RecPO learns effective reasoning strategies directly from recommendation signals using reinforcement learning.

Trajectory Sampling

Sample diverse reasoning sequences using top-K sampling with controlled temperature for stochastic exploration.

Fused Rewarding

Combine discrete ranking rewards (NDCG) with continuous similarity scores to introduce granular learning signals.

Joint RL Objective

Treat "reasoning-then-recommend" as a single RL trajectory, with only highest-advantage sequences contributing to final recommendation updates.

Reward Design: Balancing Signals

The fused reward scheme elegantly combines two complementary signals:

Discrete Rewards R_d

NDCG@k metrics based on ground-truth item rankings. Provides strong alignment with final recommendation quality.

Continuous Rewards R_c

Softmax similarity scores providing fine-grained learning signals that guide the model through diverse reasoning paths.

$$R = \beta R_c + (1 - \beta) R_d$$

This balance enables the model to explore diverse reasoning strategies while maintaining focus on recommendation accuracy.

Joint Training Objective

The entire "reasoning-then-recommend" sequence is treated as a single RL trajectory, symbolically represented as:

$$x_u \xrightarrow{\pi_\theta} o_1 \xrightarrow{\pi_\theta} \dots \xrightarrow{\pi_\theta} o_T \xrightarrow{\pi_\theta} v^+$$

Here, x_u denotes the user input, o_i represents the *i*-th token of reasoning, and v^+ signifies the final recommended item.

$$\pi_\theta(v^+ | x_u, o_i) = \frac{\exp(s_\theta(v^+))}{\sum_{v \in B} \exp(s_\theta(v))}$$

where B represents the batch of candidate items.

$$\mathcal{J}(\theta) = \mathbb{E}_{\{u,v^+\} \sim \mathcal{D}, \{o_i\}_{i=1}^{G} \sim \pi_{\theta_{\mathrm{old}}}(\cdot | x_u)} \frac{1}{G} \sum_{i=1}^{G} \left[\sum_{t=1}^{T_i} \ell_\epsilon(r_{i,t}(\theta), A_i) + \delta_{i,i^*} \ell_\epsilon(r_{i,T+1}(\theta), A_i) \right]$$

- PPO clipped-ratio loss employed.
- Only the **best** trajectory (max advantage) back-props
- **Every** sampled trajectory updates token-level policy

Main Results

Comprehensive evaluation across multiple domains demonstrates consistent superiority.

