Yearning for Better Metrics: Revisiting the ReDial Dataset for Evaluating Conversational Recommender Systems

Michael Müller[1,*], Amir Reza Mohammadi[1], Andreas Peintner[1], Beatriz Barroso Gstrein[1], Eva Zangerle[1] and Günther Specht[1]

[1]Department of Computer Science, University of Innsbruck, Austria

Abstract

Current evaluation of Conversational Recommender Systems (CRS) relies heavily on automatic metrics such as Accuracy, BLEU and ROUGE, often using datasets like ReDial. However, these metrics and benchmarks overlook essential user-centric aspects of conversational quality. In this paper, we take a user-focused perspective by applying Large Language Model (LLM) annotators and the CRS-Que framework to over 10,000 ReDial conversations. Our analysis uncovers significant limitations: ReDial exhibits strong popularity bias, conversations are brief and lack qualities such as adaptability, humanness, and rapport, and traditional metrics fail to capture these dimensions. These results highlight the need for richer, multi-dimensional evaluation protocols and improved datasets that better reflect authentic user experience. We discuss implications for future CRS research and the development of user-centric assessment frameworks.

Keywords

Conversational Recommender Systems, User Centric

1. Introduction

Conversational recommender systems (CRS) are increasingly central to online platforms, enabling users to express preferences and receive personalized suggestions through natural, multi-turn dialogue [1, 2]. Over the past decade, CRS have evolved from rule-based and retrieval models to sophisticated neural architectures, culminating in the integration of LLMs that offer unprecedented fluency and contextual understanding [3, 4, 5]. These advances have enabled CRS to handle complex user queries, adapt to evolving preferences, and generate more engaging conversational experiences.

Despite these technical improvements, evaluating CRS remains a significant challenge. Traditional evaluation methods—such as BLEU, ROUGE, and offline accuracy metrics—focus on reference-based comparisons and often overlook crucial user-centric aspects like adaptability, rapport, and conversational quality [6, 7, 3]. While these metrics provide reproducible benchmarks, they fail to capture the nuanced qualities that define effective human-computer interaction, such as empathy, transparency, and trust.

Benchmark datasets such as ReDial [8] are widely used for CRS evaluation, but may not fully reflect the diversity and richness of real-world conversations. Recent frameworks, including CRS-Que [6] and CAFE [9], highlight the need for multi-dimensional, user-focused assessment protocols that go beyond surface-level metrics. However, conducting large-scale user studies is resource-intensive, and the reliability of LLMs as scalable annotators or user simulators remains an open question [7, 10].

In this paper, we present our first steps towards a more user-centric evaluation of CRS. We critically examine the ReDial dataset and its evaluation methodology by leveraging LLM annotators and the CRS-Que framework to analyze conversational quality across multiple dimensions. Our exploratory analysis reveals several limitations in both the dataset and prevailing metrics, motivating the need for richer evaluation protocols and improved benchmarks. These initial findings aim to inform future research

36th GI-Workshop on Foundations of Databases (Grundlagen von Datenbanken), September 29 - October 01, 2025, Regensburg, Germany

[*] Corresponding author.

michael.m.mueller@uibk.ac.at (M. Müller)

0009-0005-0918-8833 (M. Müller); 0000-0003-3934-6941 (A. R. Mohammadi); 0000-0001-7337-524X (A. Peintner); 0000-0003-3195-8273 (E. Zangerle); 0000-0003-0978-7201 (G. Specht)

on authentic user-centric assessment and the design of next-generation conversational recommender systems.

2. Related Work

CRS have attracted growing research interest, with evaluation protocols evolving alongside advances in system design. Early work focused on system-centric metrics, relying on datasets such as ReDial [8] for benchmarking recommendation accuracy and conversational fluency. ReDial remains a foundational resource, enabling large-scale evaluation and supporting both automatic and user-centric analysis. Extensions like TG-ReDial [11] and E-ReDial [12] have introduced topic-guided threads and explanation annotations, broadening the scope for context-rich and explainable CRS evaluation.

Traditional evaluation methods, including BLEU and ROUGE [8, 1, 11, 12], provide reproducible benchmarks but often fail to capture subjective qualities critical to user experience. Recent studies highlight that these metrics do not correlate well with user satisfaction or conversational quality [7]. For example, Manzoor et al. [7] demonstrate weak alignment between automated scores and human ratings, underscoring the limitations of reference-based evaluation.

