Multi-Agentic Recommender Systems: Foundations, Design Patterns, and E-Commerce Applications — An Industrial Tutorial

Reza Yousefi Maragheh
Walmart Global Tech
Sunnyvale, California, USA
reza.yousefimaragheh@walmart.com

Yashar Deldjoo
Polytechnic University of Bari
Bari, Italy
yashar.deldjoo@poliba.it

Chi Wang
Google DeepMind
Redmond, Washington, USA
wangchi@google.com

Jason Cho
Walmart Global Tech
Sunnyvale, CA, USA
jason.cho@walmart.com

Derek Cheng
Google DeepMind
Mountain View, CA, USA
zcheng@google.com

Abstract

The goal of this tutorial is to provide our perspective on the most recent advances in LLM-powered agents for recommender systems. Building on our extensive experience deploying agentic tools in large-scale environments, this tutorial hopes to deepen the understanding of participants with diverse backgrounds on the alphabets that underpin multi-agentic frameworks. Organized by the founders of leading agentic tools, the tutorial will highlight how these frameworks are being applied to create next-generation recommender systems in diverse applications. The examples include context-aware recommendation, dynamic multi-step orchestration, and personalized recommendation systems. To provide a solid foundation, we begin with a brief background on the evolution of recommender systems and how recent breakthroughs in large language models (LLMs) have shifted the paradigm toward more interactive, adaptive, and autonomous systems. The hands-on session will allow participants to directly engage with state-of-the-art techniques, bridging the gap between theoretical concepts and practical implementations.

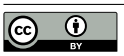
CCS Concepts

• **Information systems** → *Recommender systems.*

Keywords

Agentic Recommender Systems, Large Language Models, Multi-Agent AI, Industrial Applications, Autonomous Recommendation, Interactive Recommendation, LLM-Powered Agents

ACM Reference Format:
Reza Yousefi Maragheh, Yashar Deldjoo, Chi Wang, Jason Cho, and Derek Cheng. 2025. Multi-Agentic Recommender Systems: Foundations, Design Patterns, and E-Commerce Applications — An Industrial Tutorial. In *Proceedings of the Nineteenth ACM Conference on Recommender Systems (RecSys '25), September 22–26, 2025, Prague, Czech Republic.* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3705328.3748008



RecSys '25, Prague, Czech Republic

ACM ISBN 979-8-4007-1364-4/25/09
https://doi.org/10.1145/3705328.3748008

1 Tutorial Length, Level, and Audience

We propose a session of approximately **3 hours** (half-day format), including a **hands-on session**.

2 Motivation for Proposing this Tutorial

Despite major breakthroughs in recommender systems, most large-scale and user-facing industrial solutions still follow static or "one-shot" models that produce recommendations with limited interaction or personalization beyond user profiles. As user demands evolve toward more intuitive, interactive, and context-aware experiences, there is an urgent need for these systems to evolve into dynamic autonomous systems capable of reasoning, *multi-step orchestration*, and proactive engagement. Recent developments in LLM enable new capabilities to reason dialogically, maintain longer contexts (memory), query external tools autonomously, and orchestrate *multi-step* workflows [11, 24], paving the way for 'autonomous' or 'agentic' recommender systems.

In short, these *agentic* solutions, or more precisely **multi-agent systems**, can adapt to diverse user behaviors, moving beyond the simple question–response paradigm typical of non-agentic chatbots. They can proactively reason [13, 19, 21, 28], query external APIs [18, 23], refine suggestions based on user and system constraints [3, 9, 25], and even simulate user behavior to evaluate recommendations on broader criteria [26, 29]. However, implementing such advanced pipelines also raises concerns around scalability, reliability, and transparency. Balancing interactive, multi-step reasoning with performance and safety constraints requires best design frameworks and best practices.

Against this backdrop, this tutorial clarifies the core concepts, architectures, and workflows of agentic LLM-based recommender systems. By illustrating the building blocks and showcasing real-world implementation experiences, and providing formal definitions, we aim to equip participants with actionable insights into how these powerful paradigms fundamentally impact the way we think about the design of next wave of (generative) recommender systems.[1]

[1]It is worth mentioning that the authors are currently preparing a visionary survey paper that expands on the topics discussed in this tutorial. This complementary work will be shared alongside our tutorial materials, including dedicated Colab files and slides, which will be made available on a dedicated GitHub page.

