

Advancing User-Centric Evaluation and Design of Conversational Recommender Systems

Michael Müller

University of Innsbruck

Innsbruck, Austria

michael.m.mueller@uibk.ac.at

Abstract

Conversational Recommender Systems (CRS) are rapidly evolving with advancements in large language models (LLMs), enabling richer, more adaptive user interactions. However, existing evaluation practices remain largely system-centric, underestimating nuanced factors like conversational quality, empathy, and real-world user satisfaction. This doctoral research aims to bridge that gap by advancing holistic, user-centric evaluation frameworks for CRS. The work pursues four directions: (1) identifying key drivers of user satisfaction through targeted user studies and dataset analyses; (2) systematically investigating LLMs as annotators and user simulators to support scalable CRS assessment; (3) developing scalable, standardized evaluation protocols that balance objective accuracy with subjective conversational experience; and (4) deriving actionable design guidelines by comparing strategies for preference elicitation and context integration. Ultimately, this research seeks to provide reproducible methods, and evidence-based guidance to foster the development of CRS that genuinely center the user.

CCS Concepts

• Information systems → Recommender systems.

Keywords

Conversational Recommender Systems, User-Centric, Evaluation, User Interaction, Large Language Models, Personalization

ACM Reference Format:

Michael Müller. 2025. Advancing User-Centric Evaluation and Design of Conversational Recommender Systems. In *Proceedings of the Nineteenth ACM Conference on Recommender Systems (RecSys '25)*, September 22–26, 2025, Prague, Czech Republic. ACM, New York, NY, USA, 6 pages. <https://doi.org/10.1145/3705328.3748759>

1 Introduction

Recommender systems have become an indispensable component of modern online services, guiding users through an ever-growing sea of content and products. Traditional approaches—predominantly collaborative filtering and content-based methods—rely on implicit feedback (e.g., clicks, ratings) and static interfaces, which limit the system’s ability to capture users’ evolving needs and contextual

nuances. CRS address these shortcomings by supporting natural, multi-turn dialogues that allow users to articulate preferences in their own words, receive interactive explanations, and refine recommendations in real time [16]. These systems are commonly built on three main modules—recommendation engine, dialogue management, and natural language processing—that together enable dynamic, user-adaptive conversations [9]. The two primary goals of CRS are therefore (1) to elicit accurate and fine-grained user preferences through dialogue, and (2) to generate personalized suggestions that adapt dynamically as the conversation unfolds.

Recent advances in LLMs have significantly expanded the capabilities of CRS, enabling more fluent conversations, richer context understanding, and the integration of external knowledge [29, 30, 35]. However, these technical advances also expose new challenges in how to evaluate and design CRS from a truly user-centric perspective. Traditional evaluation methods—often centered on offline accuracy metrics such as precision, recall, or BLEU scores—fail to capture crucial aspects of the conversational experience, such as empathy, rapport, adaptability, and overall user satisfaction [18, 25, 35]. In many scenarios, especially for high-involvement or cold-start items, users may not have well-defined preferences at the outset, or may even construct their preferences dynamically during the conversation [16]. Thus, the quality of the dialogue—how well the system understands, adapts, and responds to the user—becomes as important as the accuracy of the recommendations themselves.

Despite growing recognition of these issues, the evaluation of CRS remains an open problem. Existing datasets and offline benchmarks provide limited insight into real user experience, and there is a lack of standardized frameworks for assessing conversational quality and user satisfaction. Recent research has begun to address this gap by developing user-centric evaluation frameworks, such as CRS-Que [18], which extend established models (e.g., ResQue [26]) to include conversation-specific dimensions like understanding, adaptability, response quality, attentiveness, rapport, and humaneness. Yet, conducting large-scale user studies is resource-intensive, and the potential for LLMs to serve as scalable annotators or user simulators introduces both opportunities and open questions regarding their reliability and validity compared to human feedback [6, 25]. Importantly, the recent *CAFE* framework [2] also identifies this gap, emphasizing that current evaluation practices often overlook user-centric aspects and proposing a comprehensive structure for evaluating conversational agents that centers stakeholder goals, user tasks, evaluation criteria, and methodologies.

