Part IV: Towards Agentic LLMs and NeSys AI for RecSys

RecSys - Past, Present, and Future (Challenges) Course

Markus Reiter-Haas, Elisabeth Lex

July 2026

Course Website: *https://aisocietylab.github.io/recsys-ppf-course/*

Lecture Structure

- 5min Recap
- 20min NeSy Theory + Examples
- 30min Agentic Theory + Demo
- 25min Open Challenges/Discussion
- 10min Wrap Up/Final Questions

Recap: Modern and Ethical RecSys

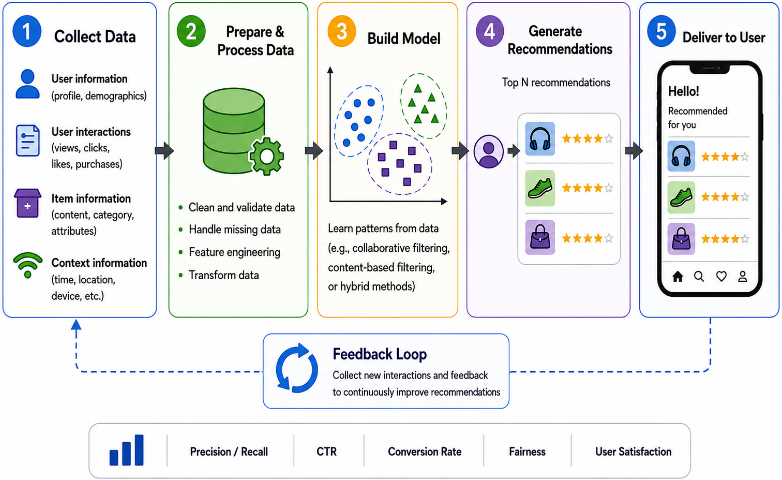
- Transition to Deep Learning
- Implementation of Two Tower Model
- Ethical and Legal Considerations
- Trustworthy and Inclusive RecSys

Recap: Modern and Ethical RecSys

- Transition to Deep Learning
- Implementation of Two Tower Model
- Ethical and Legal Considerations
- Trustworthy and Inclusive RecSys

Ice Breaker Question: How are POI Recommendations different from Route Navigation?

How Recommender Systems Work



NeSy and Agentic AI

- Neurosymbolic (NeSy) AI [6]: combines neural (nets) + symbolic (explicit representations) approaches of AI
- Agentic AI[1]: limited supervision of (proactive) AI agents to achieve a goal (cf. a reactive AI assistant)

[1] https://www.ibm.com/think/topics/agentic-ai

Learning goals

By the end of the session, you should be able to:

- Distinguish Conversational Recommender Systems from AI assistants.
- Describe three distinct types of Neurosymbolic Recommender Systems.
- Explain the different stages in Agentic Recommender Systems.
- Identify and reason about open challenges for future Recommender Systems.

Neurosymbolic RecSys

Deep Integration

What is Neurosymbolic AI?

Subsymbolic AI

- neural networks
- learned representations
- pattern recognition from data

$\leftrightarrow$

Symbolic AI

- knowledge representation
- explicit structures
- reasoning over represented knowledge

Neurosymbolic AI combines learning from data with structured knowledge and reasoning.

Why Neurosymbolic Recommender Systems?

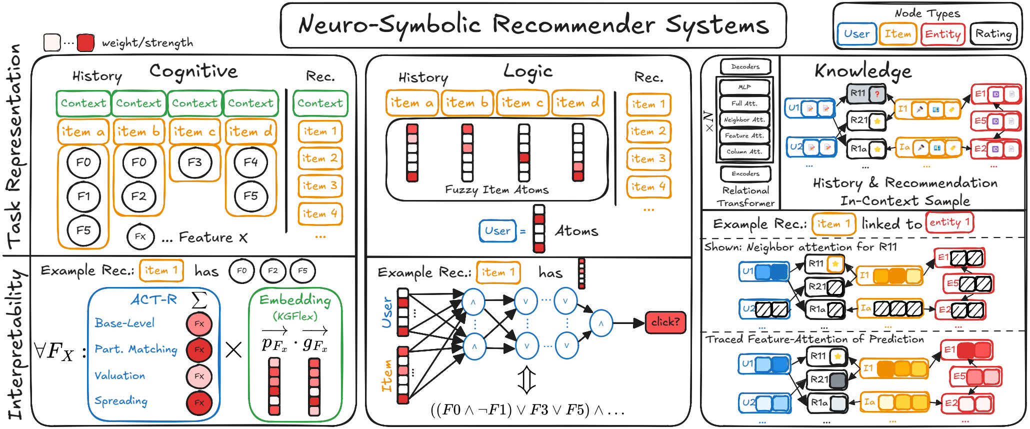
Neural recommenders

- Effective at ranking and prediction
- Learn from high-dimensional user-item signals
- Often difficult to inspect or debug

