MEMARD: A Conversational Recommendation Dataset via Memory-Enhanced Multi-Agent Dialogue Synthesis

Anonymous ACL submission

Abstract

The modeling and evaluation of conversational recommender systems (CRSs) depend critically on high-quality multi-turn dialogue datasets. Yet crowdsourced datasets often incur the substantial cost of large-scale annotation. Recent advances in LLM agents enable low-cost synthesis of CRS datasets through profile-driven role-play and coordinated multi-turn interaction between user and system agents. However, existing synthetic datasets still compress user modeling into flat textual personas and lack turn-level quality control during generation. We present MEMARD[1] (Memory-Enhanced Multi-Agent Recommendation Dialogue), a synthetic conversational recommendation dataset of 50,166 dialogues built from Amazon reviews. We build a multi-agent framework to synthesize multi-turn dialogues with turn-level supervision during generation. From real review histories, we dynamically track users' fine-grained preferences and model dialogue style from user reviews; by equipping user agents with such fine-grained and comprehensive memory, our method enables more realistic user simulation and yields specific user preferences, metadata-grounded recommendations, and diverse inter-user behavior. Experiments with human and LLM dialogue quality evaluation and recommendation ranking demonstrate effectiveness. MEMARD provides high-fidelity synthetic data for benchmarking and training conversational recommender systems.

1 Introduction

With the rapid rise of large language models (LLMs), the recommendation paradigm is gradually shifting from collaborative filtering toward interactive recommendation assistants. In conversational recommender systems (CRSs), the system acts as a recommendation assistant that converses with users to clarify needs and explain suggestions, rather than only ranking items. Unlike traditional recommendation, which infers preferences largely from implicit behavioral signals, CRS requires the assistant to interpret user needs from natural language. Progress in both modeling and evaluation therefore depends critically on high-quality multi-turn dialogue datasets that supply realistic preference expression, grounded recommendations, and scalable supervision. Yet many existing CRS datasets suffer from vague user preferences and noisy dialogues, highlighting the urgent need for better datasets.

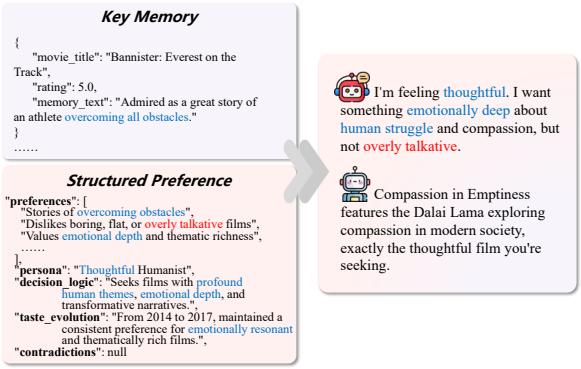

Figure 1: An illustrative MEMARD dialogue excerpt, where seeker behavior is based on key memories and structured preferences.

High-quality conversational recommendation datasets are a key bottleneck for developing personalized CRSs. Existing CRS research still relies heavily on manually collected datasets such as ReDial (Li et al., 2018) and INSPIRED (Hayati et al., 2020). However, crowdsourced CRS datasets more generally (Zhou et al., 2020b; Liu et al., 2020, 2021) are constructed from dialogues between two human participants rather than interactions between a user and a recommendation assistant. As a result, dialogue samples are of-

[1] https://anonymous.4open.science/r/MEMARD

Table 1: Dataset statistics across MEMARD and comparison datasets (Kim et al., 2024; Martina et al., 2025; Li et al., 2018; Hayati et al., 2020). Following Kim et al. (2024), 4-gram specificity is the average number of unique 4-grams in seeker utterances per dialogue; higher values indicate less generic preference language.

Dataset	# of dialogues	# of utterances	# of words	4-gram specificity	Memory-enhanced user simulation	Turn-level quality review
ReDial	11,348	205,670	1,376k	59.3	✗	✗
INSPIRED	999	21,122	283k	122.9	✗	✗
PEARL	57,277	532,923	18,725k	153.4	✗	✗
DistillRecDial	21,039	208,082	10,283k	164.4	✗	✗
MEMARD	**50,166**	**480,433**	**27,414k**	**332.7**	✓	✓

ten segmented whenever products are mentioned, even though such mentions may not faithfully reflect the user's true preferences. This weak grounding between product mentions and user interests limits their effectiveness for training personalized CRSs.