		Instruments						CDs and Vinyl						Video Games					
	Method	H@5	N@5	H@10	N@10	H@20	N@20	H@5	N@5	H@10	N@10	H@20	N@20	H@5	N@5	H@10	N@10	H@20	N@20
	GRU4Rec	0.0171	0.0135	0.0193	0.0142	0.0201	0.0144	0.0067	0.0037	0.0104	0.0041	0.0156	0.0051	0.0109	0.0070	0.0181	0.0093	0.0301	0.0123
	Caser	0.0109	0.0141	0.0115	0.0149	0.0127	0.0155	0.0045	0.0029	0.0067	0.0037	0.0089	0.0042	0.0124	0.0083	0.0191	0.0103	0.0279	0.0126
	SASRec	0.0175	0.0144	0.0201	0.0162	0.0223	0.0210	0.0076	0.0104	0.0081	0.0119	0.0086	0.0141	0.0129	0.0080	0.0206	0.0105	0.0326	0.0135
	TIGER	0.0171	0.0128	0.0184	0.0132	0.0193	0.0134	0.0067	0.0045	0.0097	0.0055	0.0156	0.0069	0.0123	0.0085	0.0222	0.0116	0.0323	0.0142
Qwen	BigRec	0.0052	0.0033	0.0111	0.0052	0.0189	0.0072	0.0045	0.0025	0.0089	0.0039	0.0141	0.0052	0.0008	0.0004	0.0016	0.0006	0.0128	0.0034
	D^3	0.0042	0.0020	0.0094	0.0037	0.0192	0.0062	0.0082	0.0057	0.0141	0.0076	0.0253	0.0104	0.0054	0.0028	0.0104	0.0044	0.0197	0.0067
	LangPTune	0.0127	0.0083	0.0224	0.0115	0.0348	0.0145	0.0074	0.0053	0.0156	0.0080	0.0208	0.0094	0.0049	0.0027	0.0088	0.0040	0.0140	0.0140
	R^{2}ec	**0.0237***	**0.0154***	**0.0374***	**0.0198***	**0.0615***	**0.0259***	**0.0513***	**0.0372***	**0.0647***	**0.0414***	**0.0818***	**0.0457***	**0.0288***	**0.0185***	**0.0532***	**0.0264***	**0.0827***	**0.0337***
	% Improve.	35.43%	6.94%	66.96%	22.22%	52.61%	23.33%	46.57%	58.30%	37.95%	51.09%	20.83%	40.62%	42.36%	34.05%	51.13%	41.29%	31.56%	33.53%
Gemma	BigRec	0.0068	0.0048	0.0101	0.0058	0.0130	0.0066	0.0030	0.0030	0.0052	0.0037	0.0119	0.0053	0.0156	0.0105	0.0260	0.0138	0.0430	0.0182
	D^3	0.0072	0.0038	0.0202	0.0080	0.0339	0.0114	0.0216	0.0129	0.0327	0.0164	0.0446	0.0194	0.0117	0.0068	0.0210	0.0141	0.0478	0.0224
	SDPO*	0.0066	0.0034	0.0098	0.0054	0.0144	0.0071	0.0022	0.0018	0.0037	0.0025	0.0162	0.0094	0.0166	0.0122	0.0298	0.0155	0.0466	0.0222
	Llara*	0.0078	0.0055	0.0137	0.0074	0.0159	0.0080	0.0097	0.0039	0.0127	0.0049	0.0202	0.0152	0.0275	0.0173	0.0428	0.0223	0.0677	0.0299
	SPRec	0.0070	0.0033	0.0111	0.0062	0.0142	0.0077	0.0029	0.0022	0.0037	0.0025	0.0124	0.0063	0.0152	0.0113	0.0244	0.0133	0.0566	0.0211
	LangPTune	0.0130	0.0079	0.0221	0.0107	0.0403	0.0152	0.0350	0.0235	0.0469	0.0274	0.0677	0.0325	0.0068	0.0053	0.0120	0.0059	0.0195	0.0094
	R^{2}ec	**0.0264***	**0.0161***	**0.0397***	**0.0203***	**0.0615***	**0.0257***	**0.0573***	**0.0398***	**0.0804***	**0.0472***	**0.1042***	**0.0527***	**0.0326***	**0.0205***	**0.0531***	**0.0271***	**0.0835***	**0.0347***
	% Improve.	50.86%	11.81%	77.23%	25.31%	52.61%	22.38%	63.71%	69.36%	71.43%	72.26%	53.91%	62.15%	18.98%	19.19%	24.07%	21.52%	23.34%	16.25%

Table 1: The overall performance of baselines and R²ec on three datasets. The best results in each group are marked in Bold, while the second-best results are underlined. * implies the improvements over the second-best results are statistically significant (p-value < 0.05). % improve represents the relative improvement achieved by R²ec over the best baseline