3 Instructors

- Dr. Reza Yousefi Maragheh (ryousefimaragheh@acm.org), Staff Data Scientist, Walmart Global Tech, Sunnyvale, USA,
- Dr. Yashar Deldjoo (deldjooy@acm.org), Senior Researcher and Associate Professor, Polytecyhnic University of Bari, Italy
- Dr. Chi Wang (wangchi@google.com), Senior Staff Research Scientist, Google DeepMind,
- Dr. Jason Cho (jhdcho@gmail.com), Director of Data Science and ML, Walmart Global Tech, Sunnyvale, USA,
- Dr. Derek Cheng, (zcheng@google.com), Senior Staff Research Scientist, Google DeepMind.

Collectively, the instructors have presented tutorials at international venues (RecSys, KDD, WWW), authored refereed articles on LLM-based recommender systems, and developed industrial-scale agentic AI pipelines [1, 2, 2, 4–8, 10, 12, 14–17, 20, 22, 27]. *At least one instructor from each institution (Walmart Global Tech, Google DeepMind, and Polytechnic University of Bari) will attend and present in person.*

4 Outline

We structure this 3-hour tutorial into five modules:

A. **Introduction and Background (15 minutes)**
- Core properties of modern RecSys: accuracy, policy alignment, context awareness, scalability (cost efficiency, development time efficiency)
- Temporal view of how evolving LLM-oriented technique address or fail to address the above goals (LLM Vanilla prompting, Chain of Thought, Self-Refine, prompt chaining, single agent frameworks, multi-agent frameworks)
- Motivating scenario (e.g., **a personalized birthday-planner**) illustrating multi-step, autonomous workflows.

B. **Alphabets of Multi-Agentic-AI (45 minutes)**
- **Memory Moderation & Retrieval Mechanisms:** Discussion of different memory types and how/when to retrieve them
- **Function Calling & Tool Usage:** Extending LLM pipelines with external APIs, databases, and knowledge bases
- **Model Context Protocols:** The standardization of the orchestrations.
- **Reasoning Load Balancing:** Splitting complex tasks into manageable segments for efficient model usage, mainstream industrial orchestrations.
- **Revisiting the Running Example**: System Design and how it can be improved using above components.

C. **Industrial Agentic RecSys Implementations (60 minutes)**
- Prominent tasks and design patterns:
  - (i) Conversational Recommendation
  - (ii) Context-Aware Autonomous Recommendation
  - (iii) Recommendation Evaluation and User Behavior Simulation
  - (iv) Explanation Generation
- Best practices for large-scale agentic pipelines: standards, pitfalls, and optimization

D. **Hands-On Demonstration (30–45 minutes)**
- Overview of open-source or industrial frameworks: AG2,[2] LangChain/LangGraph,[3] CrewAI,[4] Agent Development Kit (ADK)[5]
- Live demonstration of a multi-step recommendation scenario (e.g., "Personalized Birthday Planner") in a shared notebook environment
  - System Components
  - System and Architecture Design
  - Initial Implementation
  - Debugging and failure point diagnoses
  - Refinement

E. **Challenges & Future of Agentic AI for RecSys (15 minutes)**
- Common pitfalls: communication complexity and protocol, scalability, hallucinations, error propagation, fairness, and bias
- Privacy/fairness/unintended behavior concerns and possible mitigation strategies
- Future directions: multi-agent synergy, advanced memory systems, self-improving agentic systems.

5 Relevance and Target Audience

Relevance. Agentic AI represents an emerging, rapidly evolving field that has attracted substantial research attention due to its potential to shape the future of recommender systems. Given this growing significance, our tutorial aims to equip the community with practical knowledge and proven techniques drawn from both industry and academia. The instructors collectively bring extensive experience in agentic architectures, including contributions to foundational open-source frameworks and deployments in industrial-scale applications. Our goal is to ensure that attendees gain insights from leading research in the field.

Target Audience and Prerequisites: The tutorial is positioned between intermediate and advanced levels, with content prepared at an **intermediate level** to ensure accessibility. We believe the tutorial will benefit a wide audience, including PhD researchers exploring this emerging topic, senior researchers with experience in generative LLM models, and industry practitioners seeking practical, scalable solutions.

References

[1] Vito Walter Anelli, Luca Belli, Yashar Deldjoo, Tommaso Di Noia, Antonio Ferrara, Fedelucio Narducci, and Claudio Pomo. 2021. Pursuing privacy in recommender systems: the view of users and researchers from regulations to applications. In *Proceedings of the 15th ACM Conference on Recommender Systems*. 838–841.