To reduce the cost of manual data collection, recent studies have begun to synthesize CRS dialogues with agents (Kim et al., 2024; Liang et al., 2024; Martina et al., 2025; He et al., 2023), where user-simulation agents interact with recommender agents to transform existing recommendation data into conversational form. This agent-based paradigm improves scalability, but existing pipelines still suffer from two major limitations. First, **user modeling remains insufficient**: existing methods usually represent users with fixed role settings or coarse preference descriptions, making generated user behaviors relatively homogeneous and unable to reflect the dynamic preferences and diverse behavioral patterns of real-world users. Second, **fine-grained process control remains underdeveloped**: since generation is primarily LLM-driven, it often produces seeker–recommender conversations that appear plausible on the surface, while the ways in which user intents are revealed, preferences are clarified, and historical interactions support each conversational decision remain insufficiently realistic.

Accordingly, this work targets two goals, namely finer-grained, diverse simulation of real user behavioral patterns and turn-level generation quality control, and proposes MEMARD (Memory-Enhanced Multi-Agent Recommendation Dialogue). We build a memory-enhanced user simulation pipeline that distills memories from real behavioral traces to ground preferences in concrete interaction evidence and support diverse, realistic seeker expression. Figure 1 shows a MEMARD dialogue excerpt with seeker behavior based on key memories and structured preferences. On this basis, we design a multi-agent dialogue framework comprising a user agent, a system agent, and a controller agent. We equip the user agent with the extracted, refined real-world memories and semantic retrieval over the memory store, couple the system agent with structured metadata retrieval over an item catalog, and decouple each role into a planner–executor structure. We further add an iterative review module that performs multi-dimensional binary checks on user and system outputs and drives retries with structured feedback at each turn. This module suppresses low-quality generations.

We instantiate the data layer on Amazon reviews (Ni et al., 2019) and conduct comprehensive evaluations. Human and LLM dialogue quality evaluation against multiple baselines verifies improvements in user behavior and recommendation semantics as well as the effect of turn-level review; recommendation ranking and multi-turn dialogue experiments demonstrate ranking performance and behavioral diversity. Results support MEMARD as a behaviorally diverse, quality-controlled synthetic training resource.

Table 1 summarizes how MEMARD compares with representative CRS datasets in scale and key construction features.

2 Related Work

2.1 CRS Datasets

CRSs aim to understand user preferences through multi-turn natural language interaction and provide personalized recommendations, and high-quality annotated datasets remain central infrastructure. Early datasets were built via crowdsourcing: ReDial (Li et al., 2018) provides over 10,000 movie dialogues (averaging ∼18 turns) but

users often give generic preferences; INSPIRED (Hayati et al., 2020) labels recommender strategies from social science but scales to only about 1,000 dialogues; and TG-ReDial (Zhou et al., 2020b) extends ReDial with topic tags and subtasks while relying on predefined topics and subjective paths. Collectively, these datasets offer indispensable benchmarks but often trade preference specificity, recommendation expertise, and scale for annotation cost, motivating more efficient synthetic construction.

As large language models advance, automatic CRS dialogue generation has become a prominent alternative to crowdsourcing: carefully designed prompts let an LLM alternate as recommender and seeker, producing large datasets at far lower marginal cost. He et al. (He et al., 2023) construct large-scale movie recommendation conversations from Reddit but crawled thread quality can be uneven and user preference modeling remains limited. AUGUST (Lu et al., 2023) uses data-to-text generation over user–item interaction graphs before dialogue synthesis, reducing annotation cost but inheriting biases of the CRS datasets used to train the generator. PEARL (Kim et al., 2024) builds review-derived personas and item knowledge with a dual-simulator to produce over 57k dialogues, reporting gains over crowdsourced sets on preference specificity and recommendation relevance, although seeker-side user modeling remains relatively homogeneous. LLM-REDIAL (Liang et al., 2024) further scales LLM-generated CRS dialogues across domains, and controllable in-context simulation (Li et al., 2022) offers steerable synthesis with light human annotation. To diversify seeker behavior beyond flat personas, DistillRecDial (Martina et al., 2025) applies knowledge distillation and multiple user prototypes with tailored prompts so dialogues exhibit richer seeker behavior than PEARL. Such methods nonetheless leave user behavior coarse-grained at the memory, expressive, and interaction-logic levels, and most pipelines lack turn-level quality control, allowing low-quality turns to enter the dataset uncorrected.

