MemRec: Collaborative Memory-Augmented Agentic Recommender System

Bridging Isolated Agentic Memory with Dynamic Collaborative Graphs

Weixin Chen
https://weixinchen.com

Joint work between:

Hong Kong Baptist University (HCI-RecSys Group)
Rutgers University (WISE Lab)
SNAP Research (User Modeling & Personalization Team)

March 3, 2026

Agenda

1. **Motivation:** The Evolution of RecSys and the Missing "Collaboration"
2. **The Dual Challenges:** Cognitive Overload & Update Bottlenecks
3. **The MemRec Framework:** Architectural Decoupling
4. **Main Results:** Performance across Benchmarks
5. **Industrial Insights:** Cost-Efficiency, Deployment, & Robustness
6. **Conclusion & Q&A**

Motivation: The Paradigm Shift in RecSys

The evolution of preference storage and reasoning:

- **CF Era:** Sparse rating matrices. (Pure collaborative, no semantics)
- **Deep Learning Era:** Dense latent embeddings. (Implicit representation)
- **Agentic Era (Emerging):** Semantic memory powered by LLMs. (Explicit reasoning)

The Promise of Agentic RecSys

Memory is transformed into a semantic format, enabling LLMs to perform complex reasoning, follow natural language constraints, and evolve their understanding over time.

The Critical Gap: Isolated Memory

Current SOTA Agents (e.g., iAgent, AgentCF)

- Rely on **isolated, self-reflective memory**.
- They update understanding based only on direct interactions.
- **The Flaw:** They miss the most potent signal in RecSys: **Global Collaboration**.

Analogy: Trying to recommend books by only asking what the user read before, completely ignoring what millions of similar readers enjoy.

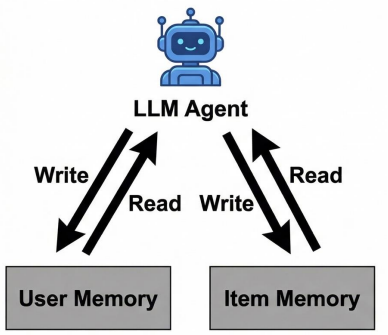


Figure 1: Current isolated memory paradigm.

Why not just feed all neighbors to the LLM?

Injecting raw collaborative neighborhoods into an agent's memory seems intuitive, but it fails in practice due to **Two Dual Challenges**:

1. Cognitive Overload (Retrieval Phase)

The sheer volume of textual and structural signals from raw neighbors exceeds context windows. It bombards the LLM with noise, causing severe information bottlenecks and hallucinations.

2. Prohibitive Computational Cost (Update Phase)

Synchronously propagating updates throughout the graph requires independent LLM calls for every interaction. This scales linearly $O(|N|)$ and destroys online inference efficiency.

The Solution: Architectural Decoupling

Core Idea: Separate memory management from high-level reasoning.

The "Brain" (LLM_{Rec})

- Focuses purely on **Grounded Reasoning**.
- Uses heavy models (e.g., GPT-4o).
- Takes synthesized, high-signal context.

The "Librarian" (LM_{Mem})

- Dedicated to **Graph Management**.
- Uses fast, cost-effective models (e.g., 4o-mini).
- Curates, synthesizes, and asynchronously propagates updates.

System Architecture Overview

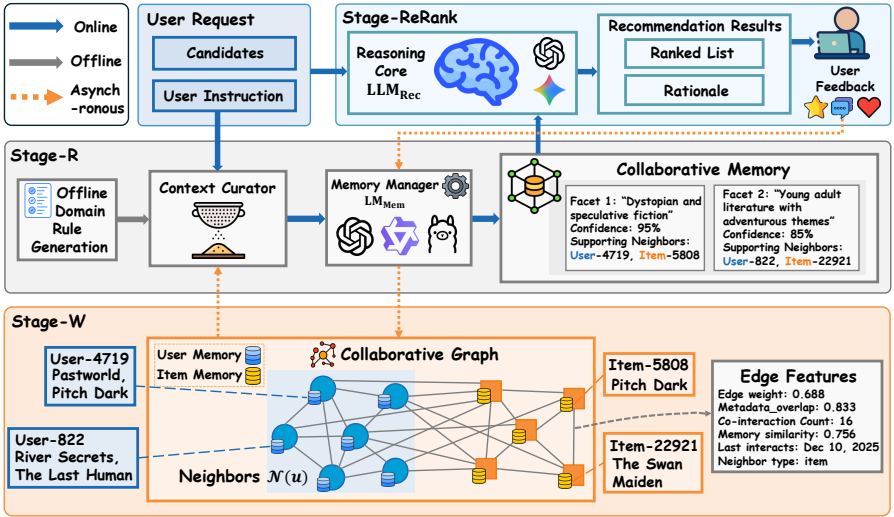


Figure 2: The MemRec Three-Stage Pipeline.