To address these gaps, user-centric frameworks such as ResQue [13] and CRS-Que [6, 14] have been developed, introducing multi-dimensional criteria like adaptability, rapport, humanness, and response quality. These frameworks have been validated in controlled user studies with music and mobile phone CRS agents [6], showing that conversational dimensions significantly influence satisfaction and trust. Large-scale studies, including those by Yun et al. [15] and Manzoor et al. [7], further demonstrate the importance of longitudinal and comparative evaluation, revealing that LLM-powered CRS can aid preference clarification and outperform retrieval models on perceived response quality.

Recent papers have leveraged ReDial for both objective and subjective evaluation protocols. For instance, Zhang et al. [3] use the ReDial dataset to assess recommendation accuracy and the emotional quality of generated responses in empathetic CRS, employing metrics such as Recall@N and AUC, as well as human and LLM-based ratings for user satisfaction. Similarly, Wang et al. [16] utilize ReDial as a benchmark for contextual and time-aware CRS, focusing on extracting internal knowledge from conversation context. Their evaluation combines automatic metrics with user-centric human assessments, where crowd-workers rate generated responses for fluency, informativeness, and coherence on a 0–2 scale, with final scores averaged across annotators.

Crowdsourced platforms such as CRS Arena [17] offer scalable alternatives for user studies, though with reduced experimental control. Meanwhile, the emergence of LLMs has enabled new paradigms for CRS evaluation. LLMs can serve as annotators and user simulators, automating the assessment of conversational quality across multiple dimensions. Frameworks like RecUserSim [18], SimpleUserSim [19], and CONCEPT [20] leverage agent-based simulation and LLM annotation to benchmark CRS agents at scale, integrating criteria such as social intelligence and adaptability.

Despite these advances, challenges remain. LLM-based simulation can suffer from prompt brittleness, data leakage, and limited behavioral diversity [19]. While some studies report that LLM-generated annotations closely match human ratings [7], others find that crowdsourced judgments are more reliable and that reference-based metrics often fail to capture human preferences [21]. The reliability and validity of LLMs as annotators and simulators require further systematic comparison with human studies [18, 10].

In summary, CRS evaluation is shifting toward richer, user-centric protocols that combine large-scale automated annotation with targeted user studies. Our work builds on these trends by critically assessing ReDial and its evaluation methodology, leveraging LLM annotators and CRS-Que to highlight the limitations of current benchmarks and metrics [22, 23, 6, 7, 3]. This motivates the development of improved datasets and multi-dimensional evaluation frameworks for authentic user-centric assessment in CRS research.

3. Background

3.1. ReDial Dataset

The ReDial dataset [8] was developed to facilitate research on conversational recommender systems at the intersection of goal-directed and free-form dialogue. It contains 11,348 human-human conversations about movie recommendations, collected via Amazon Mechanical Turk (AMT) between December 2017 and June 2018.

For data collection, pairs of qualified AMT workers from English-speaking countries were matched in real time and assigned the roles of *seeker* (requesting recommendations) and *recommender*. Workers were required to have a high approval rate and over 1,000 approved HITs. Before each task, participants provided informed consent via a dedicated form outlining the study's purpose and methodology. The custom interface enforced tagging of movie mentions with the '@' symbol, offering a searchable list of movies sourced from DBpedia, with the option to add new titles. Instructions emphasized formal language, a minimum of ten messages per conversation, and discussion of at least four distinct movies. Dialogues not meeting these criteria, or containing offensive or off-topic content, were removed.

Each conversation is stored as a JSON object in `jsonl` format, with fields for conversation and worker IDs, a list of messages, movie mentions, and per-movie labels indicating whether the seeker or recommender has seen, liked, or suggested each movie. Messages include text, sender ID, and time offset. No additional preprocessing was performed on the released data.

3.2. BLEU and ROUGE Scores

BLEU (BiLingual Evaluation Understudy) and ROUGE (Recall-Oriented Understudy for Gisting Evaluation) are the most widely adopted automatic metrics for evaluating natural language generation, including CRS.