In this work, I focus on advancing the user-centric evaluation and design of conversational recommender systems. My research

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).

RecSys '25, Prague, Czech Republic

© 2025 Copyright held by the owner/author(s).

ACM ISBN 979-8-4007-1364-4/25/09

<https://doi.org/10.1145/3705328.3748759>

aims to answer the following key questions: What do users truly expect from CRS? How can we systematically evaluate CRS to ensure these expectations are met? Can LLMs reliably support user-centric evaluation at scale? And, ultimately, how should CRS be designed to maximize user satisfaction, trust, and engagement during interaction? By addressing these questions through a combination of user studies, analysis of existing datasets, and exploration of LLM-based evaluation methods, this work seeks to contribute both methodological insights and practical guidelines for the next generation of conversational recommender systems.

2 Related Work

The evaluation and design of CRS have attracted increasing attention in recent years. Prior research has explored both system-centric and user-centric perspectives, highlighting the limitations of traditional accuracy-based metrics and the importance of user experience, empathy, and adaptability. With the rise of LLMs, new opportunities and challenges have emerged for both the development and assessment of CRS. In this section, we review key contributions in these areas, including the integration of LLMs, methods for preference elicitation, the incorporation of contextual information, and recent advances in evaluation protocols.

2.1 Evaluation of CRS

Existing evaluation methods for conversational recommender systems fall into three paradigms: offline studies, user studies, and user simulation with LLM annotation.

2.1.1 Offline Studies. Offline studies remain a foundational paradigm for benchmarking CRS, relying on established conversational datasets and automated metrics. Prominent corpora such as ReDial [23] enable evaluation at scale by providing naturalistic dialogues annotated with recommendation intents. TG-ReDial [37] expands on this by introducing topic-guided threads, facilitating more proactive and context-rich conversational flows, while E-ReDial [11] offers human-curated explanations to enhance the assessment of system-generated explanations.

Evaluation in this setting primarily utilizes objective, reproducible metrics—such as precision, recall, BLEU, and ROUGE—to quantify system performance [11, 16, 23, 37]. Recent advances like RecBench+ [13] further broaden benchmark complexity to better reflect the diversity of real user queries.

However, despite their scalability, offline studies are limited in capturing subjective dimensions critical to user-centric evaluation. These approaches typically overlook conversational coherence across multiple turns, adaptability to emerging user preferences, and factors such as empathy, rapport, and trust [14, 15, 18, 25, 35]. As a result, gains in automated metrics often fail to translate into improved real-world experiences.

2.1.2 User Studies. User studies are widely regarded as the definitive method for evaluating real-world user satisfaction, trust, and engagement with CRS. Early efforts, such as the ResQue model [26], established robust instruments for measuring perceived recommendation quality. CRS-Que [17, 18] extends this by introducing six key conversation-specific dimensions—*understanding*, *adaptability*, *response quality*, *attentiveness*, *rapport*, and *humanness*—which are

now commonly used to systematically assess dialogue quality and user experience.

CRS-Que has been validated with controlled user studies using music (MusicBot) and mobile phone (PhoneBot) CRS [18]. Participants interacted with varied conversational behaviors, including explanation styles and personalization, then rated their experiences using CRS-Que. Results showed that dimensions such as attentiveness, adaptation, and perceived humanness significantly contributed to satisfaction and trust. Moreover, features like human-like cues and tailored explanations influenced perceptions of rapport and behavioral intention, underlining the impact of design choices on subjective experience.

Recent work demonstrates the depth achievable with longitudinal, user-driven studies. A three-week diary study with daily users of music recommender services interacting with GPT-based CRS agents was conducted [31]. The study design allowed participants to explore and optimize interactions in open-ended, personally relevant scenarios. Findings revealed that LLM-powered CRS aided preference clarification, supported unique behaviors, and fostered greater self-understanding during recommendations.

A large-scale user study [25] benchmarked LLM-based CRS (ChatGPT) against strong retrieval models. Users rated dialogue responses for meaningfulness, and ChatGPT outperformed competitors on perceived response quality, though occasionally missed multi-topic context. However, there remained only weak alignment between automated metrics (BLEU, ROUGE) and user satisfaction. This underscores the continued necessity of human-in-the-loop studies for understanding nuanced conversational quality.