2.2 Multi-Agent Simulation and Synthesis

Recent work also explores multi-agent simulation and synthesis for recommendation-related settings (Chen et al., 2025b). MUSE (Wang et al., 2025b) automatically constructs a multimodal conversational recommendation dataset using a multi-agent framework powered by multimodal LLMs, deriving user profiles from real-world scenarios and optimizing synthesized conversations in the clothing domain. Generative agents (Park et al., 2023) model believable behavior through streaming memory, reflection, and planning in open-ended social simulation. RecAgent (Wang et al., 2025a) simulates recommendation-related user behaviors with profiling, memory, and action modules in a large-scale agent sandbox. RecUserSim (Chen et al., 2025a) offers a CRS-oriented user simulator with profile, memory, action, and refinement modules for realistic, diverse evaluation of conversational recommenders. Motivated by these developments, our work focuses on large-scale movie CRS dataset construction, combining memory-enhanced user simulation from reviews, a metadata-grounded system agent, and a user/system/controller stack with turn-level review when writing training dialogues.

3 Dataset Construction

As illustrated in Figure 2, we propose MEMARD, a multi-agent framework for memory-enhanced conversational recommendation dataset construction. Profile-driven synthesis often relies on static personas and lacks turn-level quality control, limiting grounded preference expression during generation. MEMARD counters this with review-derived memories, structured preferences, and dialogue style, together with multi-agent per-turn review. Within the Conversation Generator, a user agent and a system agent alternate to generate seeker and recommender turns. A controller agent in Turn Level Control supervises this multi-agent loop via Trajectory Judge and Acceptance Judge, triggering regeneration when a turn fails.

3.1 Data Sources and Preprocessing

The data layer consists of a user review dataset and an item metadata store. Reviews are drawn from the "Movies and TV" category of the Amazon Reviews dataset[2] (Ni et al., 2019). After retaining entries alignable with TMDB, each record contains a user id, a movie id, a star rating (1–5), and review text. We then match and filter records by interaction scale and metadata completeness; the resulting dataset contains 40,676 items, 53,494 users, and 959,983 ratings in total. Item-side data come from TMDB[3] (The Movie Database, 2024), in-