BLEU was originally developed for machine translation [24]. It measures the precision of n-gram overlaps between a candidate response and one or more reference responses, penalizing overly short outputs via a brevity penalty. BLEU is computed as:

$$\text{BLEU} = \text{BP} \cdot \exp\Big(\sum_{n=1}^{N} w_n \log p_n\Big),$$

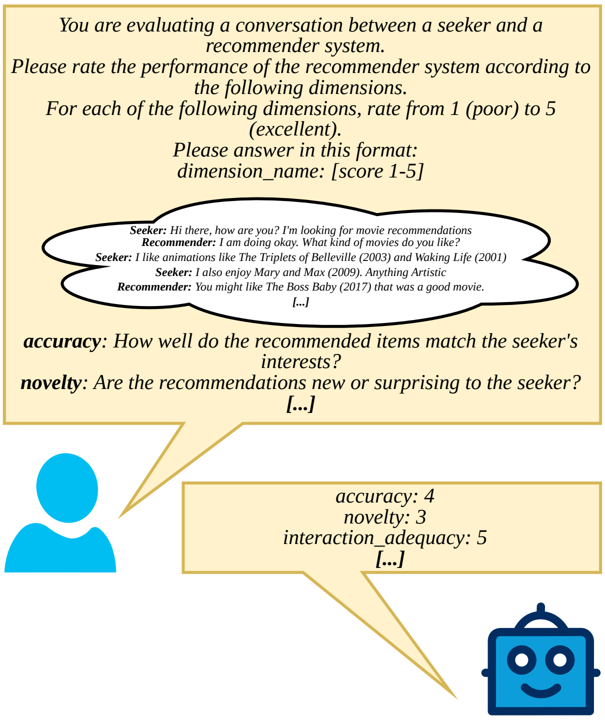
where p_n is the precision of n-grams, w_n are weights (typically uniform), and BP is the brevity penalty. BLEU can be calculated for different n-gram orders (e.g., BLEU-1, BLEU-4), with higher n capturing more contextual similarity.

ROUGE was introduced for automatic summarization [25]. Unlike BLEU, ROUGE is recall-oriented, focusing on how much of the reference content is covered by the candidate. Common variants include ROUGE-N, which measures recall of matching n-grams; ROUGE-L, which is based on the longest common subsequence (LCS); and ROUGE-S, which measures skip-bigram overlap. For example, ROUGE-N is computed as:

$$\text{ROUGE-N} = \frac{\sum_{\text{ngram} \in \text{Ref}} \min(\text{Count}_{\text{cand}}(\text{ngram}), \text{Count}_{\text{ref}}(\text{ngram}))}{\sum_{\text{ngram} \in \text{Ref}} \text{Count}_{\text{ref}}(\text{ngram})}$$

In CRS research, BLEU and ROUGE can be used to evaluate the quality of generated responses by comparing them to human-written reference replies in datasets such as ReDial [8, 7]. These metrics are easy to compute and allow for reproducible benchmarking across models and datasets.

Despite their popularity, BLEU and ROUGE have notable limitations in the context of CRS. Both metrics rely on surface-level n-gram matching, which means they often fail to account for semantic equivalence, paraphrasing, or contextually appropriate responses. They do not measure conversational qualities such as coherence, empathy, adaptability, or user satisfaction, which are critical in CRS.



You are evaluating a conversation between a seeker and a recommender system.
Please rate the performance of the recommender system according to the following dimensions.
For each of the following dimensions, rate from 1 (poor) to 5 (excellent).
Please answer in this format:
dimension_name: [score 1-5]

Seeker: *Hi there, how are you? I'm looking for movie recommendations*
Recommender: *I am doing okay. What kind of movies do you like?*
Seeker: *I like animations like The Triplets of Belleville (2003) and Waking Life (2001)*
Seeker: *I also enjoy Mary and Max (2009). Anything Artistic*
Recommender: *You might like The Boss Baby (2017) that was a good movie.*
[...]

accuracy*: How well do the recommended items match the seeker's interests?*
novelty*: Are the recommendations new or surprising to the seeker?*
[...]

accuracy: 4
novelty: 3
interaction_adequacy: 5
[...]

Figure 1: Overview of the automated annotation process: Each ReDial conversation is formatted and rated by an LLM across multiple user-centric dimensions using the CRS-Que framework.

4. Methods

We began with an exploratory analysis of the ReDial dataset to characterize its conversational structure and recommendation patterns. This included examining the distribution of mentioned movies, identifying potential popularity bias, and analyzing message lengths to assess the depth and diversity of interactions. These insights informed our subsequent annotation and evaluation procedures.