Ablation Study

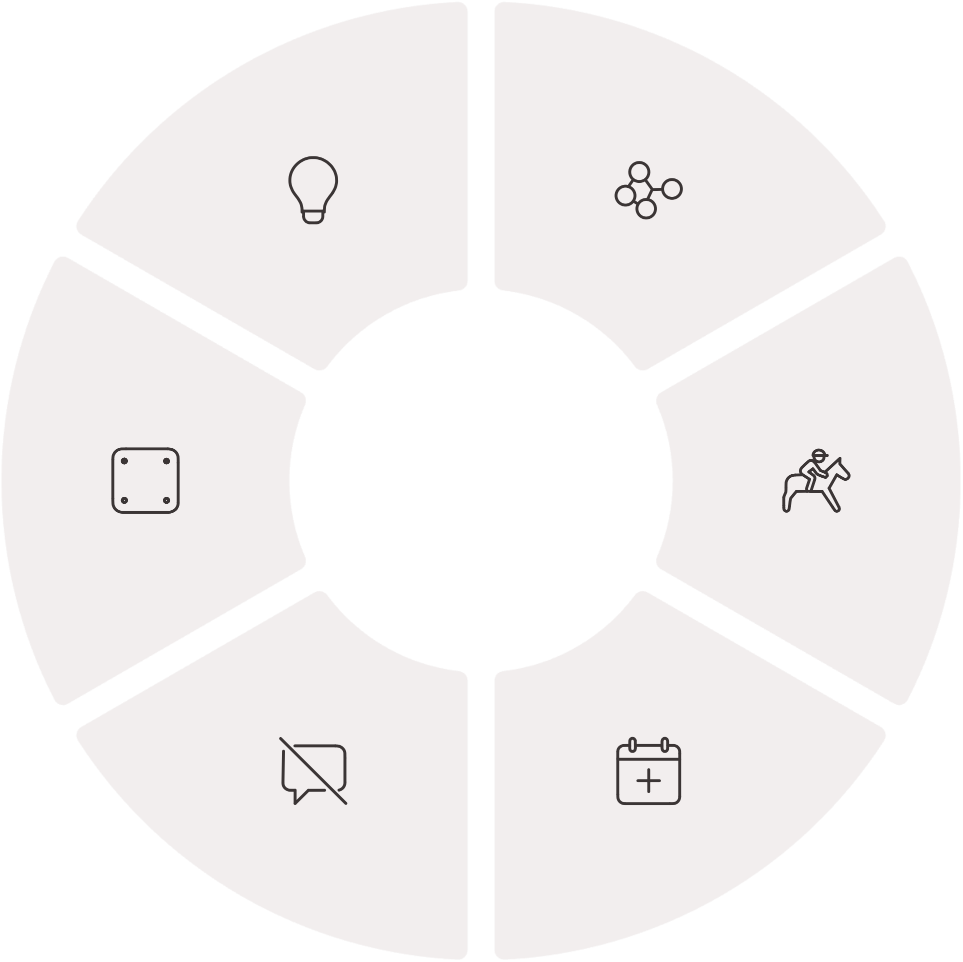
	Instruments						CDs and Vinyl						Video Games					
Method	H@5	N@5	H@10	N@10	H@20	N@20	H@5	N@5	H@10	N@10	H@20	N@20	H@5	N@5	H@10	N@10	H@20	N@20
w/ ClsHead	0.0044	0.0023	0.0102	0.0033	0.0179	0.0067	0.0030	0.0025	0.0045	0.0027	0.0095	0.0044	0.0012	0.0008	0.0022	0.0011	0.0133	0.0032
w/o Reasoning	0.0176	0.0121	0.0296	0.0153	0.0511	0.0200	0.0469	0.0321	0.0692	0.0393	0.0945	0.0456	0.0277	0.0174	0.0441	0.0227	0.0748	0.0303
w/o R_d	0.0198	0.0124	0.0338	0.0164	0.0560	0.0224	0.0521	0.0338	0.0766	0.0404	0.0974	0.0486	0.0302	0.0196	0.0487	0.0254	0.0798	0.0332
w/o R_c	0.0244	0.0160	0.0394	**0.0208**	0.0605	**0.0258**	0.0543	0.0382	0.0774	0.0456	0.1012	0.0515	0.0316	0.0202	**0.0534**	0.0264	0.0814	0.0355
R^{2}ec	**0.0264**	**0.0161**	**0.0397**	0.0203	**0.0615**	0.0257	**0.0588**	**0.0388**	**0.0804**	**0.0457**	**0.1086**	**0.0525**	**0.0326**	**0.0205**	0.0531	**0.0271**	**0.0853**	**0.0363**

Table 2: Ablation study on key components of R^{2}ec.

- **Reasoning Impact**: Removing reasoning tokens resulted in an average 15% performance drop across all metrics, confirming the substantial benefit of explicit reasoning for recommendations.
- **Architectural Coupling**: Using a separate classification head instead of the tightly-coupled recommendation head led to significantly worse performance, highlighting the importance of shared hidden-state spaces.
- **Reward Design**: The fused reward scheme outperformed using either discrete or continuous rewards alone, with discrete rewards showing stronger alignment to recommendation objectives.

