Beyond Homogeneous Users: Simulating Diverse User Personas for Conversational Recommendation Datasets

Alessandro Francesco Maria Martina[1,2,*,†], Alessandro Petruzzelli[1,†], Cataldo Musto[1], Marco de Gemmis[1], Pasquale Lops[1] and Giovanni Semeraro[1]

[1]*University of Bari Aldo Moro, Bari, Italy*
[2]*University of Pisa, Pisa, Italy*

Abstract

Conversational Recommender Systems (CRSs) leverage multi-turn interactions to assist users in decision-making. The quality of these systems is heavily dependent on the datasets used for their training. Existing datasets, often created by crowdworkers or proprietary Large Language Models (LLMs), frequently lack conversational coherence, fail to model diverse user behaviors, and suffer from reproducibility issues. To address these limitations, we introduce DistillRecDial, a conversational recommendation dataset that incorporates varied user personas and interaction patterns. The dataset is generated using a Large-to-Small Language Model Distillation pipeline, enabling dialogue synthesis without reliance on closed, resource-intensive LLMs. We model user heterogeneity in preferences, goals, and dialogic behavior, resulting in a more realistic and diverse corpus. Human and automated evaluations indicate that DistillRecDial surpasses existing datasets in dialogue quality and diversity. To promote reproducible research, the dataset and generation code are publicly released and integrated into the CRSLab framework. This article is a discussion paper for our work accepted to the Reproducibility Track at *ACM RecSys 2025*; the code and resources can be found in [12].

Keywords

Conversational Recommendation, Large Language Models, Datasets

1. Introduction

Conversational Recommender Systems (CRSs) enhance user engagement by providing personalized recommendations through multi-turn dialogue [6]. The development of end-to-end CRSs [9] has intensified the need for high-quality conversational datasets. However, existing corpora present notable challenges.Human-generated datasets, such as ReDial [9], often result from role-playing exercises that lack genuine user intent, leading to superficial conversations [8, 2]. While recent methods using Large Language Models (LLMs) for synthetic data generation [4, 8] have improved coherence, they introduce two primary issues. First, they typically depend on proprietary LLMs, which limits accessibility and reproducibility. Second, they model users as a homogeneous group, failing to capture the behavioral diversity observed in real-world interactions where users may have varying levels of preference clarity and initiative [1, 14]. To overcome these gaps, we present DistillRecDial, a conversational recommendation dataset designed around explicit user personas. Our contributions are:

1. A dataset that models heterogeneous user behaviors by defining five distinct user stereotypes based on preference and intention clarity.
2. A Large-to-Small knowledge distillation pipeline that enables scalable, high-quality dialogue generation using open-source models, ensuring reproducibility.

IIR2025: 15th Italian Information Retrieval Workshop, 3th - 5th September 2025, Cagliari, Italy

[*] Corresponding author.

[†] These authors contributed equally.

alessandro.martina@uniba.it (A. F. M. Martina); alessandro.petruzzelli@uniba.it (A. Petruzzelli); cataldo.musto@uniba.it (C. Musto); marco.degemmis@uniba.it (M. d. Gemmis); pasquale.lops@uniba.it (P. Lops); giovanni.semeraro@uniba.it (G. Semeraro)

0009-0003-9609-1583 (A. F. M. Martina); 0009-0008-2880-6715 (A. Petruzzelli); 0000-0001-6089-928X (C. Musto); 0000-0002-2007-9559 (M. d. Gemmis); 0000-0002-6866-9451 (P. Lops); 0000-0001-6883-1853 (G. Semeraro)

3. A comprehensive evaluation demonstrating that DistillRecDial exhibits superior dialogue quality and diversity compared to established datasets.
4. The public release of the dataset, generation code, and its integration into the CRSLab benchmark [15].

2. Related Works

Early CRS datasets like ReDial [9], GoRecDial [7], and INSPIRED [3] were generated by human annotators. While foundational, these datasets are difficult to scale and can have inconsistent quality.

Synthetic data generation using LLMs has emerged as a scalable alternative. Initial efforts converted single-turn interactions to dialogues [4, 11], but often inherited the limitations of the source data. More recent work, such as PEARL [8] and LLM-REDIAL [10], employed GPT-3.5 as a user simulator. These methods produce more coherent dialogues but rely on closed-source APIs, posing challenges for reproducibility. Furthermore, they do not explicitly model the diversity of user interaction patterns.