[2] Ashmi Banerjee, Adithi Satish, Fitri Nur Aisyah, Wolfgang Wörndl, and Yashar Deldjoo. 2025. SynthTRIPs: A Knowledge-Grounded Framework for Benchmark Data Generation for Personalized Tourism Recommenders. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2025, Padua, Italy, July 13-18, 2025*. ACM, 3743–3752. doi:10.1145/3726302.3730321

[3] Jiao Chen, Kehui Yao, Reza Yousefi Maragheh, Kai Zhao, Jianpeng Xu, Jason Cho, Evren Korpeoglu, Sushant Kumar, and Kannan Achan. 2025. CARTS: Collaborative Agents for Recommendation Textual Summarization. *arXiv preprint arXiv:2506.17765* (2025).

[2] https://ag2.ai/
[3] https://www.langchain.com/langgraph
[4] https://www.crewai.com/
[5] https://google.github.io/adk-docs/

[4] Jason Cho and Reza Yousefi Maragheh. 2025. Productionizing OSS agents: Best practices for agent frameworks and Vertex AI. In *Google Next Conference*. Google, Las Vegas, NV, USA.

[5] Yashar Deldjoo, Zhankui He, Julian McAuley, Anton Korikov, Scott Sanner, Arnau Ramisa, René Vidal, Maheswaran Sathiamoorthy, Atoosa Kasirzadeh, and Silvia Milano. 2024. A Review of Modern Recommender Systems using Generative Models (Gen-RecSys). In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 6448–6458.

[6] Yashar Deldjoo, Zhankui He, Julian McAuley, Anton Korikov, Scott Sanner, Arnau Ramisa, René Vidal, Maheswaran Sathiamoorthy, Atoosa Kasirzadeh, and Silvia Milano. 2025. Tutorial on Recommendation with Generative Models (Gen-RecSys). In *Proceedings of the Eighteenth ACM International Conference on Web Search and Data Mining*. 1002–1004.

[7] Yashar Deldjoo, Nikhil Mehta, Maheswaran Sathiamoorthy, Shuai Zhang, Pablo Castells, and Julian J. McAuley. 2025. Toward Holistic Evaluation of Recommender Systems Powered by Generative Models. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2025, Padua, Italy, July 13-18, 2025*. ACM, 3932–3942. doi:10.1145/3726302.3730354

[8] Victor Dibia, Jingya Chen, Gagan Bansal, Suff Syed, Adam Fourney, Erkang Zhu, Chi Wang, and Saleema Amershi. 2024. Autogen studio: A no-code developer tool for building and debugging multi-agent systems. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. 72–79.

[9] Jiabao Fang, Shen Gao, Pengjie Ren, Xiuying Chen, Suzan Verberne, and Zhaochun Ren. 2024. A multi-agent conversational recommender system. *arXiv preprint arXiv:2402.01135* (2024).

[10] Najmeh Forouzandehmehr, Reza Yousefi Maragheh, Sriram Kollipara, Kai Zhao, Topojoy Biswas, Evren Korpeoglu, and Kannan Achan. 2025. CAL-RAG: Retrieval-Augmented Multi-Agent Generation for Content-Aware Layout Design. *arXiv preprint arXiv:2506.21934* (2025).

[11] Taicheng Guo, Xiuying Chen, Yaqi Wang, Ruidi Chang, Shichao Pei, Nitesh V Chawla, Olaf Wiest, and Xiangliang Zhang. 2024. Large language model based multi-agents: A survey of progress and challenges. *arXiv preprint arXiv:2402.01680* (2024).

[12] Wang-Cheng Kang, Jianmo Ni, Nikhil Mehta, Maheswaran Sathiamoorthy, Lichan Hong, Ed Chi, and Derek Zhiyuan Cheng. 2023. Do llms understand user preferences? evaluating llms on user rating prediction. *arXiv preprint arXiv:2305.06474* (2023).

[13] Reza Yousefi Maragheh and Yashar Deldjoo. 2025. The Future is Agentic: Definitions, Perspectives, and Open Challenges of Multi-Agent Recommender Systems. *arXiv preprint arXiv:2507.02097* (2025).

[14] Reza Yousefi Maragheh, Chenhao Fang, Charan Chand Irugu, Parth Parikh, Jason Cho, Jianpeng Xu, Saranyan Sukumar, Malay Patel, Evren Korpeoglu, Sushant Kumar, et al. 2023. LLM-TAKE: Theme-aware keyword extraction using large language models. In *2023 IEEE International Conference on Big Data (BigData)*. IEEE, 4318–4324.