Finally, crowdsourced evaluation platforms (e.g., CRS Arena [3]) provide accessible but less controlled alternatives to lab studies, enabling broader participation at the cost of experimental rigor.

As research on CRS evaluation advances, continued investment in well-designed user studies will be critical to ensure that improvements in system capabilities translate into authentic gains in user experience. They provide both the ground truth for constructing new metrics and the foundation for designing systems that genuinely center user needs.

2.1.3 User Simulation and LLM Annotation. User simulation and LLM annotation have emerged as scalable alternatives to traditional user studies for evaluating CRS. These approaches leverage large language models to both simulate user interactions and automate the assessment of conversational quality, aiming to address the high cost and limited scalability of human-in-the-loop evaluation.

Recent frameworks employ agent-based simulators, such as RecUserSim [5], which generate persona-driven dialogues and provide multi-dimensional ratings, and SimpleUserSim [38], which introduces intent-aware prompting and data-leakage controls for greater realism. Automated toolkits can now assess conversational flow, task success, and user satisfaction without exhaustive human annotation [1, 34, 39]. Protocols like CONCEPT [12] integrate both system- and user-centric criteria—including social intelligence and adaptability—using LLM-based simulation and assessment.

Despite these advances, several challenges remain. LLM-based simulators can suffer from prompt brittleness, data leakage, and limited behavioral diversity [38]. While some studies report that

Table 1: Comparison of CRS Evaluation Approaches

Approach	User-Centric	Scalable	LLM-Based	Reproducible	Validated
Offline Metrics	✗	✓	✗	✓	✗
User Studies	✓	✗	✗	✗	✓
LLM Annotation/Simulation	?	✓	✓	?	?
Synthetic Dataset	?	✓	✓	✓	?

LLM-generated annotations closely match human ratings for subjective dimensions [25], others find that crowdsourced human judgments are more reliable, and that reference-based metrics often fail to capture human preferences [10]. As a result, the reliability and validity of LLMs as user simulators and annotators require further systematic comparison with human studies across diverse CRS scenarios [5, 6].

In summary, user simulation and LLM annotation offer promising paths toward scalable, nuanced CRS evaluation, but rigorous validation and methodological refinement are needed to ensure their effectiveness and trustworthiness.

2.1.4 Synthetic Datasets. Synthetic conversational data remains under-studied in CRS evaluation. Most existing literature focuses on generating synthetic datasets for training recommender systems, with less attention paid to their role in evaluation. A multi-dimensional framework from recommender systems—combining human realism ratings, classifier-based real-vs-fake tests, and entropy or divergence statistics—has been adapted to assess the quality of synthetic utterances [20]. For example, the TtWMusic pipeline can generate over one million music-domain dialogues; with carefully crafted prompts, these synthetic conversations closely match human data in terms of fluency and next-item prediction [21].

In parallel, GAN-based frameworks have been proposed for generating fully parameterized synthetic datasets for collaborative filtering, enabling controlled experimentation and benchmarking [4].

Table 1 summarizes these paradigms across five key criteria: user-centricity, scalability, use of LLMs, reproducibility, and validation. Offline metrics offer reproducibility and scalability but lack user-centric depth. User studies provide validated, user-centered insights but are resource-intensive. LLM annotation and simulation promise scalable, LLM-driven evaluation, though their user-centricity and validation remain under investigation. Synthetic datasets enable scalable, reproducible experiments, but their alignment with real user experience is still being established. Developing a unified, standardized evaluation framework that balances these criteria remains an open challenge.

2.2 User-Centric Design of CRS

A truly user-centric Conversational Recommender System (CRS) must go beyond technical optimization to address the nuanced needs, expectations, and behaviors of real users. Effective design requires a holistic approach that integrates preference elicitation, empathetic and adaptive interaction, and contextual grounding of recommendations. These elements are essential for fostering trust, engagement, and satisfaction—core goals highlighted in our research questions.