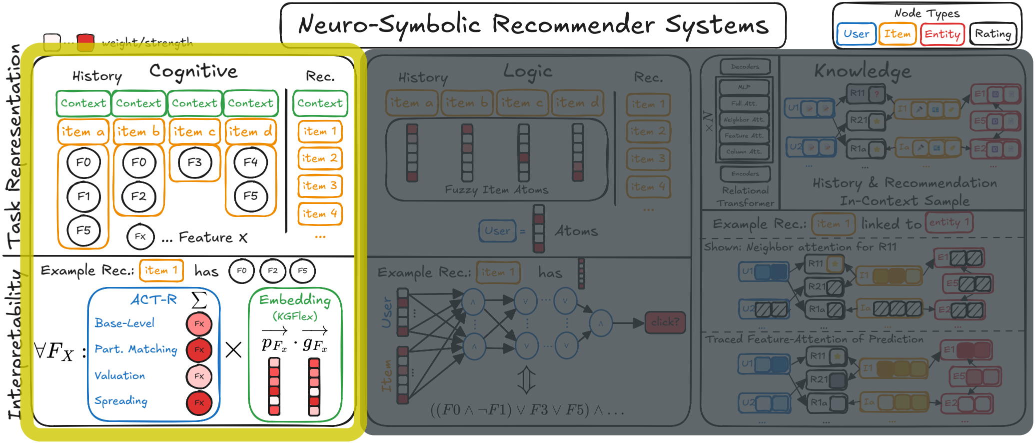
Symbolic and cognitive methods

- Represent explicit features, concepts, and rules
- Enable explanations grounded in user or item attributes
- Support bias analysis and transparent decision-making