[15] Reza Yousefi Maragheh, Lalitesh Morishetti, Ramin Giahi, Kaushiki Nag, Jianpeng Xu, Jason Cho, Evren Korpeoglu, Sushant Kumar, and Kannan Achan. 2023. Llm-based aspect augmentations for recommendation systems. (2023).

[16] Fatemeh Nazary, Yashar Deldjoo, and Tommaso Di Noia. 2025. Poison-RAG: Adversarial Data Poisoning Attacks on Retrieval-Augmented Generation in Recommender Systems. In *Advances in Information Retrieval - 47th European Conference on Information Retrieval, ECIR 2025, Lucca, Italy, April 6-10, 2025, Proceedings, Part IV (Lecture Notes in Computer Science, Vol. 15575)*. Springer, 239–251. doi:10.1007/978-3-031-88717-8_18

[17] Noveen Sachdeva, Benjamin Coleman, Wang-Cheng Kang, Jianmo Ni, Lichan Hong, Ed H Chi, James Caverlee, Julian McAuley, and Derek Zhiyuan Cheng. 2024. How to train data-efficient llms. *arXiv preprint arXiv:2402.09668* (2024).

[18] Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. 2023. Toolformer: Language models can teach themselves to use tools. *Advances in Neural Information Processing Systems* 36 (2023), 68539–68551.

[19] Yubo Shu, Haonan Zhang, Hansu Gu, Peng Zhang, Tun Lu, Dongsheng Li, and Ning Gu. 2024. RAH! RecSys–Assistant–Human: A Human-Centered Recommendation Framework With LLM Agents. *IEEE Transactions on Computational Social Systems* (2024).

[20] Ali Tourani, Fatemeh Nazary, and Yashar Deldjoo. 2025. RAG-VisualRec: An Open Resource for Vision- and Text-Enhanced Retrieval-Augmented Generation in Recommendation. *arXiv preprint arXiv:2506.20817* (2025).

[21] Zhefan Wang, Yuanqing Yu, Wendi Zheng, Weizhi Ma, and Min Zhang. 2024. Macrec: A multi-agent collaboration framework for recommendation. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2760–2764.

[22] Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, et al. 2023. Autogen: Enabling next-gen llm applications via multi-agent conversation. *arXiv preprint arXiv:2308.08155* (2023).

[23] Siyu Yuan, Kaitao Song, Jiangjie Chen, Xu Tan, Yongliang Shen, Ren Kan, Dongsheng Li, and Deqing Yang. 2024. Easytool: Enhancing llm-based agents with concise tool instruction. *arXiv preprint arXiv:2401.06201* (2024).

[24] Ceyao Zhang, Kaijie Yang, Siyi Hu, Zihao Wang, Guanghe Li, Yihang Sun, Cheng Zhang, Zhaowei Zhang, Anji Liu, Song-Chun Zhu, et al. 2024. Proagent: building proactive cooperative agents with large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 17591–17599.

[25] Jizhi Zhang, Keqin Bao, Wenjie Wang, Yang Zhang, Wentao Shi, Wanhong Xu, Fuli Feng, and Tat-Seng Chua. 2024. Prospect personalized recommendation on large language model-based agent platform. *arXiv preprint arXiv:2402.18240* (2024).

[26] Junjie Zhang, Yupeng Hou, Ruobing Xie, Wenqi Sun, Julian McAuley, Wayne Xin Zhao, Leyu Lin, and Ji-Rong Wen. 2024. Agentcf: Collaborative learning with autonomous language agents for recommender systems. In *Proceedings of the ACM Web Conference 2024*. 3679–3689.

[27] Lemei Zhang, Peng Liu, Yashar Deldjoo, Yong Zheng, and Jon Atle Gulla. 2024. Understanding Language Modeling Paradigm Adaptations in Recommender Systems: Lessons Learned and Open Challenges. In *The 27th European Conference on Artificial Intelligence (ECAI'24)*.

[28] Yuyue Zhao, Jiancan Wu, Xiang Wang, Wei Tang, Dingxian Wang, and Maarten De Rijke. 2024. Let me do it for you: Towards llm empowered recommendation via tool learning. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1796–1806.

[29] Lixi Zhu, Xiaowen Huang, and Jitao Sang. 2024. A LLM-based Controllable, Scalable, Human-Involved User Simulator Framework for Conversational Recommender Systems. *arXiv preprint arXiv:2405.08035* (2024).