2.2.1 Preference Elicitation. Recent advances demonstrate that dynamic, multi-turn questioning can substantially improve the quality of elicited preferences. Frameworks such as Vague Preference Policy Learning (VPPL) explicitly model uncertainty in user goals and adapt question selection throughout the dialogue [32, 33]. Multi-agent designs like MACRS coordinate specialized agents (questioner, recommender, chit-chat) and use reflection loops to update user profiles in real time [8]. LLMs can support prompt engineering for elicitation tasks [22] and generate synthetic user-item interactions for cold-start scenarios [28]. However, many users lack the domain knowledge to answer attribute-specific queries; automatic generation of usage-based questions from product reviews has been proposed to bridge this gap [19]. Empirical studies confirm that while elicitation method (button, free-text, open-ended) may not drastically affect overall user experience, satisfaction with the recommendation itself is the strongest predictor of a positive experience [16, 40].

2.2.2 Leveraging LLMs for Interaction. LLMs now play a central role in enhancing CRS user interaction. They enable more natural, adaptive, and empathetic dialogues by understanding nuanced user input, generating contextually appropriate responses, and maintaining conversational coherence [24]. Incorporating in-domain collaborative signals—via fine-tuning or retrieval—improves recommendation relevance without sacrificing fluency. At the same time, LLM-driven recommenders can inherit biases; prompt strategies are being developed to mitigate these effects [7]. Ongoing research explores how LLMs can be guided to exhibit greater empathy and adaptivity, directly impacting user trust and engagement. The ReFICR model [30] unifies retrieval and generation tasks through instruction tuning, achieving state-of-the-art results on public CRS datasets.

2.2.3 Contextual and Semantic Integration. Grounding recommendations in rich context is crucial for user satisfaction. Knowledge graphs inject structured metadata into responses [36, 41], while unified architectures like PECRS represent items as natural-language descriptions and leverage a single pre-trained LM for both semantic matching and dialogue generation [27]. Recent work also demonstrates that contextual and time-aware modeling, with less reliance on domain-specific knowledge, can improve CRS performance and generalizability [29]. These approaches simplify system design, reduce semantic misalignment, and enable CRS to adapt to diverse conversational contexts without requiring separate modules for knowledge retrieval and ranking.

Table 2: Progress of Dissertation Research

Phase	Status	Key Activities	Deliverables
RQ1: User Satisfaction	In progress	• Design and run user studies • Investigate limitations of existing datasets	• Analysis of satisfaction drivers • Draft design guidelines
RQ2: LLM Annotation	Planned	• Build LLM annotation and simulation pipeline • Conduct reliability tests	• LLM vs. human comparison report • Validation protocol
RQ3: Evaluation Framework	Planned	• Integrate system- and user-centric metrics • Evaluate the use of synthetic datasets • Develop open-source evaluation toolkit	• Reproducible evaluation protocol • Synthetic datasets • Toolkit release
RQ4: User-Centric Design	Planned	• Prototype preference-elicitation and context modules	• Evidence-based design guidelines • Prototype code

3 Research Objectives and Questions

Although recent work has advanced both user-centric evaluation frameworks and LLM-based simulation methods for CRS, key gaps remain: (1) there is no unified, scalable evaluation framework that balances system-centric metrics with conversational quality; (2) available datasets and benchmarks do not fully capture real user behavior or conversational nuance; (3) the reliability of large language models as annotators and simulators for CRS evaluation needs systematic validation; (4) design guidelines for integrating preference elicitation and contextual knowledge into CRS are still underdeveloped. Therefore we will address the following research questions (RQs):

- RQ1:** What are the key factors driving user satisfaction and conversational quality in CRS?
- RQ2:** How can large language models be leveraged as reliable annotators and simulators to support CRS evaluation at scale?
- RQ3:** How can we develop a scalable, standardized evaluation framework that balances user-centric and system-centric criteria?
- RQ4:** How should CRS be designed to integrate preference elicitation and contextual information for enhanced user engagement?