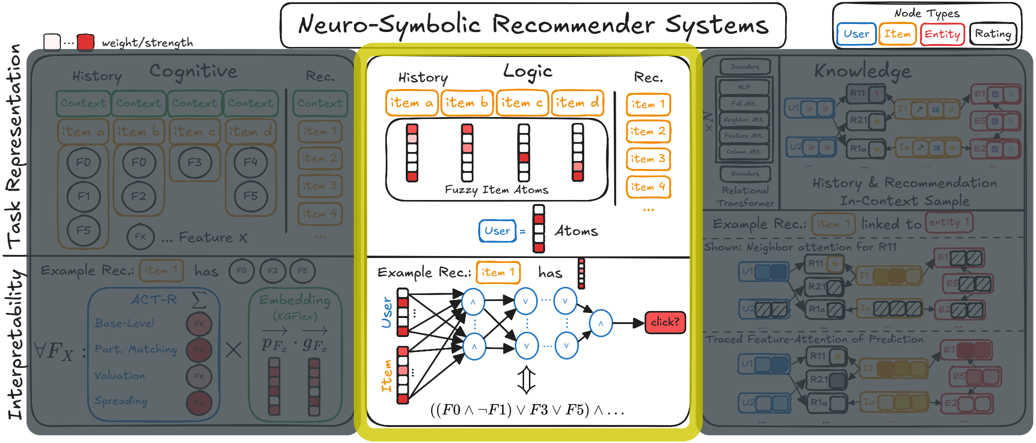
Types of Neurosymbolic RecSys



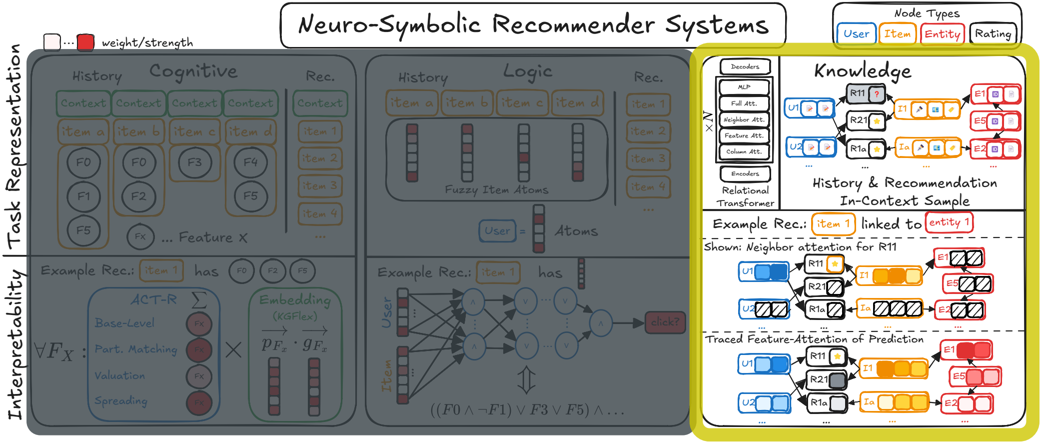
Types of Neurosymbolic RecSys



Types of Neurosymbolic RecSys



Types of Neurosymbolic RecSys



Cognitive Architectures

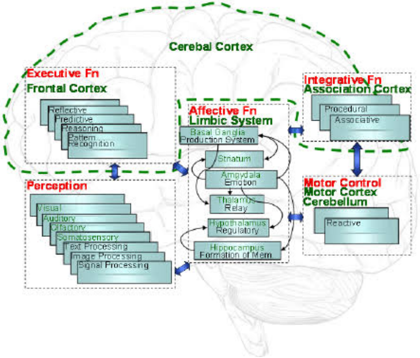


Figure 1: Cognitive Architecture mapped on Brain.

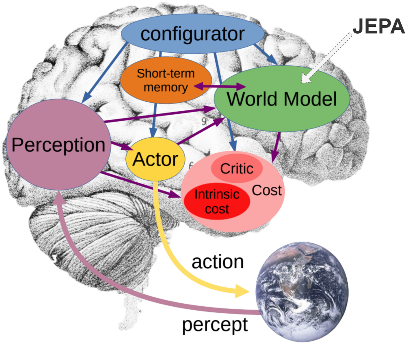


Figure 2: World models as envisioned by LeCun.

Example Cognition-based NeSy RecSys: ACT-R

Core Concept: Recommender systems that emulate human decision-making processes to enhance personalization and explainability.

Dual-Process Theory:

- **System 1 (Fast):** Intuitive, associative, and reactive (e.g., immediate item clicks).
- **System 2 (Slow):** Deliberate, logical, and effortful (e.g., long-term preference evaluation).

Example in RecSys:

- **ACT-R Integration:** Uses symbolic/sub-symbolic production rules to model memory activation and retrieval.
- **Example:** Model music relistening patterns [12].

Logic and Reasoning

Logical representations can express explicit assumptions about recommendation behavior.

They can support:

- rule-based constraints
- inspection of decision processes
- explanation generation

Example rule

If a user repeatedly interacts with a topic, then the system may increase the relevance of related items.

Examples Logic-Based NeSy RecSys: LTNRec

LTNrec[2] [2] applies *Logic Tensor Networks (LTNs)* to top-N recommendation, bridging the gap between statistical deep learning and symbolic reasoning.

Core Mechanism:

- **Grounding:** Maps symbolic entities (users, items, genres) to continuous embedding vectors.
- **Fuzzy Logic:** Uses differentiable T-norms to implement FOL operators ($\wedge, \vee, \implies$) in $[0, 1]$.
- **Knowledge Injection:** Axioms enforce domain rules (e.g., *"User likes Item $\implies$ User likes Genre"*).

Optimization Goal:

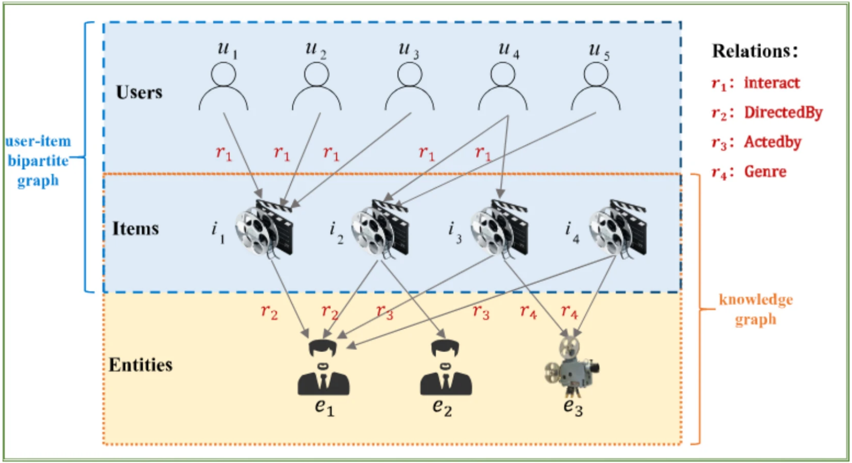
Minimize a combined loss function:

$$\mathcal{L} = \mathcal{L}_{\text{data}} + \lambda \mathcal{L}_{\text{logic}}$$

- $\mathcal{L}_{\textbf{data}}$: Predictive accuracy (e.g., MSE).
- $\mathcal{L}_{\textbf{logic}}$: Dissatisfaction of logical axioms

[2] `https://github.com/tommasocarraro/LTNrec`

Knowledge Graphs and Representations



collaborative knowledge graph

Example Knowledge-Based NeSy RecSys: KGFlex

Most approaches *embed knowledge graphs*.