[2] https://nijianmo.github.io/amazon/index.html
[3] https://api.themoviedb.org

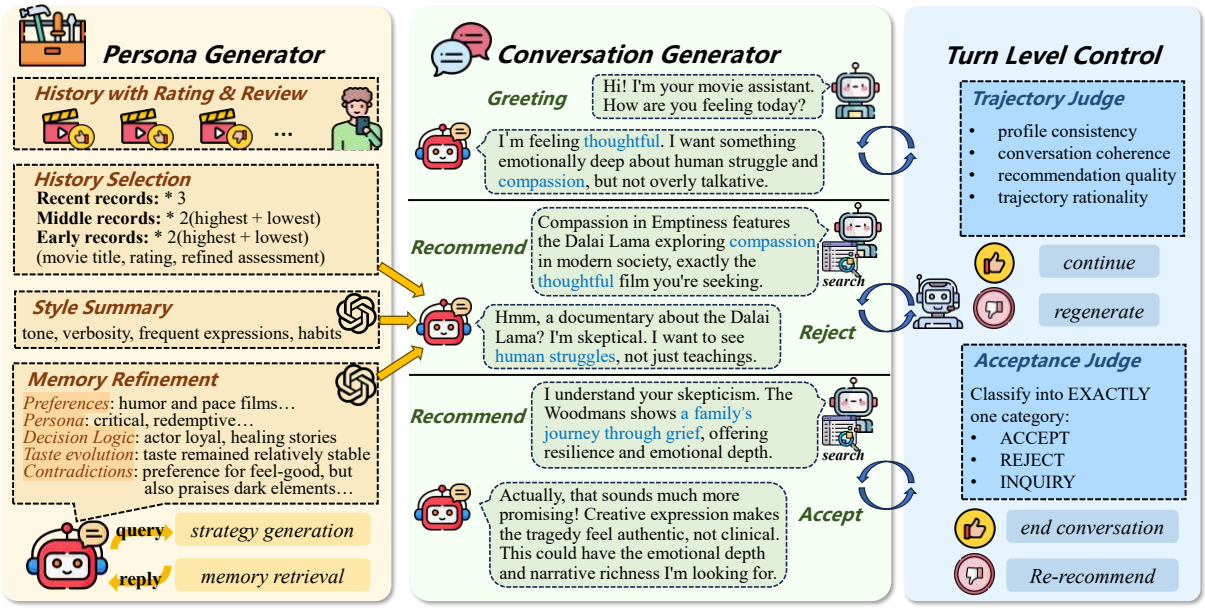

Figure 2: Overview of MEMARD. The Persona Generator constructs seeker profiles from review histories through History Selection, Style Summary, and Memory Refinement. The Conversation Generator alternates a user agent and a system agent to synthesize seeker and recommender turns. Turn Level Control deploys a controller agent with Trajectory Judge and Acceptance Judge to supervise the multi-agent loop and classify user intent.

cluding structured fields such as title, genres, keywords, cast and crew, and plot summaries, providing retrievable domain knowledge for the system agent during dialogue synthesis.

3.2 Multi-Agent Dataset Construction

If a single model alternates between seeker and recommender, it is difficult to jointly honor user-specific evidence and item-side knowledge. Explicit turn-level control over dialogue progression and quality is also difficult to maintain. We address this with a multi-agent design comprising a user agent, a system agent, and a controller agent, each responsible for a distinct conversational function. Within each agent, we further use a decoupled planner–executor structure so that high-level reasoning is separated from retrieval: the planner reasons over dialogue context and issues structured instructions, while the executor performs retrieval and returns evidence for the final utterance. Prompt templates are provided in Appendix B.

3.2.1 Persona Generator

To better reflect inter-user differences in preference structure and interaction expression, we adopt a memory-enhanced user simulation design. Unlike flat textual profiles, memories here are compact, retrievable units distilled from observed reviews. Each comprises a title, rating, and refined assessment, grounding what the user agent can say during dialogue in concrete behavioral evidence rather than generic claims; the user agent may query them on demand when disclosing preferences, challenging recommendations, or justifying accept/reject decisions. We further induce structured preferences and dialogue style atop these memories to capture taste evolution, contradictions, and heterogeneous expression. The pipeline comprises three stages: history selection, style summary, and memory refinement. Staged sampling selects representative comments across a user's timeline to form comparable key memories and structured fields such as contradictory preferences. This addresses limitations of flat personas or fixed stereotypes when modeling individualized, contextualized, and sometimes inconsistent users.

History Selection. For each user, we sort interactions chronologically and divide them into early, middle, and recent phases. We retain the three most recent interactions from the recent phase to reflect current interests. From both the middle and early phases, we select one highest-rated and one lowest-rated interaction, capturing stable preference boundaries in the middle stage and initial interests and trajectories in the early stage. This yields 7 core interactions per user, spanning preference breadth and supplying multi-

stage, comparable evidence. An LLM structurally parses the sampled reviews into key memory units, each a triple: (movie title, rating, refined subjective/objective assessment), compressing redundancy and emotional noise.

Style Summary. We further induce stylistic attributes from raw reviews, including tone, verbosity, frequent expressions, and habits, so surface language features differ across users rather than collapsing to a homogeneous template.

Memory Refinement. Building on key memories, structured preferences are summarized by an LLM as (1) viewing preference tags; (2) aesthetic preference lists; (3) decision logic; (4) taste evolution; and (5) contradictions among coexisting preferences that appear conflicting yet realistic, to constrain user agent consistency and interpretability.

3.2.2 Conversation Generator

In the Conversation Generator, a user agent and a system agent alternate to produce multi-turn dialogues. Each session typically begins with a system greeting; the user agent then expresses preferences, and the system agent responds with a metadata-grounded suggestion conditioned on the dialogue history and revises its recommendation based on user replies.

The user agent simulates the seeker and remains consistent with this memory-enhanced user simulation. Static profiles often yield generic preference talk; to better approximate real users, the agent retrieves relevant memories on demand from a pre-built FAISS-based user-memory index (Johnson et al., 2021), so that disclosures, challenges, and accept/reject decisions can be supported by concrete interaction evidence rather than improvisation alone. Turn policy is also staged to reflect common CRS interaction patterns: early turns emphasize broad preferences, mid-dialogue turns are more likely to hesitate or object once recommendations appear, and later turns bias toward acceptance after sufficient clarification.