Emergent Reasoning Strategies

R²ec can adaptively adopt reasoning strategies based on context and domain characteristics:



Attribute Abstraction

Identifying and generalizing item features

Pattern Recognition

Grouping similar items and preferences

Negative Exclusion

Explicitly avoiding unfavorable types

Scenario Reasoning

Context-aware, role-based recommendations

Self-Explanation

Articulating preference rationales

Temporal Reasoning

Time-based patterns and trends

Domain-Specific Adaptation

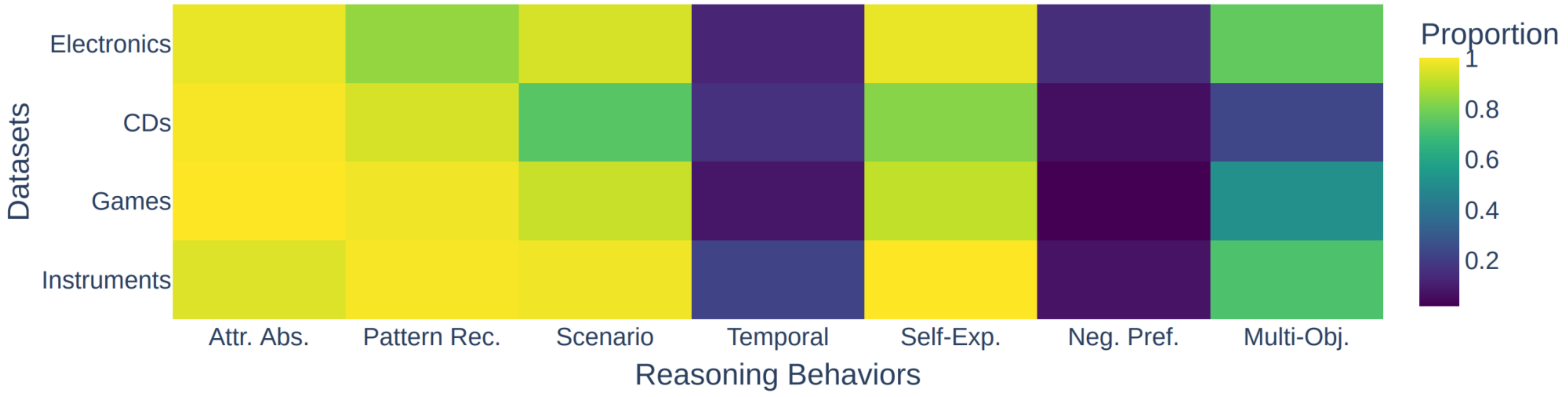


Figure 4: Distribution of reasoning behaviors across datasets. Each bar represents the proportion of reasoning outputs exhibiting a given reasoning behavior within a dataset.

Insight: This adaptive reasoning demonstrates R²ec's ability to self-organize its decision-making process based on domain characteristics and user contexts, leading to improved interpretability and more appropriate recommendations.

Efficiency Analysis

Method	Latency (s)
SASRec	0.014
LangPTune	1.90
D^3	4.62
LLaRA	5.23
R^2ec	**1.67**
R^2ec (with VLLM)	**0.0945**

Table 4: Average inference latency (in seconds) across models.

Architectural Advantages

- Outperforms non-reasoning Large recommender models
- Dual-head design avoids expensive autoregressive item decoding

Deployment with VLLM

- Significantly reduces efficiency gap with traditional sequential models
- Maintains expressiveness of reasoning-enhanced recommendations

Result: Competitive inference efficiency among LLM-based recommenders while preserving superior performance

Summary

- **R²ec**: Unified large recommender model with intrinsic reasoning capabilities
- **RecPO**: Learns effective reasoning strategies without human annotations
- **Performance**: Superior recommendation quality with competitive efficiency
- **Adaptability**: Self-organizes reasoning strategies across domains

Takeaway

Reasoning and recommendation can be effectively unified in a single model, achieving both performance and efficiency.

Thank You!

Access All Resources:

Scan the QR code to explore the details and reproduce our findings on Paper page - R^2ec: Towards Large Recommender Models with Reasoning

- Full Research Paper
- GitHub Repository (Code & Data)
- Model Checkpoints