4 Methods

The progress plan (Table 2) comprises four phases addressing drivers of user satisfaction, evaluation-framework development, LLM annotation/simulation reliability, and user-centric CRS design. In phase 1, we will run controlled user studies building on existing evaluation frameworks [5, 12, 18], collecting multi-dimensional ratings on conversational behaviors (e.g., empathy, adaptivity) and annotating existing dialogues (e.g., ReDial) to identify attributes most strongly correlated with user satisfaction. In phase 2, we will

conduct comparative experiments where LLMs and human annotators independently rate the same dialogue sessions, computing inter-rater agreement and qualitatively analyzing strengths and limitations; concurrently, we will evaluate LLM-driven user simulators for behavioral fidelity, diversity, and prompt robustness. Phase 3 focuses on building a modular evaluation framework that combines system-centric metrics with user-centric dimensions, supporting both manual and automated assessments via LLMs. In this phase, we will investigate the suitability and impact of incorporating synthetic conversational datasets into the evaluation process—systematically comparing results from real and synthetic dialogues to assess their alignment and utility for CRS evaluation. Open-source tools and protocols will be published to ensure transparency. Phase 4 involves prototyping CRS variants with different preference-elicitation (e.g., direct queries, usage-based questions) and context-integration (e.g., knowledge graphs) strategies, measuring engagement, trust, and satisfaction through surveys and behavioral metrics to derive actionable design guidelines. In parallel, we will continue to explore the generation and assessment of synthetic datasets, informed by findings from phase 3.

5 Conclusion

Advancing the evaluation and design of user-centered Conversational Recommender Systems from a user-centric perspective remains a critical and open challenge. This work aims to address persistent gaps in current evaluation practices by systematically integrating user studies, LLM-based annotation, and synthetic datasets into a unified framework. The proposed approach will clarify which conversational and system factors truly drive user satisfaction and trust, develop scalable methodologies for robust evaluation, and

derive practical guidelines for CRS design. The resulting tools, protocols, and findings are expected to support reproducible research, foster user-centered innovation, and guide practitioners in building CRS that deliver enhanced real-world value.

Acknowledgments

The author would like to thank Dr. Eva Zangerle and Prof. Günther Specht for their valuable feedback and support throughout this project.