KGFlex [3] optimizes recommendation by integrating *Knowledge Graphs* with *Sparse Feature Factorization*.

Knowledge Graph Definition

A heterogeneous KG is defined as $\mathcal{G} = (\mathcal{E}, \mathcal{R})$, where:

- $\mathcal{E}$ is the set of entities (Users, Items, Attributes).
- $\mathcal{R}$ is the set of relation types.
- $\mathcal{F} \subseteq \mathcal{E} \times \mathcal{R} \times \mathcal{E}$ are the observed facts (triples).

The predicted rating $\hat{r}_{ui}$ in KGFlex is defined by aggregating the user's affinity towards the set of features associated with an item:

$$\hat{r}_{ui} = \sum_{f \in \mathcal{F}_{ui}} k_{uf}(p_u^T g_f + b_f) \qquad (1)$$

where k_{uf} is modeled by the information gain

Hybrid NeSy RecSys

- Combine logic and knowledge, e.g., FOL rules with knowledge embeddings [14].
- Feature-level (extracted from knowledge graphs) retrieval of items based on cognitive-inspired equations (WIP).
- ...?

→ research gap to be explored!

Opportunity to Publish Your Research

Second International Workshop on Hybrid AI for Human-Centric Personalization and Recommendation (HyPeR 2026)
held at 35th International ACM Conference on Knowledge and Information Management (CIKM 2026)
November 7-11, 2026 | Rome, ITALY

- all kinds of NeSy integration for personalization/recommendations
- not only algorithmic: behavioral analysis, evaluation, trustworthiness, . . .

Official call we be released soon; expected submission deadline August 24TH

Agentic RecSys

All the Rave!

Excercise: LLMs as recommender

Try out the prepared gem[3] or any of LLMs of your own choosing.

What are they good at? Where is room for improvement?

[3] *https://gemini.google.com/gem/1UuOuQ4COG0vwSSUHtNTyJOc63bAqBrHn?usp=sharing*

Why Agentic AI

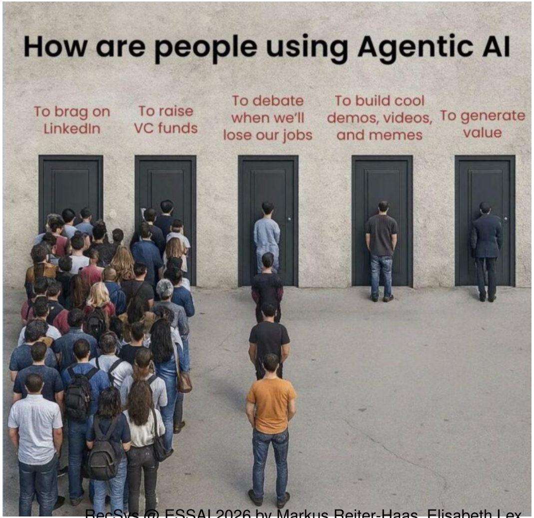
GPTs and Assistants are precursors to agents. They will gradually be able to plan and perform more complex actions on your behalf. These are our first step toward AI Agents.

By Sam Altman (2023)

Pathway to Agentic AI

- → large language models (LLMs)
- → retrieval augment generation (RAG)
- → tool calling/function chaining (via JSON schema)
- → model context protocol (MCP) services
- → (autonomous) agents
- → multi-agent systems

Motivation for Agentic AI



The Shift to Agentic Recommender Systems

Classical/Deep Learning algorithms rely on pattern matching

Properties: reactive; past behavior but no explicit working memory; single-step recommendation

Conversational/Generative models use semantic reasoning

Properties: responsive; context (in context window); multi-step answer

LLM-based/Agentic approaches are goal-driven

Properties: proactive/guidance; situational with structured long-term memory; multi-step planning

Condensed and adapted information from "Multi-Agentic Recommender Systems: Foundations, Design Patterns, and E-Commerce Applications An Industrial Tutorial" [18]

What are Agentic Recommender Systems?

*We define agentic recommender systems as recommendation pipelines in which one or more stateful agents **observe, plan, call tools, and verify**, rather than score in a single shot, while **operating over users, item catalogs, candidate sets, and recommendation objectives**.*

"The Future is Agentic: Definitions, Perspectives, and Open Challenges of Multi-Agent Recommender Systems" [10]

Components of Agentic RecSys

Core capabilities [10]:

- **Planning & Task Decomposition** (candidate selection, constraint checking, bundle retrieval, re-ranking, and verification)
- **Tool Use & Action Execution** (retrieval systems, inventory databases, policy checkers, vision models, or constraint solvers)
- **Memory & State Management** (working, episodic, semantic, or procedural memory)
- **Autonomy & Goal-Driven Behavior** (closed loop until recommendation objective is satisfied)

Analogies for Agentic RecSys

- Bring in context: context-aware recommender systems
- Distinct function responsibility: cognitive architectures.
- Split the task: microservices (rather than monolithic)

LLM-driven Recommendations

- User Intent as Natural Language (rather than implicit) $\rightarrow$ Conversational Recommender Systems
- Leverage internal reasoning capabilities.

