Exploring the Impact of Data Quality on Agentic Recommender Systems

Marco Valentini[1,*], Antonio Ferrara[1] and Tommaso Di Noia[1]

[1]Politecnico di Bari, Bari, Italy

Abstract

Agentic Artificial Intelligence is transforming recommender systems through agent-based architectures that integrate diverse data modalities and advanced reasoning. While recent work emphasizes agent design, the quality of input data remains largely overlooked. This paper aims to provide the basic concepts to assess how data quality impacts recommendation process and user satisfaction from different perspectives in agentic settings.

Keywords

Recommender Systems, Agentic Artificial Intelligence, Multimodal Data Quality

1. Introduction

Every day, users face vast and ever-growing catalogs of items, struggling with *information overload*, which degrades their experience and threatens platform revenues. Recommender Systems (RSs) [1] have emerged as the main approach to mitigate this issue, matching user preferences with relevant content to ease decision making and boost engagement [2].

Recommendation algorithms have undergone many innovations and advancements, with the latest being *Agentic AI* architectures, based on Large Language Models (LLMs)-empowered agents, which have gained traction in recommender systems research [3]. These systems decompose the recommendation task into smaller, more manageable subtasks [4, 5], each handled by specialized agents that collaborate through various communication paradigms to achieve a shared objective. The agents composing such systems can be specialized for different roles using In-Context Learning (ICL)[6], avoiding retraining through prompt-based role assignment [7, 8]. Prompt design determines the agent's behavior, domain expertise, and interaction capabilities, including tool usage via frameworks like Reason and Act (ReAct) [9]. These agentic systems can seamlessly operate over multimodal data, e.g., text, images, user-item interactions, and dynamically retrieve relevant information during inference. This ability to integrate and reason over heterogeneous sources allows them to be interpreted as **multimodal recommender systems**, forming the foundation for the evaluation principles we propose.

While the performance of traditional recommender systems is known to depend heavily on input data quality [10], this relationship remains largely underexplored in the context of LLM-based agentic recommenders. In conventional pipelines, issues such as outdated metadata, noisy user logs, or biased knowledge sources are known to degrade model performance and user satisfaction [11]. In multi-agent settings, the concept of data quality, and its influence on the performance, lacks both a clear definition and systematic study. In these systems, each agent may rely on different data modalities, and poor-quality input may either be mitigated through collaborative reasoning or propagate through the agent pipeline, compounding its effects. Despite growing attention to agent roles and architectural coordination [12, 13, 3], the evaluation of multimodal data quality as a key factor in these systems remains a crucial, yet insufficiently addressed, research direction.

To bridge this gap, we propose evaluation criteria that aim to decouple the influence of input data quality from algorithmic performance. These principles aim to support more transparent evaluation

DaQuaMRec:1st International Workshop on Data Quality-Aware Multimodal Recommendation, September 22–26, 2025, Prague, CZ

[*]Corresponding author.

m.valentini7@phd.poliba.it (M. Valentini); antonio.ferrara@poliba.it (A. Ferrara); tommaso.dinoia@poliba.it (T. Di Noia)

0009-0004-1652-0745 (M. Valentini); 0000-0002-1921-8304 (A. Ferrara); 0000-0002-0939-5462 (T. Di Noia)

practices, foster fairer comparisons across systems, and encourage thoughtful data and prompt design in future research on agentic recommender systems.

2. Discussion

The rise of LLM-powered agentic systems in recommendation tasks [4] offers new opportunities to rethink how we evaluate not only their architectures but, more crucially, the data they rely on. These systems typically consist of specialized agents coordinated by a central orchestrator [3], each operating over different knowledge sources, e.g., user-item interactions, item metadata, textual reviews, or visual content, and collaborating to produce recommendations through cooperative reasoning.