Our work differs by systematically modeling user variation through stereotypes. We utilize a knowledge distillation pipeline [5], transferring the capabilities of a large "teacher" model to a smaller, open-source "student" model. This approach allows for the creation of a high-quality, diverse, and reproducible dataset without dependence on proprietary models.

3. Dataset Construction

The generation of DistillRecDial was designed to produce high-quality dialogues reflecting diverse user behaviors. The pipeline involves grounding conversations in real user data, defining user stereotypes to guide dialogue flow, and using knowledge distillation for scalable generation. A summary of dataset statistics is provided in Table 1.

Table 1
Comparative Dataset Statistics.

Metric	ReDial	INSPIRED	PEARL	Ours
# Dialogues	11,347	999	57,277	21,039
Avg. turns	18.16	35.72	9.30	9.89
Avg. tokens/utt.	8.32	9.24	42.47	58.16

3.1. Data Grounding and User Stereotypes

To ground dialogues in realistic preferences, we utilized the Amazon Reviews dataset, focusing on the *Movies and TV* category. After applying a 12-core filter to ensure user profiles were sufficiently dense, the data contained 23,456 users and 15,597 items. We enriched item metadata with information from The Movie Database (TMDB), including genres, keywords, and cast, to enable content-based conversations.

A core feature of our work is the modeling of user heterogeneity. We defined five user stereotypes by combining two behavioral dimensions: *Preference Expression* (None, Implicit, Explicit) and *Intention Clarity* (None, Vague, Specific), based on prior user modeling literature [14]. (1) **Curious Newcomer**: No history, vague goal. (2) **Focused Newcomer**: No history, specific goal. (3) **History-Based Browser**: Has history, no specific goal. (4) **Guided Explorer**: Has history, seeks novelty. (5) **Preference-Driven Searcher**: Has history, expresses explicit and complex preferences.

3.2. Prompt Design and Generation Pipeline

For each of the 23,456 users, we randomly assigned one of the five stereotypes. A structured prompt was created containing: (1) a description of the user's persona and goal, (2) a target item to be recommended, and (3) the user's interaction history (if applicable), with features conditioned on the assigned stereotype (Table 2).

Table 2
Feature Inclusion by User Stereotype. History-based features are aggregated from the user's past interactions.

Feature	Curious Newc.	Focused Newc.	History-Browser	Guided Explorer	Preference-Driven
Genres/Keywords	Target	Target	History	Both	Both
Actors/Directors	✗	Target	✗	✗	Both
Liked/disliked	✗	✗	✓	✓	✓

We employed a Large-to-Small knowledge distillation pipeline for dialogue generation:

1. **Teacher Generation**: A teacher model (LLaMA 3.3 70B) generated high-quality dialogues for 10% of the prompts. These dialogues served as exemplars.
2. **Knowledge Distillation**: A smaller student model (LLaMA 3.1 8B) was fine-tuned on the teacher-generated dialogues. This offline distillation transfers the teacher's stylistic and structural capabilities to the student [13].
3. **Scalable Generation**: The fine-tuned student model generated dialogues for all 23,456 prompts. After filtering for hallucinations, the final dataset contains 21,039 high-quality dialogues. Example snippets are shown in Table 3.

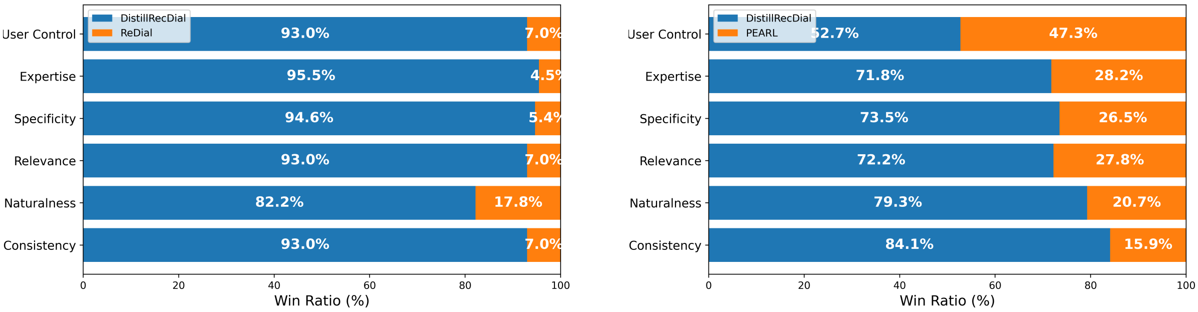
4. Evaluation

We evaluated DistillRecDial to address three research questions regarding dialogue quality (RQ1), diversity (RQ2), and downstream task performance (RQ3). We compared our dataset with ReDial, INSPIRED, and PEARL.