Retrieval Augmented Generation (RAG)

- Retrieval Augmented Generation [8, 5]
  - AI framework combining LLMs with external knowledge base
  - LLM can reference specific, adequate documents and/or parts
  - Output becomes more accurate and relevant

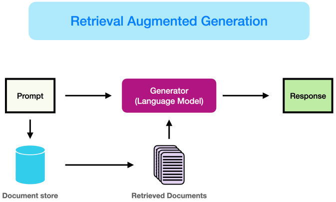


Figure 5: Naïve RAG Overview

Credit: Promptinguide.ai

RAG System

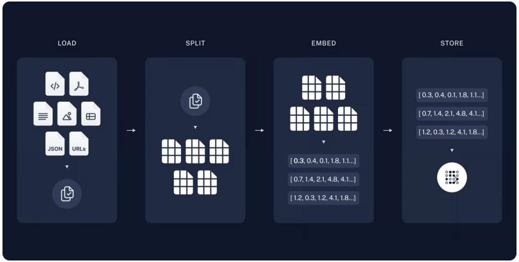


Figure 6: Indexing in Vector Store using LangChain: Load → Split → Embed → Store

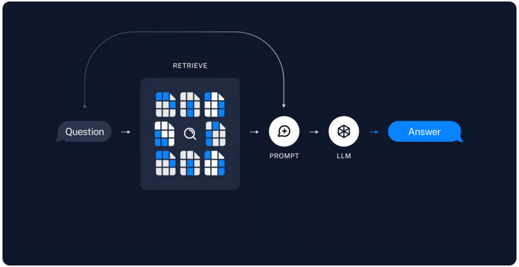


Figure 7: Retrieval process using Semenatic Search in LangChain: Question → Retrieve (+Question) → Prompt → LLM →Answer

The RAG triad

- **Problem:** Traditional IR metrics (precision/recall) are not enough to evaluate agentic frameworks

The RAG triad

- **Problem:** Traditional IR metrics (precision/recall) are not enough to evaluate agentic frameworks
- **Solution:** RAG triad, evaluating three distinct edges of the system:

The RAG triad

- **Problem:** Traditional IR metrics (precision/recall) are not enough to evaluate agentic frameworks
- **Solution:** RAG triad, evaluating three distinct edges of the system:
  1. **Context Relevance:** Is the retrieved chunk actually relevant?
  2. **Groundedness:** Is the answer derived from retrieved text, or did the model hallucinate?
  3. **Answer relevance:** Does the generated answer actually address and answer the query?

The RAG triad cont.

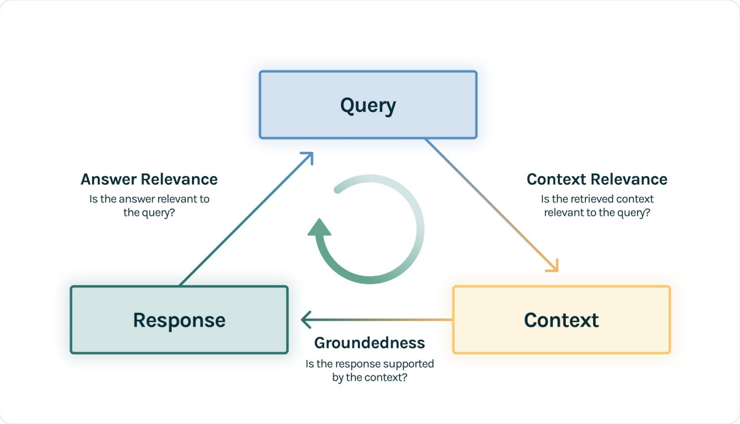


Figure 8: The RAG Triad (Truera)

Tool/Function calling/chaining for Recommender System

Demo of system[4]

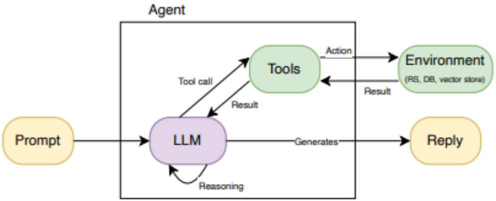


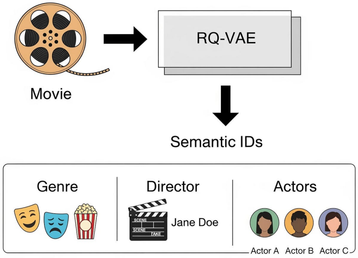
Figure 9: Tool Calling workflow from [1]

Challenge: structured model outputs (JSON schema)

Limitations, e.g., on MovieLens?