While recent works such as MACRec [12], AgentRecBench [14], AgentCF [13], and RecMind [15] have advanced our understanding of agent collaboration and overall system performance, they fall short in addressing a critical gap: the lack of a clear definition of data quality and systematic methods to assess its impact on recommendation outcomes. This includes quality variations in retrieved content, prompt design, and knowledge base reliability, all of which are core elements of multimodal, agentic systems. We argue for a shift from model-centric benchmarking to a **data-centric perspective** [16, 17], where the quality of both input data and prompts is a pivotal factor in evaluating system behavior.

We focus on two key and underexplored dimensions of this issue:

1. the quality of the **input data** fed into the system;
2. the quality of the **prompts** used to define agent roles, behavior, and coordination.

In contrast to traditional recommender systems, where noisy or incomplete data directly undermines performance [18], agentic systems may compensate for poor input through the reasoning abilities and pretrained knowledge of LLMs [19]; however, this does not make data quality irrelevant, rather, it suggests the need for rethinking it as a **minimal threshold** that enables effective agent collaboration.

To study this, we first introduce **Relative Data Quality (RDQ)** for measuring performance variation across different versions of the same dataset (e.g., differing in annotation quality, modality richness, or encoding), while keeping architecture and prompting fixed. Comparing across different datasets would introduce confounding factors like domain or sparsity, making such comparisons less informative. RDQ instead isolates the impact of specific data quality dimensions under controlled settings, allowing us to decouple the contribution of data quality from that of the algorithm itself. For example, experiments could assess how changes in image resolution, text length, or preprocessing impact system performance, and to what extent LLMs' reasoning capabilities make them robust to such variations.

Prompting strategies assume a central role in agentic systems, where the prompts define the roles and the allowed actions of each agent [3]. Thus, we propose the **Relative Prompt Quality (RPQ)** to capture the impact of different prompting strategies, e.g., role definitions or in-context examples, on system output. Once fixed an architecture and a version of the data, this measure will highlight how the prompt quality affects the agent coordination and recommendation results. By comparing different prompt versions, we can quantify the sensitivity of the system to prompt formulation, allowing us to identify best practices for guiding agent behavior.

It is worth noticing that absolute evaluation is not meaningful in this context, as removing prompts or input data would render the system non-functional. Instead, we advocate for assessing relative performance variations under controlled and fixed conditions. This approach offers more actionable insights by isolating the impact of specific elements, such as structured metadata or image inputs, on agents' reasoning and decision-making capabilities. These observations can inform system design, highlight critical dependencies, and support more transparent development and debugging processes.

3. Conclusion

This paper laid the groundwork for evaluating how input data quality influences agentic recommender systems, highlighting the often-overlooked role of data, particularly in multimodal, agentic settings.

By analyzing the impact of different modalities and prompt configurations, we aim to encourage more transparent and data-aware evaluation practices. Future work may explore modality-specific quality assessment and adaptive agent routing based on evaluated data quality, positioning data quality as a central concern in agentic RSs.

Acknowledgments

The authors acknowledge the partial support provided by the “Huawei PhD Grant”.

Declaration on Generative AI

The authors have not employed any Generative AI tools.

References

[1] F. Ricci, L. Rokach, B. Shapira (Eds.), Recommender Systems Handbook, Springer US, 2022.

[2] D. Jannach, M. Jugovac, Measuring the business value of recommender systems, ACM Trans. Manag. Inf. Syst. 10 (2019) 16:1–16:23.

[3] Q. Peng, H. Liu, H. Huang, Q. Yang, M. Shao, A survey on llm-powered agents for recommender systems, CoRR abs/2502.10050 (2025).

[4] C. Huang, J. Wu, Y. Xia, Z. Yu, R. Wang, T. Yu, R. Zhang, R. A. Rossi, B. Kveton, D. Zhou, J. J. McAuley, L. Yao, Towards agentic recommender systems in the era of multimodal large language models, CoRR abs/2503.16734 (2025).

[5] M. Valentini, Cooperative and competitive llm-based multi-agent systems for recommendation, in: ECIR (5), volume 15576 of *Lecture Notes in Computer Science*, Springer, 2025, pp. 204–211.