4.1. Human Evaluation of Dialogue Quality (RQ1)

We conducted a head-to-head human evaluation where 10 evaluators compared 100 dialogues from DistillRecDial against 100 from each baseline. As shown in Figure 1, DistillRecDial was consistently preferred across all criteria, including naturalness, relevance, and consistency. It outperformed human-generated datasets (ReDial, INSPIRED) by a wide margin and also demonstrated higher quality than the LLM-generated PEARL dataset.

Figure 1: Head-to-head human evaluation results.



(a) DistillRecDial vs. ReDial

(b) DistillRecDial vs. PEARL

4.2. Automated Evaluation of Dialogue Diversity (RQ2)

To assess diversity, we computed the average pairwise cosine similarity of utterances at each turn for DistillRecDial and PEARL. As illustrated in Figure 2, DistillRecDial exhibits significantly lower

Table 3

Partial dialogue snippets from DistillRecDial, demonstrating various user stereotypes.

User Stereotype	Dialogue Snippet
Curious Newcomer	**User**: "I'm in the mood for something that's going to make me feel uneasy, you know? Something that's a little dark and maybe even a bit scary. I don't know, maybe something with a twist?"
Preference-Driven Searcher	**User**: "Hi, I'm looking for a movie that has a similar tone to Mean Girls and The Wedding Date, with a mix of adventure, comedy, and a touch of romance. I enjoy movies that have complex characters and unexpected plot twists."

turn-level similarity (0.422) compared to PEARL (0.598), indicating greater linguistic diversity. PEARL's high similarity in early turns is due to repetitive openings, whereas our stereotype-driven approach generates more varied initial user requests.

4.3. Downstream Task Performance (RQ3)

We benchmarked standard CRS models on DistillRecDial using the CRSLab framework.

Recommendation Performance: As shown in Table 4, BERT, which leverages rich textual context, achieved the strongest recommendation performance (Hit@10 = 0.1728). Sequential models like SASRec performed less favorably, suggesting that the context-rich dialogues in our dataset are well exploited by pre-trained language models. Integrated CRS models like KBRD, INSPIRED, and ReDial struggled, indicating that they do not generalize well to the complex conversational structures present in DistillRecDial. This highlights the challenge of jointly optimizing for recommendation and dialogue on realistic, user-aware data and points to opportunities for developing more advanced CRS architectures.

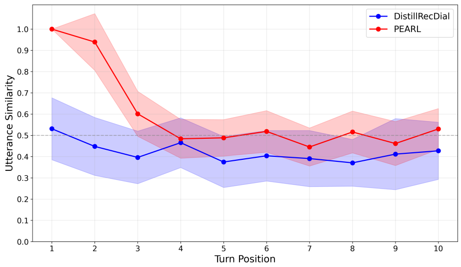


Figure 2: Average pairwise cosine similarity of utterances at each dialogue turn for PEARL and DistillRecDial.

Model	Hit@1	Hit@10	MRR@10	NDCG@10
Dialogue + Recommendation Models				
ReDial	0.0000	0.0075	0.0013	0.0027
INSPIRED	0.0014	0.0090	0.0030	0.0043
KBRD	0.0009	0.0104	0.0026	0.0043
Recommendation-Only Models				
BERT	**0.0763**	**0.1728**	**0.1039**	**0.1202**
SASRec	0.0040	0.0195	0.0077	0.0104

Table 4: Recommendation metrics on DistillRecDial.

5. Conclusion

We introduced DistillRecDial, a large-scale, diverse, and reproducible conversational recommendation dataset. By modeling distinct user stereotypes and employing a knowledge distillation pipeline with open-source models, we address key limitations of prior datasets related to behavioral homogeneity and reliance on proprietary LLMs. Our evaluations confirm that DistillRecDial exhibits higher dialogue quality and diversity. By integrating it into the CRSLab framework, we provide a robust benchmark to spur the development of next-generation CRSs capable of adapting to a wide range of user interaction styles.

Acknowledgement

This research is partially funded by the PNRR project FAIR—Future AI Research (PE00000013), Spoke 6—Symbiotic AI, under the NRRP MUR program supported by NextGenerationEU (CUP H97G22000210007), and the PHaSE project — Promoting Healthy and Sustainable Eating through

Interactive and Explainable AI Methods, funded by MUR under the PRIN 2022 program - Finanziato dall'Unione europea - NextGeneration EU, Missione 4 Componente 1 (CUP H53D23003530006).