[4] https://github.com/tommasocarraro-sony/recsys-agent

Semantic IDs



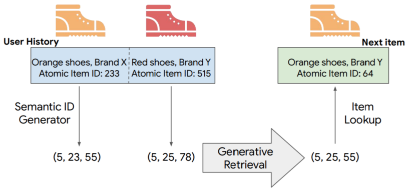


Figure 11: How semantic IDs can be used [11]. Generative Retrieval is similar to language modeling.

Figure 10: Intuition behind Semantic IDs.
Residual Quantized Variational Autoencoders convert the items (IDs) to meaningful token-based semantic IDs.

Why Semantic IDs?

- Semantic relation between IDs
- Leverage prior knowledge of items/users
- Smaller number of parameters (i.e., vocabulary) required

From Single-Step Prediction to Multi-Step Reasoning (CoT)[17]

- **The shift:** Moving from "input $\rightarrow$ answer" to "input $\rightarrow$ reasoning steps $\rightarrow$ answer"
- **Decomposition:** The ability of an LLM to break down complex contexts into intermediate steps

From Single-Step Prediction to Multi-Step Reasoning (CoT)[17] cont.

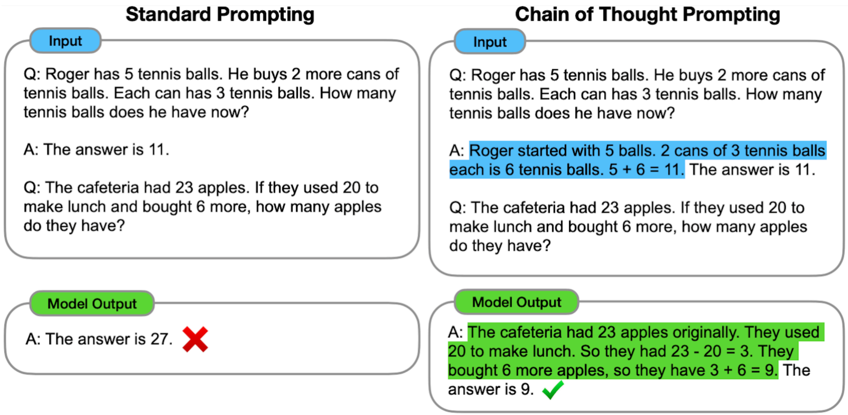


Figure 12: Chain of Thought Overview (Google Research)

ReAct Pattern/Loop

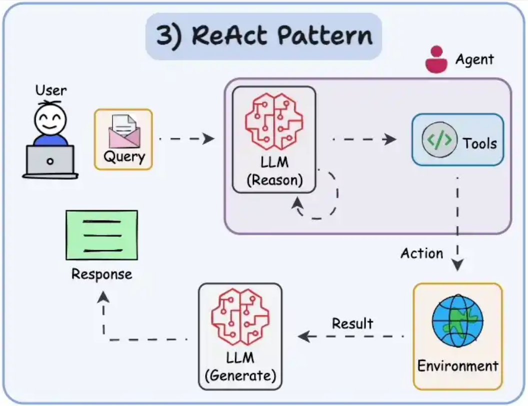


Figure 13: ReAct Pattern

Model Context Protocol (MCP)

Proposed by Anthropic to solve NxM integration problem.

- JSON-RPC 2.0
- Session-based
- Many libraries available (e.g., FastMCP)
- Adopted by major providers (e.g., OpenAI, Google)
- Donated to Agentic AI Foundation (AAIF)

Demo MCP Server for Recommender System

Weather-based Movie Recommendations:

1. request
2. reason → ask for information
3. reason → call (geo + weather) APIs
4. reason → call movie recommender
5. reason → response

Agents

- **Agent:** An autonomous system that uses an LLM as a "cognitive engine" to perceive, reason, and act towards achieving a goal
- Agents usually automate simple tasks

Agents

- **Agent:** An autonomous system that uses an LLM as a "cognitive engine" to perceive, reason, and act towards achieving a goal
- Agents usually automate simple tasks
- **Key components:**
  - **Brain:** The LLM which does the "planning"
  - **Hands:** Tools and APIs as connections
  - **Memory:** Short-term (context window) and long term (vector DB)

The "Router" Architecture[9]