To assess conversational quality in the ReDial dataset, we developed an automated annotation pipeline using LLMs. We processed the 10,003 training conversations with the OpenAI GPT-4.1nano model via the batch API. Each conversation was preprocessed by replacing movie mention tokens with their corresponding movie titles for readability, and formatted as a transcript with each turn labeled by role (*Seeker* or *Recommender*).

For each dialogue, we generated a standardized annotation prompt (see Figure 1) instructing the LLM to rate the recommender's performance across multiple user-centric dimensions [6]. These included both ResQue and CRS-Que criteria: accuracy, novelty, interaction adequacy, explainability, adaptability, understanding, response quality, attentiveness, ease of use, usefulness, user control, transparency, humanness, rapport, trust, satisfaction, and behavioral intention. The LLM assigned a score from 1 (poor) to 5 (excellent) for each dimension, following a specified response format.

Model outputs were parsed to extract scores, and results were validated to ensure all required dimensions were present and within the valid range. Dialogues with incomplete or malformed ratings were excluded from analysis. The annotation process with GPT-4.1nano required approximately 18 hours and cost $0.76; using different model sizes would proportionally affect cost and throughput.

We configured GPT-4.1nano with a temperature of 0.0 to ensure deterministic outputs and reduce annotation variance. The entire annotation process was conducted in a single batch session to maintain consistency. GPT-4.1nano was selected for its cost-effectiveness while maintaining sufficient capabilities for multi-dimensional assessment. The annotation prompt was embedded as a system message to establish consistent evaluation criteria across all conversations.

5. Results

5.1. Exploratory Analysis

We conducted an exploratory analysis of the ReDial dataset to better understand its conversational structure and the diversity of recommendations it contains. Figure 2 illustrates the distribution of all mentioned movies, revealing a pronounced long-tail pattern: a small number of movies are referenced very frequently, while the majority appear only rarely. This concentration suggests a potential popularity bias, where conversations are dominated by well-known or currently popular titles, such as many Marvel movies. As a result, the dataset may be less suitable for evaluating CRS models that aim to recommend diverse or niche items, since most dialogues revolve around a limited set of mainstream movies.

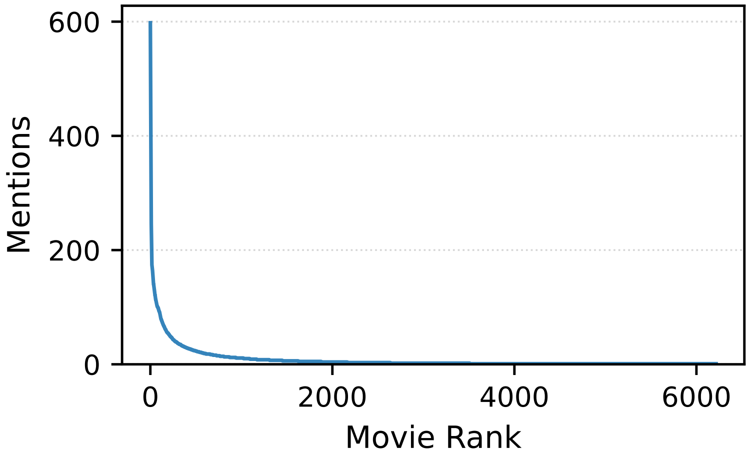


Figure 2: Distribution of all mentioned movies. The distribution is long-tailed, indicating a few movies are mentioned very frequently, while most are mentioned rarely.

This popularity bias raises questions about the representativeness of user preferences and the diversity of recommendations in ReDial. It is possible that recommenders in the dataset tend to suggest movies they have recently watched or that are trending at the time, rather than exploring a broader range of options. Figure 3 further highlights this effect by showing the top 10 most frequently mentioned movies.