[6] Q. Dong, L. Li, D. Dai, C. Zheng, J. Ma, R. Li, H. Xia, J. Xu, Z. Wu, B. Chang, X. Sun, L. Li, Z. Sui, A survey on in-context learning, in: EMNLP, Association for Computational Linguistics, 2024, pp. 1107–1128.

[7] G. Li, H. Hammoud, H. Itani, D. Khizbullin, B. Ghanem, CAMEL: communicative agents for "mind" exploration of large language model society, in: NeurIPS, 2023.

[8] M. Valentini, A. Ferrara, T. Di Noia, G. Illuzzi, P. Colacicco, Leveraging llm-powered multi-agent systems to enhance customer experience in complex product domains (2025).

[9] S. Yao, J. Zhao, D. Yu, N. Du, I. Shafran, K. R. Narasimhan, Y. Cao, React: Synergizing reasoning and acting in language models, in: ICLR, OpenReview.net, 2023.

[10] Y. Wang, K. Ding, X. Liu, J. Kang, R. A. Rossi, T. Derr, Data quality-aware graph machine learning, in: CIKM, ACM, 2024, pp. 5534–5537.

[11] D. Malitesta, E. Rossi, C. Pomo, T. D. Noia, F. D. Malliaros, Do we really need to drop items with missing modalities in multimodal recommendation?, in: CIKM, ACM, 2024, pp. 3943–3948.

[12] Z. Wang, Y. Yu, W. Zheng, W. Ma, M. Zhang, Macrec: A multi-agent collaboration framework for recommendation, in: SIGIR, ACM, 2024, pp. 2760–2764.

[13] J. Zhang, Y. Hou, R. Xie, W. Sun, J. J. McAuley, W. X. Zhao, L. Lin, J. Wen, Agentcf: Collaborative learning with autonomous language agents for recommender systems, in: WWW, ACM, 2024, pp. 3679–3689.

[14] Y. Shang, P. Liu, Y. Yan, Z. Wu, L. Sheng, Y. Yu, C. Jiang, A. Zhang, F. Xu, Y. Wang, M. Zhang, Y. Li, Agentrecbench: Benchmarking LLM agent-based personalized recommender systems, CoRR abs/2505.19623 (2025).

[15] Y. Wang, Z. Jiang, Z. Chen, F. Yang, Y. Zhou, E. Cho, X. Fan, Y. Lu, X. Huang, Y. Yang, Recmind: Large language model powered agent for recommendation, in: NAACL-HLT (Findings), Association for Computational Linguistics, 2024, pp. 4351–4364.

[16] F. Tian, D. Ganguly, C. Macdonald, Is relevance propagated from retriever to generator in rag?, in: ECIR (1), volume 15572 of *Lecture Notes in Computer Science*, Springer, 2025, pp. 32–48.

[17] C. Bauer, L. Chen, N. Ferro, N. Fuhr, A. Anand, T. Breuer, G. Faggioli, O. Frieder, H. Joho, J. Karlgren, J. Kiesel, B. P. Knijnenburg, A. Lipani, L. Michiels, A. Papenmeier, M. S. Pera, M. Sanderson, S. Sanner, B. Stein, J. R. Trippas, K. Verspoor, M. C. Willemsen, Manifesto from dagstuhl perspectives workshop 24352 - conversational agents: A framework for evaluation (CAFE), CoRR abs/2506.11112 (2025).

[18] Y. Li, Q. Zhao, C. Lin, J. Su, Z. Zhang, Who to align with: Feedback-oriented multi-modal alignment in recommendation systems, in: SIGIR, ACM, 2024, pp. 667–676.

[19] L. Wu, Z. Zheng, Z. Qiu, H. Wang, H. Gu, T. Shen, C. Qin, C. Zhu, H. Zhu, Q. Liu, H. Xiong, E. Chen, A survey on large language models for recommendation, World Wide Web (WWW) 27 (2024) 60.