References

- [1] Jafar Afzali, Aleksander Mark Drzewiecki, Krisztian Balog, and Shuo Zhang. 2023. UserSimCRS: A User Simulation Toolkit for Evaluating Conversational Recommender Systems. In *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining (WSDM '23)*. Association for Computing Machinery, New York, NY, USA, 1160–1163. doi:10.1145/3539597.3573029
- [2] Christine Bauer, Li Chen, Nicola Ferro, and Norbert Fuhr. 2025. Conversational Agents: A Framework for Evaluation (CAFE) (Dagstuhl Perspectives Workshop 24352). *Dagstuhl Reports* 14, 8 (2025), 53–58. doi:10.4230/DagRep.14.8.53
- [3] Nolwenn Bernard, Hideaki Joko, Faegheh Hasibi, and Krisztian Balog. 2025. CRS Arena: Crowdsourced Benchmarking of Conversational Recommender Systems. In *Proceedings of the Eighteenth ACM International Conference on Web Search and Data Mining (Hannover, Germany) (WSDM '25)*. Association for Computing Machinery, New York, NY, USA, 1028–1031. doi:10.1145/3701551.3704120
- [4] Jesús Bobadilla, Abraham Gutiérrez, Racié Yera, and Luis Martínez. 2023. Creating synthetic datasets for collaborative filtering recommender systems using generative adversarial networks. *Knowledge-Based Systems* 280 (2023), 111016. doi:10.1016/j.knosys.2023.111016
- [5] Luyu Chen, Quanyu Dai, Zeyu Zhang, Xueyang Feng, Mingyu Zhang, Pengcheng Tang, Xu Chen, Yue Zhu, and Zhenhua Dong. 2025. RecUserSim: A Realistic and Diverse User Simulator for Evaluating Conversational Recommender Systems. In *Companion Proceedings of the ACM on Web Conference 2025 (Sydney NSW, Australia) (WWW '25)*. Association for Computing Machinery, New York, NY, USA, 133–142. doi:10.1145/3701716.3715258
- [6] Nuo Chen, Quanyu Dai, Xiaoyu Dong, Xiao-Ming Wu, and Zhenhua Dong. 2025. Large Language Models as Evaluators for Conversational Recommender Systems: Benchmarking System Performance from a User-Centric Perspective. doi:10.48550/arXiv.2501.09493 arXiv:2501.09493 [cs.IR]
- [7] Yashar Deldjoo. 2024. Understanding Biases in ChatGPT-based Recommender Systems: Provider Fairness, Temporal Stability, and Recency. doi:10.48550/arXiv.2401.10545 arXiv:2401.10545 [cs.IR]
- [8] Jiabao Fang, Shen Gao, Pengjie Ren, Xiuying Chen, Suzan Verberne, and Zhaochun Ren. 2024. A Multi-Agent Conversational Recommender System. doi:10.48550/arXiv.2402.01135 arXiv:2402.01135 [cs.IR]
- [9] Chongming Gao, Wenqiang Lei, Xiangnan He, Maarten de Rijke, and Tat-Seng Chua. 2021. Advances and challenges in conversational recommender systems: A survey. *AI Open* 2 (jan 2021), 100–126. doi:10.1016/j.aiopen.2021.06.002
- [10] Lukas Gienapp, Tim Hagen, Maik Fröbe, Matthias Hagen, Benno Stein, Martin Posthast, and Harrison Scells. 2025. The Viability of Crowdsourcing for RAG Evaluation. (2025). doi:10.48550/arXiv.2504.15689 arXiv:2504.15689 [cs.IR] arXiv preprint.
- [11] Shuyu Guo, Shuo Zhang, Weiwei Sun, Pengjie Ren, Zhumin Chen, and Zhaochun Ren. 2023. Towards Explainable Conversational Recommender Systems. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval (Taipei, Taiwan) (SIGIR '23)*. Association for Computing Machinery, New York, NY, USA, 2786–2795. doi:10.1145/3539618.3591884
- [12] Chen Huang, Peixin Qin, Yang Deng, Wenqiang Lei, Jiancheng Lv, and Tat-Seng Chua. 2024. Concept – An Evaluation Protocol on Conversational Recommender Systems with System-centric and User-centric Factors. doi:10.48550/arXiv.2404.03304 arXiv:2404.03304 [cs.CL]
- [13] Jiani Huang, Shijie Wang, Liang bo Ning, Wenqi Fan, Shuaiqiang Wang, Dawei Yin, and Qing Li. 2025. Towards Next-Generation Recommender Systems: A Benchmark for Personalized Recommendation Assistant with LLMs. doi:10.48550/arXiv.2503.09382 arXiv:2503.09382 [cs.IR]
- [14] Dietmar Jannach. 2023. Evaluating conversational recommender systems. *Artificial Intelligence Review* 56, 3 (mar 2023), 2365–2400. doi:10.1007/s10462-022-10229-x
- [15] Dietmar Jannach and Christine Bauer. 2020. Escaping the McNamara fallacy: Towards more impactful recommender systems research. *Ai Magazine* 41, 4 (2020), 79–95. doi:10.1609/aimag.v41i4.5312
- [16] Dietmar Jannach, Ahtsham Manzoor, Wanling Cai, and Li Chen. 2021. A Survey on Conversational Recommender Systems. *ACM Comput. Surv.* 54, 5 (may 2021), 105:1–105:36. doi:10.1145/3453154
- [17] Yucheng Jin, Li Chen, Wanling Cai, and Pearl Pu. 2021. Key Qualities of Conversational Recommender Systems: From Users' Perspective. In *Proceedings of the 9th International Conference on Human-Agent Interaction (Virtual Event, Japan) (HAI '21)*. Association for Computing Machinery, New York, NY, USA, 93–102. doi:10.1145/3472307.3484164
- [18] Yucheng Jin, Li Chen, Wanling Cai, and Xianglin Zhao. 2024. CRS-Que: A User-centric Evaluation Framework for Conversational Recommender Systems. *ACM Trans. Recomm. Syst.* 2, 1, Article 2 (March 2024), 34 pages. doi:10.1145/3631534