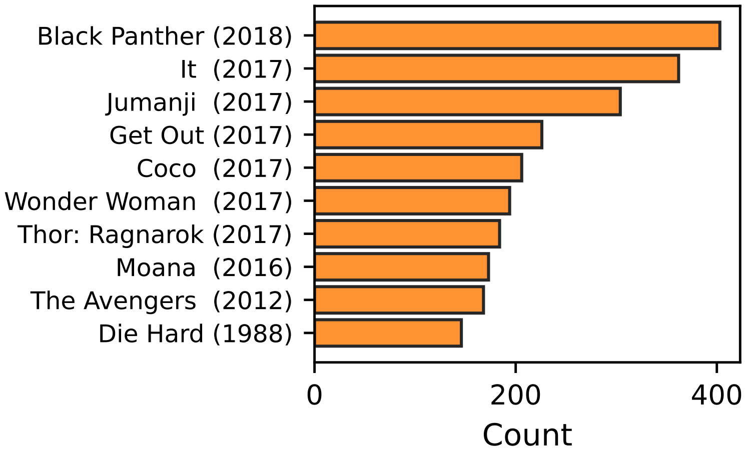


Figure 3: Top 10 most frequently mentioned movies in ReDial. The dominance of popular titles highlights the dataset's limited diversity in recommendations.

We also analyzed the length of messages exchanged in the conversations. As shown in Figure 4, the average message contains only 6.8 words, indicating that interactions are generally brief. Such short messages may not provide sufficient context for effective preference elicitation, potentially limiting the ability of CRS models to understand user needs and deliver personalized recommendations.

Moreover, message length has implications for the use of automatic evaluation metrics. BLEU, being precision-oriented, penalizes candidates that are much shorter than the reference due to the brevity penalty, while ROUGE, which is recall-oriented, may yield lower scores for short candidates that miss important n-grams from the reference. This relationship could help explain why recent studies, such as [7], have found that BLEU and ROUGE scores computed on ReDial conversations do not correlate well with user satisfaction or conversational quality.

Therefore, our exploratory analysis highlights several limitations of the ReDial dataset for user-centric CRS evaluation, including popularity bias, limited diversity, and the brevity of conversational turns. These factors should be considered when interpreting results and designing future benchmarks.

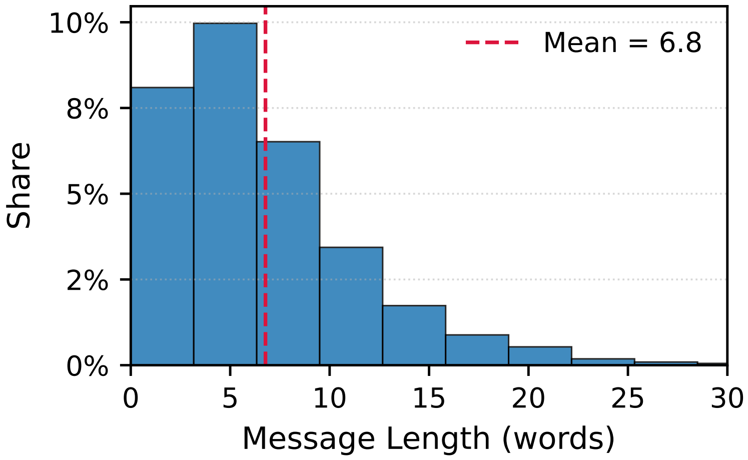


Figure 4: Distribution of message lengths in ReDial conversations. The average message contains only 6.8 words, reflecting the brevity of interactions and limited conversational depth.

