MemRec: Collaborative Memory-Augmented Agentic Recommender System

Weixin Chen[1,2] Yuhan Zhao[1] Jingyuan Huang[2]
Zihe Ye[2] Clark Mingxuan Ju[3] Tong Zhao[3]
Neil Shah[3] Li Chen[1] Yongfeng Zhang[2]

[1]Hong Kong Baptist University [2]Rutgers University [3]Snap Inc.

April 22, 2026

In Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (ACL 2026), Main Conference

Agenda

- **1. Background:** The Rise of Agentic Recommender Systems
- **2. The Evolution of Memory:** From Static Profiles to Dynamic States
- **3. Motivation:** The Missing "Collaboration"
- **4. The MemRec Framework:** Architectural Decoupling
- **5. Empirical Results**
- **6. Q&A**

The Evolution of Recommender Systems

How do we store and reason about user preferences?

- **Heuristic CF Era (e.g., User/Item-KNN):** Sparse interaction matrices.
  Directly computing similarity based on co-occurrences.
- **Representation Learning Era (SVD, FM $\rightarrow$ Deep Learning):** Dense Latent Embeddings.
  Compressing preferences into numerical vectors (implicit representation).
- **Agentic Era (Emerging):** Semantic Memory powered by LLMs.
  Storing preferences as natural language text, allowing for explicit, human-like reasoning.

What is an Agentic Recommender System?

Definition

An Agentic Recommender System acts as an **autonomous virtual assistant**, utilizing LLMs to perform complex reasoning, process semantic memory, and interact via natural language.

Key Components:

- **Brain (LLM):** Instruction following & reasoning.
- **Memory Module:** Semantic storage of user/item histories.
- **Tool Use:** APIs for searching, filtering, or retrieval.

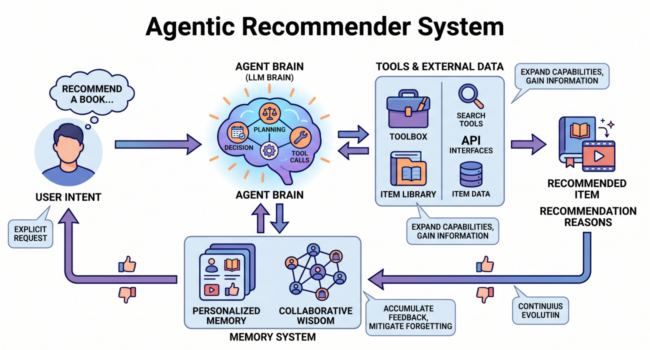


Figure 1: General Architecture of an Agentic RS.

Traditional RS vs. Agentic RS

Application Scenario Comparison:

- **Traditional RecSys (Stateless/Implicit):**
  You click a sci-fi book → The system updates your dense embedding vector → It silently pushes another sci-fi book to your homepage based on vector cosine similarity.

- **Agentic RecSys (Stateful/Explicit):**
  You tell the system: *"I want a book similar to 'Dune', but shorter, and suitable for a weekend trip."* → The Agent checks your explicit textual **Memory**, searches a database, and replies with a logically reasoned recommendation.

The Core Component: The Evolution of "Memory"

For an agent to be more than a stateless function, it must utilize persistent memory.

Three Key Milestones:

1. **No Explicit Memory:** Relying solely on the LLM's parametric knowledge or raw histories (e.g., Vanilla LLM).
2. **Static Memory:** Retrieving context from a fixed profile that does not dynamically evolve.
3. **Dynamic Memory:** Agents iteratively update their own semantic understanding over time (e.g., i^2Agent).

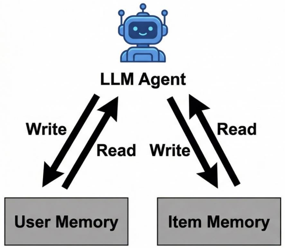


Figure 2: Current isolated memory paradigm.

The Limitation

Despite being dynamic, this paradigm remains **isolated**. It updates profiles in a silo, completely missing vital **collaborative signals** from neighbors and broader community.

The Solution: From Isolated to Collaborative Memory

Introducing the MemRec Framework

- **Bridging the Gap:** Shifts the paradigm from isolated self-reflection to a dynamic, **collaborative memory graph**.
- **Architectural Decoupling:** Separates reasoning from memory management to overcome cognitive overload.
- **Global Context:** Synthesizes high-order connectivity of community trends to augment the agent's understanding.

Analogy: The agent now consults a community library instead of just reading a single user's interaction history.

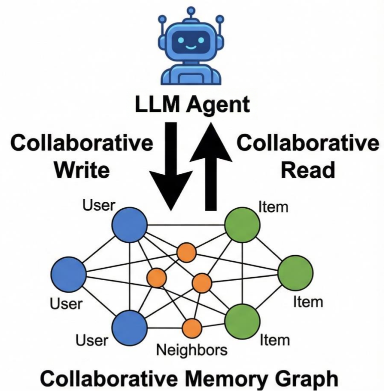


Figure 3: The proposed MemRec Framework.

Our Solution: MemRec Framework (Architectural Decoupling)

Core Idea: Architecturally decouple "Memory Management" from "High-Level Reasoning" to establish a new Pareto frontier in cost and performance.