Stage-R: Collaborative Memory Retrieval

Based on **Information Bottleneck (IB) theory**, we adopt a "Curate-then-Synthesize" strategy to minimize cognitive load.

Step 1: LLM-Guided Context Curation

- **Offline Rule Generation:** LM_{Mem} analyzes domain statistics $\mathcal{D}_{\text{domain}}$ to generate interpretable rules R_{domain}:
$$R_{\text{domain}} \leftarrow \text{LM}_{\text{Mem}}(\mathcal{D}_{\text{domain}} \parallel P_{\text{meta}})$$
- **Online Curation:** These rules act as a high-speed filter (milliseconds) to select the top-k most relevant neighbors $N'_k(u)$:
$$N'_k(u) = \text{Curate}(\mathcal{N}(u), R_{\text{domain}}, k)$$
- *Advantage:* Faster than neural scorers, more adaptive than hard-coded heuristics.

Stage-R: Memory Synthesis

Step 2: Synthesizing Collaborative Facets

- LM_{Mem} distills raw representations of curated neighbors $\text{Rep}(N'_k)$ into a compact collaborative memory M_{collab} (Preference Facets):

$$M_{\text{collab}} = \{F\} \leftarrow \text{LM}_{\text{Mem}}(\text{Rep}(N'_k) \parallel M_u^{t-1} \parallel P_{\text{synth}})$$

Example Output from LM_{Mem}

Facet 1: "Preference for dystopian fiction with strong world-building."
Confidence: 95%
Supporting Evidence: [User-4719, Item-5808]

This acts as the highly compressed, maximum-information context for the downstream ranker.

Stage-ReRank: Grounded Reasoning

- The "Brain" (LLM_{Rec}) executes the final reasoning over candidate item memories C_{info}:

$$\{s_i, r_i\}_{i=1}^{N} \leftarrow \text{LLM}_{\text{Rec}}(\mathcal{I}_u \parallel M_{\text{collab}} \parallel C_{\text{info}} \parallel P_{\text{rerank}})$$

- **Outputs:** Relevance Score $s_i \in [0, 1]$ + Natural Language Rationale r_i.

Why this matters?

By grounding the generation in M_{collab}, the rationale is factually supported by **broader community evidence**, significantly improving relevance and reducing hallucinations.

Stage-W: Asynchronous Collaborative Propagation

A static graph cannot capture evolving trends. But synchronous updates block the system.

Our Optimization: $O(1)$ Call Complexity

- Inspired by Label Propagation. When User u interacts with Item i_c, we execute a **single, batched asynchronous update**:

$$M_u^t, M_{i_c}^t, \{\Delta M_{\text{neigh}}\} \leftarrow \text{LM}_{\text{Mem}}(M_{\text{collab}} \parallel M_u^{t-1} \parallel M_{i_c}^{t-1} \parallel N_k'(u) \parallel P_{\text{update}})$$

- This simultaneously updates:
  1. The User's Memory (M_u^t)
  2. The Item's Memory ($M_{i_c}^t$)
  3. The Memories of relevant Neighbors ($\{\Delta M_{\text{neigh}}\}$)
- Ensures continuous graph evolution without disrupting the online interaction flow.

Qualitative Case Study (1/2): Grounded Reasoning

Target: User 2057 (Sparse history: Fan of YA Fantasy).
Instruction: *"I seek a visually immersive experience... stunning visuals... like a* **graphic novel***."*

Step 1: Stage-R (Collaborative Synthesis)

- **Noisy Input:** Raw histories of 16 neighbors (e.g., *Pastworld*, *The Lighthouse Land*).
- **Synthesized Facets:** The community strongly appreciates **dystopian fiction** and **complex, emotionally layered storytelling**.

Step 2: Stage-ReRank (Grounded Recommendation)

- **Recommendation:** *Attack on Titan: No Regrets*
- **Rationale:** "This is a **graphic novel** (matching your format request) known for its stunning visuals. It strongly resonates with the community's interest in **dystopian settings** and **complex storytelling** (matching collaborative facets)."