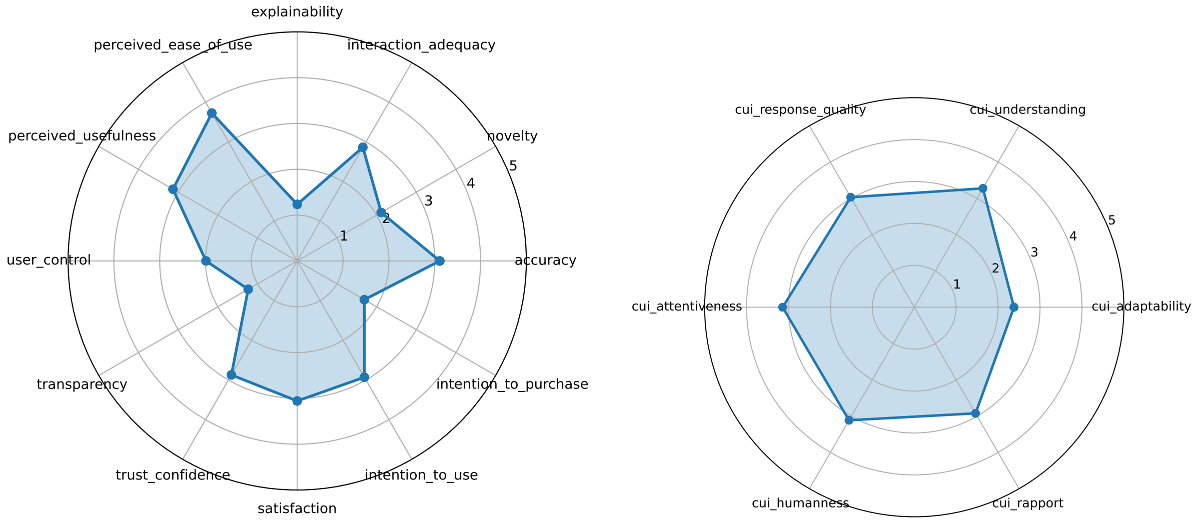
5.2. Annotation Results

Figure 5 summarizes the average LLM-annotated scores for both ResQue and CRS-Que user-centric dimensions across all ReDial conversations.

For the ResQue dimensions (Figure 5a), ratings are moderate overall. Accuracy and usefulness receive higher scores, while novelty, transparency, and explainability are rated lower. This indicates that recommendations are generally relevant but often lack diversity and clear justifications. Combined with the brevity of messages in the dataset, these low scores suggest that users may not have invested much effort in their responses, or may have simply recommended movies they had recently watched. This behavior would also help explain the pronounced long-tail distribution of mentioned movies observed in Figure 2.

The CRS-Que conversational dimensions (Figure 5b) also show middling scores. Notably, humanness is rated only 3 out of 5, despite the conversations being between two humans. This may reflect either limitations in LLM-based annotation or a lack of effort and engagement from participants.

These findings highlight several limitations of using ReDial as a benchmark for user-centric CRS evaluation. The dataset's popularity bias, limited diversity, and brief conversational turns restrict its ability to capture nuanced conversational behaviors. Moreover, the LLM-based annotations suggest that even human-human dialogues in ReDial often lack key qualities such as adaptability, humanness, and rapport—traits essential for effective user interaction. Consequently, ReDial does not provide a sufficient ground truth for evaluating CRS models on user-centric dimensions.



(a) Resque Dimensions

(b) Conversational Dimensions

Figure 5: User-centric evaluation results for ReDial. Left: ResQue dimensions; Right: CRS-Que conversational dimensions. Scores range from 1 (worst) to 5 (best), revealing moderate performance and highlighting key limitations in conversational quality.

6. Future Work

Our initial findings highlight several promising directions for future research in user-centric evaluation of CRS. A primary avenue is the systematic comparison of LLM-based annotations with human ratings through controlled user studies. While LLMs offer scalability, it remains unclear how well their judgments reflect genuine user perceptions of conversational quality. Future work should design experiments that directly compare LLM and human ratings across CRS-Que dimensions, such as adaptability, rapport, and humanness, to establish the reliability and validity of automated annotation. This will be essential for developing robust evaluation protocols that can be adopted by the broader research community.

Another important challenge is to move beyond traditional reference-based metrics like BLEU and ROUGE, which often fail to capture the nuanced qualities of effective CRS interactions. One promising approach is the creation of synthetic datasets composed of high-quality CRS responses, as identified by LLM annotation. These datasets could serve as new benchmarks for evaluating conversational systems, allowing researchers to investigate the correlation between automated metrics and user-centric quality.

The evolution of LLM providers and models also presents an opportunity for longitudinal research. As LLMs continue to improve, it will be important to monitor whether their assessments of User-Centric metrics like CRS-Que dimensions converge with or diverge from human judgments over time. Comparative studies involving multiple LLMs and versions may show the stability and generalizability of automated evaluation, and can inform best practices for using LLMs to benchmark fast or scalable CRS models in production environments.

Preference elicitation remains a critical open problem in CRS research. Our analysis of ReDial suggests that users are often hesitant to declare their preferences, resulting in sparse or ambiguous conversational data. Future work should explore conversational strategies and interface designs that encourage users to share richer and more explicit feedback. This could involve adaptive questioning, context-aware prompts, or the integration of external knowledge to support preference clarification. Understanding how to effectively elicit preferences will not only improve recommendation quality but also enhance the validity of user-centric evaluation.

Finally, broader methodological questions remain regarding the integration of LLM-based evaluation into the CRS development lifecycle. Researchers should investigate how automated annotation can be combined with traditional user studies, crowdsourcing, and simulation to create comprehensive, multi-layered evaluation frameworks. There is also a need to address potential biases in LLM annotation, ensure transparency in evaluation processes, and develop guidelines for interpreting automated scores in the context of real-world user satisfaction and trust.