- [19] Ivica Kostrić, Krisztian Balog, and Filip Radlinski. 2024. Generating Usage-related Questions for Preference Elicitation in Conversational Recommender Systems. *ACM Trans. Recomm. Syst.* 2, 2, Article 12 (April 2024), 24 pages. doi:10.1145/3629981
- [20] Harsh Lara and Manoj Tiwari. 2022. Evaluation of Synthetic Datasets for Conversational Recommender Systems. doi:10.48550/arXiv.2212.08167 arXiv:2212.08167 [cs.CL]
- [21] Megan Leszczynski, Shu Zhang, Ravi Ganti, Krisztian Balog, Filip Radlinski, Fernando Pereira, and Arun Tejasvi Chaganty. 2023. Talk the Walk: Synthetic Data Generation for Conversational Music Recommendation. doi:10.48550/arXiv.2301.11489 arXiv:2301.11489 [cs.IR]
- [22] Belinda Z. Li, Alex Tamkin, Noah Goodman, and Jacob Andreas. 2023. Eliciting Human Preferences with Language Models. doi:10.48550/arXiv.2310.11589 arXiv:2310.11589 [cs.CL]
- [23] Raymond Li, Samira Kahou, Hannes Schulz, Vincent Michalski, Laurent Charlin, and Chris Pal. 2018. Towards Deep Conversational Recommendations. In *Advances in Neural Information Processing Systems 31 (NIPS 2018) (Montréal, Canada) (NIPS '18)*. Curran Associates Inc., Red Hook, NY, USA, 9748–9758.
- [24] Jianghao Lin, Xinyi Dai, Yunjia Xi, Weiwen Liu, Bo Chen, Hao Zhang, Yong Liu, Chuhan Wu, Xiangyang Li, Chenxu Zhu, Huifeng Guo, Yong Yu, Ruiming Tang, and Weinan Zhang. 2025. How Can Recommender Systems Benefit from Large Language Models: A Survey. *ACM Trans. Inf. Syst.* 43, 2 (Jan. 2025), 28:1–28:47. doi:10.1145/3678004
- [25] Ahtsham Manzoor, Samuel C. Ziegler, Klaus Maria. Pirker Garcia, and Dietmar Jannach. 2024. ChatGPT as a Conversational Recommender System: A User-Centric Analysis. In *Proceedings of the 32nd ACM Conference on User Modeling, Adaptation and Personalization (Cagliari, Italy) (UMAP '24)*. Association for Computing Machinery, New York, NY, USA, 267–272. doi:10.1145/3627043.3659574
- [26] Pearl Pu, Li Chen, and Rong Hu. 2011. A user-centric evaluation framework for recommender systems. In *Proceedings of the fifth ACM conference on Recommender systems (RecSys '11)*. Association for Computing Machinery, New York, NY, USA, 157–164. doi:10.1145/2043932.2043962
- [27] Mathieu Ravaut, Hao Zhang, Lu Xu, Aixin Sun, and Yong Liu. 2024. Parameter-efficient conversational recommender system as a language processing task. *arXiv preprint arXiv:2401.14194* (2024).
- [28] Jianling Wang, Haokai Lu, James Caverlee, Ed H. Chi, and Minmin Chen. 2024. Large Language Models as Data Augmenters for Cold-Start Item Recommendation. In *Companion Proceedings of the ACM Web Conference 2024 (WWW '24)*. Association for Computing Machinery, New York, NY, USA, 726–729. doi:10.1145/3589335.3651532
- [29] Lingzhi Wang, Shafiq Joty, Wei Gao, Xingshan Zeng, and Kam-Fai Wong. 2024. Improving Conversational Recommender System Via Contextual and Time-Aware Modeling With Less Domain-Specific Knowledge. *IEEE Transactions on Knowledge and Data Engineering* 36, 11 (2024), 6447–6461. doi:10.1109/TKDE.2024.3397321
- [30] Ting Yang and Li Chen. 2024. Unleashing the Retrieval Potential of Large Language Models in Conversational Recommender Systems. In *Proceedings of the 18th ACM Conference on Recommender Systems (Bari, Italy) (RecSys '24)*. Association for Computing Machinery, New York, NY, USA, 43–52. doi:10.1145/3640457.3688146
- [31] Sojeong Yun and Youn-kyung Lim. 2025. User Experience with LLM-powered Conversational Recommendation Systems: A Case of Music Recommendation. doi:10.1145/3706598.3713347 arXiv:2502.15229 [cs]
- [32] Gangyi Zhang. 2023. User-Centric Conversational Recommendation: Adapting the Need of User with Large Language Models. In *Proceedings of the 17th ACM Conference on Recommender Systems (Singapore, Singapore) (RecSys '23)*. Association for Computing Machinery, New York, NY, USA, 1349–1354. doi:10.1145/3604915.3608885
- [33] Gangyi Zhang, Chongming Gao, Wenqiang Lei, Xiaojie Guo, Shijun Li, Hongshen Chen, Zhuozhi Ding, Sulong Xu, and Lingfei Wu. 2025. Vague Preference Policy Learning for Conversational Recommendation. *ACM Trans. Inf. Syst.* 43, 3, Article 78 (May 2025), 27 pages. doi:10.1145/3717831
- [34] Shuo Zhang and Krisztian Balog. 2020. Evaluating Conversational Recommender Systems via User Simulation. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD '20)*. Association for Computing Machinery, New York, NY, USA, 1512–1520. doi:10.1145/3394486.3403202
- [35] Xiaoyu Zhang, Ruobing Xie, Yougang Lyu, Xin Xin, Pengjie Ren, Mingfei Liang, Bo Zhang, Zhanhui Kang, Maarten de Rijke, and Zhaochun Ren. 2024. Towards Empathetic Conversational Recommender Systems. In *Proceedings of the 18th ACM Conference on Recommender Systems (RecSys '24)*. Association for Computing Machinery, New York, NY, USA, 84–93. doi:10.1145/3640457.3688133