The "Librarian" (LM_{Mem})

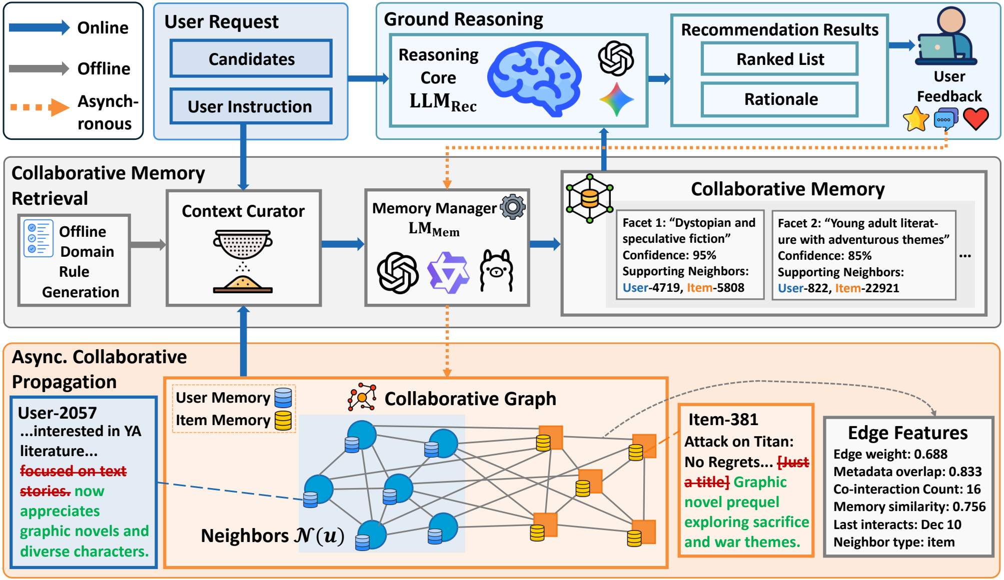
- **Role:** Memory Manager
- **Model:** Fast, cost-effective models (e.g., gpt-4o-mini or local 7B models)
- **Action:** Zero-shot curation, background graph updates, and noise filtering.

$\leftrightarrow$

The "Brain" (LLM_{Rec})

- **Role:** Reasoning Core
- **Model:** High-capacity models (e.g., gpt-4o)
- **Action:** Processes clean, synthesized collaborative context to generate grounded recommendations.

System Architecture Overview



Stage-R: Collaborative Memory Retrieval

How the Librarian prevents Cognitive Overload using a "Curate-then-Synthesize" pipeline:

Step 1: LLM-Guided Context Curation

- Rapidly filters out noisy neighbors using zero-shot domain-adaptive rules.

LLM-Generated Curation Rules (Books Snippet)

Rule 1: Content Similarity Boost
- If `metadata_overlap` > 0.6: Apply 2.5x multiplier.
- *Rationale: Books are highly content-driven...*

Rule 2: Collaborative Filtering with Threshold
- If `co_interaction` > 3: Apply 1.8x multiplier.
- *Rationale: Meaningful CF signal requires sufficient overlap.*

Step 2: Collaborative Memory Synthesis

- Distills high-signal patterns from the curated neighbors into facets.

Prompt Snippet: Collaborative Synthesis

Your Task: Analyze the target user's personal memory and the raw collaborative memories from the curated neighboring users and items. Identify distinct preference facets...

Output Requirement: For each facet, provide:
1. A concise description of the preference theme.
2. A confidence score (0-1).
3. A list of supporting neighbors providing evidence.

Stage-ReRank: Grounded Reasoning

- The "Brain" (LLM_{Rec}) executes the final ranking decision.
- **Input:** User Instruction + Synthesized Collaborative Context (M_{collab}) + Candidate Items.
- **Output:** Relevance Score $(0-1)$ + Natural Language Rationale.

The Advantage

By grounding the generation in M_{collab}, the generated rationale is factually supported by **broader community evidence**, drastically improving both specificity and relevance compared to isolated agents.

Stage-W: Asynchronous Collaborative Propagation

Solving the $O(|N|)$ computational bottleneck for continuous graph evolution:

- **Constant-time Interaction ($O(1)$):**
- When User A interacts with an item, we execute self-reflection and neighbor updates as a **single, batched asynchronous operation**.
- The LM_{Mem} simultaneously computes the update delta (ΔM) for the user, the item, and the surrounding neighborhood graph.
- **Result:** The global memory graph continuously captures shifting trends in the background without blocking the online user experience.

Experimental Setup

Datasets:

- **Books & Goodreads:** Sparse vs. Dense textual/content-driven preferences.
- **MovieTV:** Volatile, recency-sensitive user preferences.
- **Yelp:** Strong spatial and categorical constraints.

Baselines:

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

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 AgentCF	0.3457	0.4960	0.7403	0.2951	0.4654	0.7726
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%

- Significant improvements achieved across all four benchmark datasets.
- The most substantial gains are observed in top-tier precision (**Hit@1**), demonstrating that synthesizing global graph signals is vastly superior to relying solely on isolated memory.

Qualitative Case Study: A Complete Collaborative Journey

Target: User 2057 (Sparse history: Has only read a few YA Fantasy books).
Instruction: *"I seek a visually immersive experience... like a* ***graphic novel****."*

Stage-R: The Librarian checks the community

- The Librarian pulls data from 16 similar readers.
- **Synthesized Facets:** "This community is currently highly engaged with **dystopian fiction** and **emotionally layered storytelling**."

Stage-ReRank: The Brain makes the call

- **Recommendation:** *Attack on Titan: No Regrets*
- **Brain's Rationale:** "This is a **graphic novel** (matching your format request). Also, based on similar readers, it fits the community's deep interest in **dystopian settings** and **complex storytelling**."

Thank You!

Q & A

Paper available at: `https://arxiv.org/abs/2601.08816`
Code available at: `https://github.com/rutgerswiselab/memrec`
Homepage: `https://memrec.weixinchen.com`