By pursuing these directions, future research can advance the development of richer, user-focused evaluation protocols, improved datasets, and practical guidelines for authentic user-centric assessment in conversational recommender systems. This will ultimately support the design of CRS that better meet user needs and expectations in diverse application domains.

7. Conclusion

This paper presents our first steps toward a more user-centric evaluation of CRS. Our exploratory analysis of the ReDial dataset revealed a pronounced popularity bias, limited diversity in recommendations, and generally brief conversational turns, all of which constrain the dataset's suitability for evaluating nuanced CRS behaviors. Through multi-dimensional annotation of ReDial conversations using LLMs and the CRS-Que framework, we identified significant limitations in both the dataset and prevailing evaluation metrics. Our analysis shows that human conversations in ReDial often lack key qualities such as adaptability, humanness, and rapport, while traditional metrics like BLEU and ROUGE fail to capture these user-centric aspects. LLM-based annotation offers a scalable approach for multi-dimensional assessment and could be used to test whether smaller, scalable models for production can achieve good user-centric scores—assuming LLMs can reliably judge these qualities. However, accuracy cannot be fully assessed at present due to the lack of ground truth. These findings underscore the need for richer evaluation protocols and improved datasets that better reflect authentic conversational quality and user experience. Further validation of LLM-based annotation against human ratings is required. Future research should focus on developing comprehensive, user-focused evaluation frameworks, exploring new benchmarks, and advancing methods for preference elicitation and conversational design. By addressing these challenges, the CRS community can move toward systems that not only provide accurate recommendations but also foster engaging, adaptive, and satisfying user interactions.

Declaration on Generative AI

The author(s) have not employed any Generative AI tools.

References

[1] D. Jannach, A. Manzoor, W. Cai, L. Chen, A Survey on Conversational Recommender Systems, ACM Comput. Surv. 54 (2021) 105:1–105:36. doi:10.1145/3453154.

[2] C. Gao, W. Lei, X. He, M. de Rijke, T.-S. Chua, Advances and challenges in conversational recommender systems: A survey, AI Open 2 (2021) 100–126. URL: https://www.sciencedirect.com/science/article/pii/S2666651021000164. doi:10.1016/j.aiopen.2021.06.002.

[3] X. Zhang, R. Xie, Y. Lyu, X. Xin, P. Ren, M. Liang, B. Zhang, Z. Kang, M. de Rijke, Z. Ren, 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, 2024, pp. 84–93. doi:10.1145/3640457.3688133.

[4] L. Wang, S. Joty, W. Gao, X. Zeng, K.-F. Wong, Improving conversational recommender system via contextual and time-aware modeling with less domain-specific knowledge, IEEE Transactions on Knowledge and Data Engineering 36 (2024) 6447–6461. doi:10.1109/TKDE.2024.3397321.

[5] T. Yang, L. Chen, Unleashing the retrieval potential of large language models in conversational recommender systems, in: Proceedings of the 18th ACM Conference on Recommender Systems, RecSys '24, Association for Computing Machinery, New York, NY, USA, 2024, p. 43–52. doi:10.1145/3640457.3688146.

[6] Y. Jin, L. Chen, W. Cai, X. Zhao, Crs-que: A user-centric evaluation framework for conversational recommender systems, ACM Trans. Recomm. Syst. 2 (2024). doi:10.1145/3631534.

[7] A. Manzoor, S. C. Ziegler, K. M. P. Garcia, D. Jannach, Chatgpt as a conversational recommender system: A user-centric analysis, in: Proceedings of the 32nd ACM Conference on User Modeling, Adaptation and Personalization, UMAP '24, Association for Computing Machinery, New York, NY, USA, 2024, p. 267–272. doi:10.1145/3627043.3659574.

[8] R. Li, S. Kahou, H. Schulz, V. Michalski, L. Charlin, C. Pal, Towards deep conversational recommendations, in: Advances in Neural Information Processing Systems 31 (NIPS 2018), NIPS'18, Curran Associates Inc., Red Hook, NY, USA, 2018, p. 9748–9758.

[9] C. Bauer, L. Chen, N. Ferro, N. Fuhr, Conversational Agents: A Framework for Evaluation (CAFE) (Dagstuhl Perspectives Workshop 24352), Dagstuhl Reports 14 (2025) 53–58. doi:10.4230/DagRep.14.8.53.