The system agent plays the domain-knowledgeable recommender. Its goal is to pair fluent replies with recommendations grounded in retrieved item metadata rather than unconstrained LLM generation alone. Given the user utterance and persona fields, the agent retrieves candidate movies from the metadata index in Section 3.1, excludes titles already mentioned in the dialogue or incompatible with the persona, and selects the best-matching remaining candidate for recommendation. Generation further follows explicit dialogue-format constraints, including single-item recommendation per turn and avoidance of list-like templates and duplicate suggestions.

3.2.3 Turn Level Control

Multi-turn synthesis requires explicit control over scheduling, stopping, and repair. The controller agent coordinates dialogue generation and enforces turn-level quality through an integrated review module.

User Reply Review. The module examines each user reply along three aspects: (1) persona match, covering tone, persona tags, verbosity, and aesthetic lists; (2) coherence, checking response to the latest system turn and consistency with key memories and prior stance; and (3) naturalness, flagging topic jumps or abnormal repetition. Each aspect is implemented as an LLM binary judgment with an actionable failure description.

System Reply Review. The module examines each system reply along three aspects: (1) format compliance, disallowing list styles or title-like templates; (2) recommendation quality, which applies when the turn is a recommendation reply and checks alignment with user needs, avoidance of disliked types, and single-item recommendation only; and (3) naturalness, using the same criteria as for the user agent.

The controller initializes each dialogue, alternates user-agent and system-agent turns, classifies user intent (accept, reject, or question), and terminates when the user accepts or a turn budget is reached. After each generated turn, it invokes the review module; if a turn fails, the controller passes structured feedback to the corresponding user agent or system agent to regenerate the reply before the utterance is written into the dataset. If any review dimension fails, the module produces structured feedback (issue description and targeted reminders) and the controller returns it with the original prompt to the agent for retry. If all retries are exhausted, the last output is kept as a fallback to avoid dialogue breakage. This generate–review–feedback–correct loop runs every turn to suppress low-quality dialogue early.

4 Experiments

We evaluate MEMARD through three research questions on dialogue quality, recommendation utility, and behavioral diversity.

Figure 3: Pairwise win ratios against INSPIRED, PEARL, and DistillRecDial (left to right). Top row: human and LLM evaluation on matched samples (n=200). Bottom row: LLM-as-judge on 1,000 samples.

- **RQ1 (Dialogue Quality):** Does MEMARD yield higher-quality CRS datasets than existing datasets under dialogue quality evaluation (Section 4.1), and does turn-level review contribute to these gains?
- **RQ2 (Recommendation Utility):** Can MEMARD support the evaluation of CRS models on conversational recommendation?
- **RQ3 (Behavioral Diversity):** Does MEMARD exhibit greater user behavioral diversity than PEARL and DistillRecDial?

We address RQ1 with dialogue quality evaluation against three baselines and ablations (w/o Review and w/o Memory); RQ2 with recommendation experiments on multiple CRSLab backbones trained on MEMARD; and RQ3 with Distinct-n analyses for dialogue models trained on MEMARD, PEARL, and DistillRecDial under matched settings, together with dataset-level semantic diversity metrics on the same three datasets.

We compare against three representative datasets: INSPIRED (Hayati et al., 2020), a widely used English crowdsourced CRS benchmark with strategy labels; PEARL (Kim et al., 2024), a dual-simulator synthetic CRS dataset with personas and item knowledge; and DistillRecDial (Martina et al., 2025), which uses teacher–student distillation and LLM generation to emphasize behavioral heterogeneity.

Table 2: Ablation under the same LLM-as-judge setup (win ratio, %). Each row compares MEMARD against an ablated variant on 1,000 dialogues.

	w/o Review		w/o Memory	
Dimension	MEMARD	w/o Rev.	MEMARD	w/o Mem.
User Ctrl.	99.1	0.9	99.2	0.8
Expertise	55.7	44.3	70.2	29.8
Specificity	95.1	4.9	92.3	7.7
Relevance	86.7	13.3	72.0	28.0
Flow Nat.	92.8	7.2	93.2	6.8
Consistency	92.7	7.3	92.5	7.5

4.1 Dialogue Quality Evaluation

To evaluate dialogue quality, we combine human annotation with LLM-as-judge in pairwise comparisons between MEMARD and existing datasets. We follow PEARL (Kim et al., 2024) and adopt its six-dimensional rubric for assessing synthetic CRS datasets, jointly covering seeker-side engagement and preference expression, recommender-side domain knowledge and recommendation fit, and dialogue-level fluency and coherence. We apply the same six dimensions and pairwise judging setup to both evaluation tracks so that MEMARD can be compared against PEARL and other baselines rather than ad hoc criteria. In each comparison, the evaluator selects a winner on one dimension at a time and provides a brief rationale.