- [36] Kun Zhou, Wayne Xin Zhao, Shuqing Bian, Yuanhang Zhou, Ji-Rong Wen, and Jingsong Yu. 2020. Improving Conversational Recommender Systems via Knowledge Graph based Semantic Fusion. [doi:10.48550/arxiv.2007.04032](https://doi.org/10.48550/arxiv.2007.04032) [cs.CL]
- [37] Kun Zhou, Yuanhang Zhou, Wayne Xin Zhao, Xiaoke Wang, and Ji-Rong Wen. 2020. Towards Topic-Guided Conversational Recommender System. [doi:10.48550/arxiv.2010.04125](https://doi.org/10.48550/arxiv.2010.04125) [cs.CL]
- [38] Lixi Zhu, Xiaowen Huang, and Jitao Sang. 2024. How Reliable is Your Simulator? Analysis on the Limitations of Current LLM-based User Simulators for Conversational Recommendation. In *Companion Proceedings of the ACM Web Conference 2024* (Singapore, Singapore) (WWW '24). Association for Computing Machinery, New York, NY, USA, 1726–1732. [doi:10.1145/3589335.3651955](https://doi.org/10.1145/3589335.3651955)
- [39] Lixi Zhu, Xiaowen Huang, and Jitao Sang. 2025. A LLM-based Controllable, Scalable, Human-Involved User Simulator Framework for Conversational Recommender Systems. In *Proceedings of the ACM on Web Conference 2025 (WWW '25)*. Association for Computing Machinery, New York, NY, USA, 4653–4661. [doi:10.1145/3696410.3714858](https://doi.org/10.1145/3696410.3714858)
- [40] Liv Ziegfeld, Daan Di Scala, and Anita H. M. Cremers. 2025. The effect of preference elicitation methods on the user experience in conversational recommender systems. *Computer Speech & Language* 89 (Jan. 2025), 101696. [doi:10.1016/j.csl.2024.101696](https://doi.org/10.1016/j.csl.2024.101696)
- [41] Jie Zou, Cheng Lin, Weikang Guo, Zheng Wang, Jiwei Wei, Yang Yang, and Hengtao Shen. 2025. Multi-Type Context-Aware Conversational Recommender Systems via Mixture-of-Experts. [doi:10.48550/arxiv.2504.13655](https://doi.org/10.48550/arxiv.2504.13655) [cs.CL]