The models are developed using the Leonardo supercomputer with the support of CINECA-Italian Supercomputing Resource Allocation, under class C projects IscrC_LLM-REC (HP10CTEUGX) and IscrC_SYMBREC (HP10C1A4P8).

Declaration on Generative AI

During the preparation of this work, the author did not use any AI tool.

References

[1] Wanling Cai, Yucheng Jin, and Li Chen. Impacts of personal characteristics on user trust in conversational recommender systems. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI '22. ACM, 2022.

[2] Shuyu Guo, Shuo Zhang, Weiwei Sun, Pengjie Ren, Zhumin Chen, and Zhaochun Ren. Towards explainable conversational recommender systems. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2786–2795, 2023.

[3] Shirley Anugrah Hayati, Dongyeop Kang, Qingxiaoyang Zhu, Weiyan Shi, and Zhou Yu. Inspired: Toward sociable recommendation dialog systems. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8142–8152, 2020.

[4] Zhankui He, Zhouhang Xie, Rahul Jha, Harald Steck, Dawen Liang, Yesu Feng, Bodhisattwa Prasad Majumder, Nathan Kallus, and Julian McAuley. Large language models as zero-shot conversational recommenders. In *Proceedings of the 32nd ACM international conference on information and knowledge management*, pages 720–730, 2023.

[5] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

[6] Dietmar Jannach, Ahtsham Manzoor, Wanling Cai, and Li Chen. A survey on conversational recommender systems. *ACM Computing Surveys (CSUR)*, 54(5):1–36, 2021.

[7] Dongyeop Kang, Anusha Balakrishnan, Pararth Shah, Paul A Crook, Y-Lan Boureau, and Jason Weston. Recommendation as a communication game: Self-supervised bot-play for goal-oriented dialogue. In *Proceedings of the 2019 Conference on EMNLP-IJCNLP*, pages 1951–1961, 2019.

[8] Minjin Kim, Minju Kim, Hana Kim, Beong-woo Kwak, SeongKu Kang, Youngjae Yu, Jinyoung Yeo, and Dongha Lee. Pearl: A review-driven persona-knowledge grounded conversational recommendation dataset. In *Findings of the ACL 2024*, pages 1105–1120, 2024.

[9] Raymond Li, Samira Ebrahimi Kahou, Hannes Schulz, Vincent Michalski, Laurent Charlin, and Chris Pal. Towards deep conversational recommendations. *Advances in neural information processing systems*, 31, 2018.

[10] Tingting Liang, Chenxin Jin, Lingzhi Wang, Wenqi Fan, Congying Xia, Kai Chen, and Yuyu Yin. Llm-redial: A large-scale dataset for conversational recommender systems created from user behaviors with llms. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 8926–8939, 2024.

[11] Yu Lu, Junwei Bao, Zichen Ma, Xiaoguang Han, Youzheng Wu, Shuguang Cui, and Xiaodong He. August: an automatic generation understudy for synthesizing conversational recommendation datasets. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 10538–10549, 2023.

[12] Alessandro Francesco Maria Martina, Alessandro Petruzzelli, Cataldo Musto, Marco de Gemmis, Pasquale Lops, and Giovanni Semeraro. DistillRecDial: A knowledge-distilled dataset capturing user diversity in conversational recommendation. In *Proceedings of the Nineteenth ACM Conference on Recommender Systems*, RecSys '25. ACM, September 2025.

[13] Anup Shirgaonkar, Nikhil Pandey, Nazmiye Ceren Abay, Tolga Aktas, and Vijay Aski. Knowledge distillation using frontier open-source llms: Generalizability and the role of synthetic data. *arXiv preprint arXiv:2410.18588*, 2024.

[14] Junjie Zhang, Ruobing Xie, Yupeng Hou, Xin Zhao, Leyu Lin, and Ji-Rong Wen. Recommendation as instruction following: A large language model empowered recommendation approach. *ACM Trans. Inf. Syst.*, December 2024.

[15] Kun Zhou, Xiaolei Wang, Yuanhang Zhou, Chenzhan Shang, Yuan Cheng, Wayne Xin Zhao, Yaliang Li, and Ji-Rong Wen. Crslab: An open-source toolkit for building conversational recommender system. *arXiv preprint arXiv:2101.00939*, 2021.