Qualitative Case Study (2/2): Asynchronous Evolution

Step 3: Stage-W (Graph Propagation)

Trigger Event: User 2057 clicks *Attack on Titan*.

How does the memory graph evolve in the background?

- **1. User Memory Update (User 2057):**
  - **[New Insight]:** Confirmed strong interest in the "**graphic novel**" format combined with **complex, emotional storytelling**.
- **2. Item Memory Update (Attack on Titan):**
  - **[New Insight]:** Validated its appeal for YA fantasy readers seeking **immersive dystopian worlds**.
- **3. Neighbor Propagation (User-4023):**
  - **[Implicit Evolution]:** Even without direct interaction, this neighbor's memory is enriched: "*appreciates* **strong character development and emotional depth**".

Experimental Setup

Datasets (from InstructRec):

- **Books:** Sparse, stable content-driven preferences.
- **Goodreads:** Very dense, strong social/community effects.
- **MovieTV:** Sparse, recency-critical.
- **Yelp:** Context-rich, strong categorical constraints.

Baselines:

- **Traditional:** LightGCN, SASRec, P5.
- **No Memory Agents:** Vanilla LLM.
- **Static Memory Agents:** iAgent.
- **Dynamic Isolated Agents:** RecBot, AgentCF, i^2Agent.

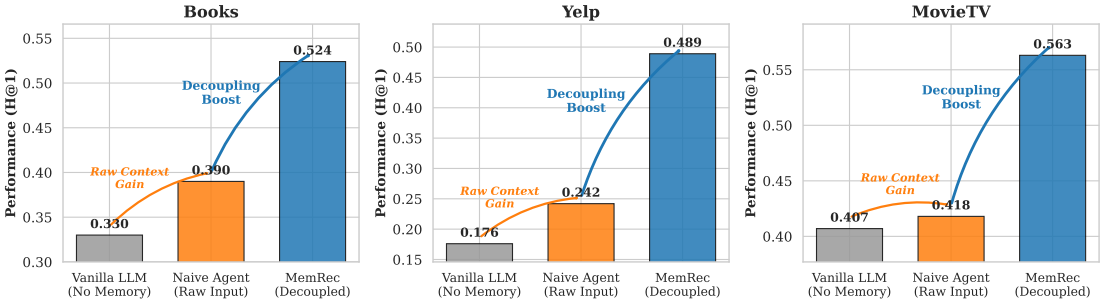
Main Results (Books & Goodreads)

Model	Books (Sparse) H@1	Books (Sparse) N@3	Books (Sparse) H@5	Goodreads (Dense) H@1	Goodreads (Dense) N@3	Goodreads (Dense) H@5
LightGCN	0.1753	0.2596	0.5703	0.2499	0.4432	0.7903
Vanilla LLM	0.3138	0.4533	0.7270	0.2864	0.3948	0.7390
Static iAgent	0.3925	0.4858	0.6905	0.2617	0.3954	0.6591
Dynamic i^2Agent	0.4453	0.5649	0.7708	0.3099	0.4825	0.7675
MemRec (Ours)	**0.5117**	**0.6152**	**0.8007**	**0.3997**	**0.5540**	**0.8052**
Improvement	+14.9%	+8.9%	+3.8%	+28.9%	+14.8%	+1.8%

- Decisive state-of-the-art performance across metrics.
- Largest gains seen in **Hit@1**, proving the precision brought by collaborative signals.

Architectural Necessity: Breaking the Bottleneck

Why decoupling is required: MemRec vs. Naive Collaborative Agent.



- **Vanilla LLM (Gray):** Baseline with no memory.
- **Naive Agent (Orange):** Ingests raw neighbor text directly. Hits a hard performance plateau due to noise and cognitive overload.
- **MemRec (Blue):** Our decoupled "Curate-then-Synthesize" approach successfully unlocks the true value of the graph.

Industrial Focus I: Deployment Flexibility & Cost

We evaluated MemRec's modularity by swapping backend models. It does **not** strictly require expensive proprietary models for every step.

Configuration	LLM_{Rec} / LM_{Mem}	Books H@1	Cost Profile	Key Takeaway
Ceiling	gpt-4o / 4o-mini	**0.580**	High	*Peak performance*
Standard	4o-mini / 4o-mini	0.524	Low	*Strong collaborative baseline*
Cloud-OSS	4o-mini / OSS-120B	<u>0.561</u>	Medium	**Near-ceiling w/ moderate cost**
Local-Qwen	4o-mini / Qwen-2.5-7B	0.470	Fixed*	*Best on-premise performance*
Vector Rerank	*Vector* / 4o-mini	0.209	<u>Low</u>	**Ultra-fast, pluggable reranker**

* Marginal cost per query is negligible (hardware amortization/electricity only).