- In an advanced system, data is scattered through different sources (e.g. SQL, Vector DB, API, Web, external tools, etc.)
- The agent functions as a **semantic router**
- **Mechanism:**
  1. The LLM analyses the *intent* of the query
  2. The LLM "routes" the retrieval task to the appropriate tool

User Query	Selected Tool	Reasoning
"Calculate 25 * 48"	Calculator	Math intent
"TUG Policy on sick leave"	Vector DB (RAG)	Internal knowledge
"Stock price of Nvidia"	Live Web API	Real-time data needed

Table 1: Routing examples based on query intent

Agentic Recommender Systems: InteRecAgent [7]

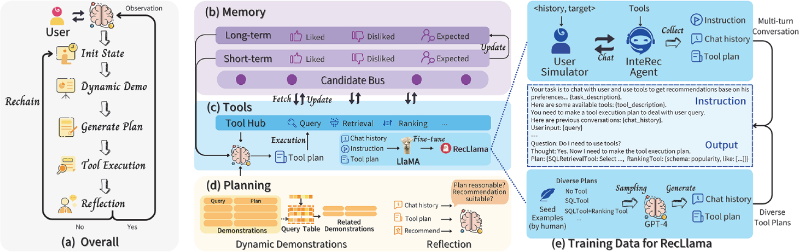


Fig. 1. InteRecAgent framework. (a) The overall pipeline of InteRecAgent; (b) the memory module, consisting of a candidate memory bus, a long-term and a short-term user profile; (c) tool module, consisting of various tools, the plan-first execution strategy and the fine-tuning of RecLlama; (d) planning module, involving the dynamic demonstrations and the reflection strategy; and (e) sources of fine-tuning data for RecLlama.

Figure 14: Interactive Recommender Agent

Agentic Recommender Systems: MACRec [16]

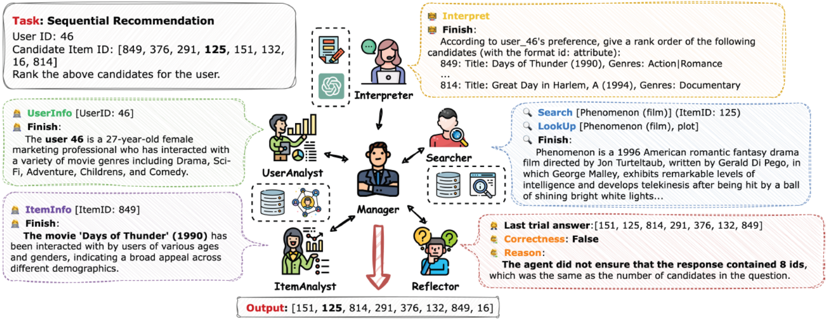


Figure 1: The Framework of MACRec. We take a sequential recommendation task as an example to show how these agents work collaboratively.

Figure 15: Multi-Agent Collaboration Framework for Recommendation

Directions for Open Challenges

How to progress?

Trinity of RecSys challenges

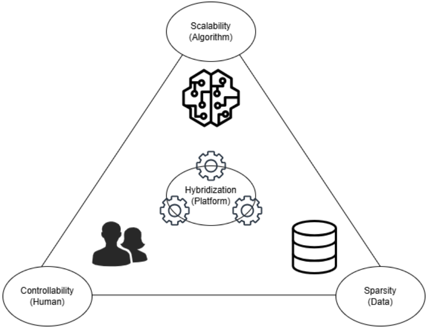


Figure 16: Most challenges fall into one of three directions: 1. (algorithmic) scalability, 2. (data) sparsity), or 3. (human) controllability. Hybridization (for actual platforms) often try to balance the trade-offs.

Discuss Data Challenges: Cold-Start (Sparsity)

How do these help? Where do they fall short?

- Knowledge Graphs
- Semantic IDs
- Conversational Systems

How do they relate to the problem of **Popularity Bias**?

Algorithmic Challenges: Cross-Domain (Scalability)

How do these help? Where do they fall short?

- Vector Databases (RAG)
- Neurosymbolic Approaches
- Agentic Systems

How do they relate to the problem of a **Lack of User Control**?

Human-Centered Challenges: Adaptability (Controllability)

How do these help? Where do they fall short?

- Cognitive Architecture
- Logical Rules
- Router Architecture

How do they relate to the problem of **Opaqueness in the Recommendation Process**?