[10] N. Chen, Q. Dai, X. Dong, X.-M. Wu, Z. Dong, Large Language Models as Evaluators for Conversational Recommender Systems: Benchmarking System Performance from a User-Centric Perspective, 2025. doi:10.48550/arXiv.2501.09493. arXiv:2501.09493.

[11] K. Zhou, Y. Zhou, W. X. Zhao, X. Wang, J.-R. Wen, Towards topic-guided conversational recommender system, 2020. doi:10.48550/arxiv.2010.04125. arXiv:2010.04125.

[12] S. Guo, S. Zhang, W. Sun, P. Ren, Z. Chen, Z. Ren, Towards explainable conversational recommender systems, in: Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '23, Association for Computing Machinery, New York, NY, USA, 2023, p. 2786–2795. doi:10.1145/3539618.3591884.

[13] P. Pu, L. Chen, R. Hu, 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, 2011, pp. 157–164. doi:10.1145/2043932.2043962.

[14] Y. Jin, L. Chen, W. Cai, P. Pu, Key qualities of conversational recommender systems: From users' perspective, in: Proceedings of the 9th International Conference on Human-Agent Interaction, HAI '21, Association for Computing Machinery, New York, NY, USA, 2021, p. 93–102. doi:10.1145/3472307.3484164.

[15] S. Yun, Y.-k. Lim, User Experience with LLM-powered Conversational Recommendation Systems: A Case of Music Recommendation, 2025. doi:10.1145/3706598.3713347. arXiv:2502.15229.

[16] J. Wang, H. Lu, J. Caverlee, E. H. Chi, M. Chen, 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, 2024, pp. 726–729. doi:10.1145/3589335.3651532.

[17] N. Bernard, H. Joko, F. Hasibi, K. Balog, Crs arena: Crowdsourced benchmarking of conversational recommender systems, in: Proceedings of the Eighteenth ACM International Conference on Web Search and Data Mining, WSDM '25, Association for Computing Machinery, New York, NY, USA, 2025, p. 1028–1031. doi:10.1145/3701551.3704120.

[18] L. Chen, Q. Dai, Z. Zhang, X. Feng, M. Zhang, P. Tang, X. Chen, Y. Zhu, Z. Dong, Recusersim: A realistic and diverse user simulator for evaluating conversational recommender systems, in: Companion Proceedings of the ACM Web Conference 2025, WWW '25, Association for Computing Machinery, New York, NY, USA, 2025, p. 133–142. doi:10.1145/3701716.3715258.

[19] L. Zhu, X. Huang, J. Sang, 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, WWW '24, Association for Computing Machinery, New York, NY, USA, 2024, p. 1726–1732. doi:10.1145/3589335.3651955.

[20] C. Huang, P. Qin, Y. Deng, W. Lei, J. Lv, T.-S. Chua, Concept – an evaluation protocol on conversational recommender systems with system-centric and user-centric factors, 2024. doi:10.48550/arxiv.2404.03304. arXiv:2404.03304.

[21] L. Gienapp, T. Hagen, M. Fröbe, M. Hagen, B. Stein, M. Potthast, H. Scells, The viability of crowdsourcing for RAG evaluation (2025). doi:10.48550/arXiv.2504.15689. arXiv:2504.15689, arXiv preprint.

[22] D. Jannach, C. Bauer, Escaping the mcnamara fallacy: Towards more impactful recommender systems research, Ai Magazine 41 (2020) 79–95. doi:10.1609/aimag.v41i4.5312.

[23] D. Jannach, Evaluating conversational recommender systems, Artificial Intelligence Review 56 (2023) 2365–2400. doi:10.1007/s10462-022-10229-x.

[24] K. Papineni, S. Roukos, T. Ward, W.-J. Zhu, Bleu: a method for automatic evaluation of machine translation, in: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, ACL '02, Association for Computational Linguistics, USA, 2002, p. 311–318. URL: https://doi.org/10.3115/1073083.1073135. doi:10.3115/1073083.1073135.

[25] C.-Y. Lin, ROUGE: A package for automatic evaluation of summaries, in: Text Summarization Branches Out, Association for Computational Linguistics, Barcelona, Spain, 2004, pp. 74–81. URL: https://aclanthology.org/W04-1013/.