- **Insight:** Swapping LM_{Mem} to **Cloud-OSS** yields near-ceiling results. Meanwhile, **Local-Qwen** provides a highly viable on-premise, privacy-preserving alternative with fixed costs.

Industrial Focus I: Token Economy (Pricing Asymmetry)

Commercial LLM APIs have **asymmetric pricing** (Output tokens cost 3x-4x more than Input tokens).

Pipeline Stage	Primary Role	Avg Input Tokens	Avg Output Tokens	I/O Ratio
Stage-R (Synthesis)	LM_{Mem}	$\sim$2,800	$\sim$400	**7.0 : 1**
Stage-W (Async Update)	LM_{Mem}	$\sim$3,800	$\sim$700	**5.4 : 1**
Stage-ReRank	LLM_{Rec}	$\sim$1,500	$\sim$500	3.0 : 1

- MemRec naturally exploits this pricing structure. The heavy lifting (processing raw graph data) is intensely **Input-biased**.
- The effective real-world cost is significantly lower than naive token counts suggest.

Industrial Focus II: Robustness to Niche Users

Question: Does collaborative memory fail for "cold-start" or highly niche users?
Answer: No. They are its greatest beneficiaries.

Sub-group analysis based on user interaction density (Books dataset):

Activity Group	Vanilla LLM	MemRec	Relative Improv.
Low (Niche: Avg 8.5 items)	**0.246**	**0.472**	**+91.4%**
Medium	0.311	0.509	+63.8%
High	0.380	0.571	+50.3%

- When personal history is extremely sparse, MemRec leverages the broader community graph to fill the "understanding gap".

Industrial Focus II: Robustness to Noisy Neighbors

Question: Can irrelevant neighbors or click-bait cause "negative propagation"?
Stress Test: We artificially injected random noise (fake interactions) into target user histories before the retrieval stage.

Noise Ratio Injected	Hit@1	Drop vs Baseline
0% (Baseline)	0.527	-
10% Noise	0.528	+0.19%
20% Noise	0.481	-8.73%
30% Noise	**0.491**	**-6.83%**

- MemRec's performance does **not** collapse.
- **Why?** The "Curate-then-Synthesize" pipeline inherently acts as a **robust noise buffer**. The LLM curator filters out semantic outliers before they reach the ranker.

Ablation Study: Are all components necessary?

We systematically removed core components to validate their individual contributions (evaluated on Books subset):

Model Configuration	H@1	H@3	N@3	H@5	N@5	Drop (H@1)
MemRec (Full Framework)	**0.527**	**0.713**	**0.634**	**0.803**	**0.670**	-
w/o Collab. Write *(Static Graph)*	0.505	0.702	0.619	0.814	0.665	-4.2%
w/o LLM Curation *(Heuristics Only)*	0.498	0.685	0.606	0.788	0.648	-5.5%
w/o Collab. Read *(Isolated Memory)*	0.475	0.650	0.575	0.769	0.624	**-9.9%**

- **Collab. Read (Retrieval):** The most critical mechanism. Reverting to isolated memory causes a severe 9.9% drop, proving the necessity of global signals.
- **LLM Curation:** Removing zero-shot LLM guidance and relying purely on rigid graph heuristics introduces noise (-5.5%).
- **Collab. Write (Propagation):** Crucial for refining top-tier precision (H@1) by capturing newly emerging community trends.

Hyperparameter Robustness: Finding the Sweet Spot

- **Number of Neighbors (k):** Controls the *breadth* of collaborative signals.
  - **Insight:** Too few neighbors degrade the system back to isolated memory. Too many introduce noise and trigger cognitive overload.
- **Number of Facets (N_f):** Controls the *compression level* during synthesis.
  - **Insight:** Requires balancing information retention (avoiding over-compression) against focus (avoiding verbosity).
- **Conclusion:** The framework exhibits a stable performance "sweet spot" around $k \in \{16, 32\}$ and $N_f = 7$, proving it does not require extreme fine-tuning to perform well.