The Role of Hybridization

Discussion points:

- How to accurately represent all kinds of users (and their intents)?
- Trade-offs: implicit, data-driven (machine) learning vs. explicit knowledge representations and logic-based reasoning (i.e., symbolic)
- "We need compositionality - predictions that can be built out of other predictions" by Rich Sutton (2015)[5]

[5] *http://www.incompleteideas.net/Talks/ncap2015.pdf*

Going back to RecSys 2007[6]

- Algorithms x2 (Collaborative Filtering + Learning)
- Privacy & Trust, User Issues in Recommender Systems

Summarized by Gemini:

> *In short, the 2007 conference was less about "what to watch next" and more about "how can we build a system that is fair, private, and smart enough to talk to the user?"*

Example paper: Conversational recommenders with adaptive suggestions [15]

Are we making much progress? [4]

[6] `https://recsys.acm.org/recsys07/program/`

Closing and Reflection

- RecSys as an application for AI
- Evolution closely mimics AI trends
- Diverse considerations in practice
- Many challenges still unsolved

Closing and Reflection

- RecSys as an application for AI
- Evolution closely mimics AI trends
- Diverse considerations in practice
- Many challenges still unsolved

Final Questions?

References for Future i

[1] Tommaso Carraro, Brijraj Singh, and Niranjan Pedanekar. **“Large Language Model-based Recommendation System Agents”.** In: *Proceedings of the Nineteenth ACM Conference on Recommender Systems*. 2025, pp. 1340–1342.

[2] Tommaso Carraro et al. **“Logic tensor networks for top-n recommendation”.** In: *International Conference of the Italian Association for Artificial Intelligence*. Springer. 2022, pp. 110–123.

[3] Antonio Ferrara et al. **“Kgflex: Efficient recommendation with sparse feature factorization and knowledge graphs”.** In: *ACM Transactions on Recommender Systems* 1.4 (2023), pp. 1–30.

References for Future ii

[4] Maurizio Ferrari Dacrema, Paolo Cremonesi, and Dietmar Jannach. **"Are we really making much progress? A worrying analysis of recent neural recommendation approaches".** In: *Proceedings of the 13th ACM conference on recommender systems*. 2019, pp. 101–109.

[5] Yunfan Gao et al. **"Retrieval-augmented generation for large language models: A survey".** In: *arXiv preprint arXiv:2312.10997* (2023).

[6] Pascal Hitzler et al. **"Neuro-symbolic approaches in artificial intelligence".** In: *National Science Review* 9.6 (2022), nwac035.

[7] Xu Huang et al. **"Recommender ai agent: Integrating large language models for interactive recommendations".** In: *ACM Transactions on Information Systems* 43.4 (2025), pp. 1–33.

References for Future iii

[8] Patrick Lewis et al. **"Retrieval-augmented generation for knowledge-intensive nlp tasks".** In: *Advances in neural information processing systems* 33 (2020), pp. 9459–9474.

[9] Jerry Liu. ***LlamaIndex: Data Framework for LLM Applications*.** Accessed: 2025-12-01. 2023. URL: `https://www.llamaindex.ai/`.

[10] Reza Yousefi Maragheh and Yashar Deldjoo. **"The future is agentic: Definitions, perspectives, and open challenges of multi-agent recommender systems".** In: *arXiv preprint arXiv:2507.02097* (2025).

[11] Shashank Rajput et al. **"Recommender systems with generative retrieval".** In: *Advances in Neural Information Processing Systems* 36 (2023), pp. 10299–10315.

References for Future iv

[12] Markus Reiter-Haas et al. “**Predicting music relistening behavior using the ACT-R framework**”. In: *Proceedings of the 15th ACM conference on recommender systems*. 2021, pp. 702–707.

[13] Zhang Rong, Liu Yuan, and Li Yang. “**Enhanced knowledge graph recommendation algorithm based on multi-level contrastive learning**”. In: *Scientific Reports* 14.1 (2024), p. 23051.

[14] Giuseppe Spillo et al. “**Recommender systems based on neuro-symbolic knowledge graph embeddings encoding first-order logic rules**”. In: *User Modeling and User-Adapted Interaction* 34.5 (2024), pp. 2039–2083.

References for Future v

[15] Paolo Viappiani, Pearl Pu, and Boi Faltings. **"Conversational recommenders with adaptive suggestions".** In: *Proceedings of the 2007 ACM conference on Recommender systems*. 2007, pp. 89–96.

[16] Zhefan Wang et al. **"Macrec: A multi-agent collaboration framework for recommendation".** In: *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2024, pp. 2760–2764.

[17] Jason Wei et al. **"Chain-of-thought prompting elicits reasoning in large language models".** In: *Advances in neural information processing systems* 35 (2022), pp. 24824–24837.

References for Future vi

[18] Reza Yousefi Maragheh et al. "**Multi-Agentic Recommender Systems: Foundations, Design Patterns, and E-Commerce ApplicationsAn Industrial Tutorial**". In: *Proceedings of the Nineteenth ACM Conference on Recommender Systems*. 2025, pp. 1427–1429.