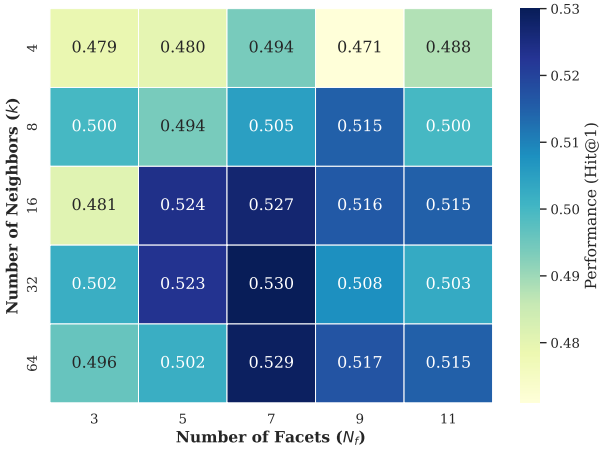


Figure 3: Performance (Hit@1) w.r.t. k and N_f.

Qualitative Impact: Rationale Quality

Using GPT-4o as an automated judge (1-5 scale) across three dimensions:

- **Specificity:** Significantly richer item details than isolated memory ($p < 0.001$).
- **Relevance:** Grounding in peer experience makes rationales feel significantly more convincing to users.
- **Factuality:** Grounding in real collaborative memories **reduces hallucinations** compared to ungrounded generation.

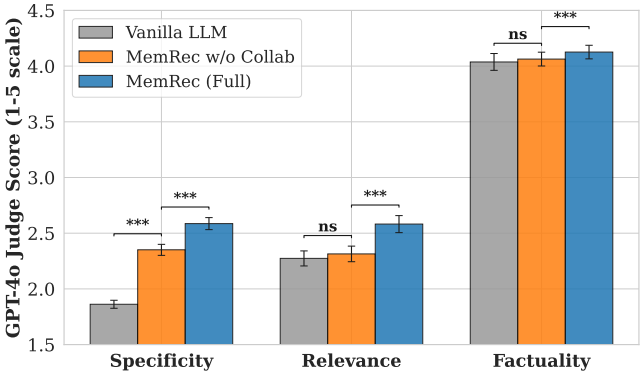


Figure 4: GPT-4o Judge Evaluation (95% CIs).

Conclusion

- **A Paradigm Shift:** MemRec pioneers the transition from isolated, self-reflective memory to a **Dynamic Collaborative Memory Graph** in Agentic RecSys.
- **Architectural Decoupling:** Successfully resolves the Information Bottleneck by separating the "Librarian" (Curation/Synthesis) from the "Brain" (Reasoning).
- **Production Ready:**
  - Resolves graph update bottlenecks via $O(1)$ asynchronous propagation.
  - Highly robust to long-tail (niche) users and noisy data.
  - Establishes a flexible Pareto frontier for real-world deployment.

About Us: HKBU HCI-RecSys Group

Principal Investigator: Prof. Li Chen
`www.comp.hkbu.edu.hk/~lichen/HCI-RecSysGroup.html`

Our Mission: Bridging Human-Computer Interaction (HCI) and Recommender Systems (RS) to build next-generation, human-centric, and trustworthy AI recommendation platforms.

A Full-Stack Research Portfolio in the LLM Era:

- We do not just focus on accuracy. We investigate the entire lifecycle of user-system interaction, from intention elicitation to beyond-accuracy evaluation.
- **Rich Industry Collaboration:** Committed to solving real-world challenges (e.g., cold-start, sparsity, bias) through cutting-edge AI techniques.

HCI-RecSys Group at HKBU

Our Core Research Directions

Our lab pioneers several cutting-edge fronts in LLM-empowered RecSys:

1. Conversational & Agentic RS (Interactive AI)

- **LLM-based CRS:** Proactively resolving user uncertainty via Multi-modal and Emotion-aware dialogues (Ting Yang).
- **Agentic Memory (My Focus today):** Building collaborative memory architectures to empower autonomous reasoning agents (Weixin Chen).

2. User Simulation & Generative Evaluation

- **LLM-based User Simulators:** Replacing costly human studies by simulating complex user behaviors and serendipity evaluation (Li Kang).

3. Trustworthy & Beyond-Accuracy Objectives

- **Robustness & Fairness:** Handling massive unlabeled samples, mitigating cross-domain overlapping bias, and ensuring heterogeneous user fairness (Dr. Jiashi Gao, Yuhan Zhao).

Thank You!

Q & A

Paper available at: https://arxiv.org/pdf/2601.08816
Code available at: https://github.com/rutgerswiselab/MemRec
Demo available at: https://memrec.weixinchen.com/demo