The six dimensions are **User Control** (the extent to which the seeker actively participates and guides the recommendation process through

Table 3: Recommendation metrics for multiple models. Hit@K, NDCG@K, and MRR@K from CRSLab.

Model	Hit@1	Hit@10	Hit@50	NDCG@10	NDCG@50	MRR@10	MRR@50
TGReDial	0.1361	0.3809	0.5602	0.2477	0.2878	0.2063	0.2152
BERT	0.1051	0.3121	0.4776	0.1990	0.2358	0.1639	0.1719
INSPIRED	0.0154	0.1090	0.2295	0.0555	0.0817	0.0392	0.0446
SASRec	0.0619	0.1633	0.2885	0.1065	0.1340	0.0891	0.0949
KGSF	0.0362	0.1633	0.3399	0.0927	0.1314	0.0709	0.0791
KBRD	0.0165	0.0773	0.1643	0.0422	0.0613	0.0316	0.0357
GRU4Rec	0.0006	0.0038	0.0085	0.0020	0.0030	0.0014	0.0016
ReDial	0.0002	0.0014	0.0075	0.0006	0.0019	0.0004	0.0006

feedback), **Expertise** (the recommender's professionalism and domain knowledge in movies), **Specificity** (the richness and detail of seeker-stated preferences for personalized recommendations), **Relevance** (alignment of recommendations with explicitly stated preferences), **Flow Naturalness** (how natural and smooth turn-taking, transitions, and overall dialogue progression appear), and **Consistency** (coherence of seeker preferences throughout the dialogue).

These criteria guide both human annotation and LLM-as-judge evaluation. For human evaluation, we randomly sample 200 matched dialogue pairs per pairing against INSPIRED, PEARL, and DistillRecDial and recruit three human judges per pair to conduct blind, dimension-wise comparisons; win ratios are aggregated by majority vote across judges. On the same matched samples, we additionally run LLM-as-judge with dialogue order randomized. We further randomly sample 1,000 dialogue pairs per pairing for LLM-as-judge at larger scale and evaluate them with gpt-3.5-turbo-0125 (OpenAI, 2022) at temperature 0.1 (Appendix A) using a unified judge prompt applied consistently across all datasets.

The pairwise comparison results against INSPIRED (Hayati et al., 2020), PEARL (Kim et al., 2024), and DistillRecDial (Martina et al., 2025) are presented in Figure 3; the top row reports human and LLM judgments on 200 matched samples, and the bottom row reports LLM-as-judge results on 1,000 samples. Across most dimensions, including user control, specificity, consistency, expertise, and relevance, MEMARD demonstrates clear advantages, which we attribute to its memory-enhanced user simulation design, retrieval-based evidence integration, and fine-grained turn-level quality control.

Ablation Analysis. To assess turn-level review and the persona generator, we compare MEMARD against w/o Review (review and retry disabled) and w/o Memory (persona generator disabled), holding other generation procedures constant. Judging follows the same six-dimensional setup as above. Table 2 shows that removing review or persona generation consistently lowers MEMARD win ratios across dimensions, supporting the importance of both turn-level iterative review and the persona generator pipeline.

4.2 Recommendation Results and Analysis

Recommendation experiments test whether MEMARD provides strong ranking supervision for mainstream models. We evaluate eight models in CRSLab (Zhou et al., 2021): **TGReDial** (Zhou et al., 2020b) for Transformer-based joint recommendation–dialogue modeling, **BERT** (Devlin et al., 2019) for contextual recommendation, the **INSPIRED** architecture (Hayati et al., 2020), **SASRec** (Kang and McAuley, 2018) and **GRU4Rec** (Hidasi et al., 2016) for sequential recommendation, **KBRD** (Chen et al., 2019) and **KGSF** (Zhou et al., 2020a) for knowledge-graph–augmented CRS, and **ReDial** (Li et al., 2018) as a dialogue-generation baseline.

As shown in Table 3, TGReDial and BERT achieve the highest performance, indicating that MEMARD provides effective support for both topic-enhanced end-to-end architectures and models emphasizing rich contextual encoding for recommendation tasks. This aligns well with the design choices of MEMARD regarding structured metadata retrieval, explicit recommendation formatting, and turn-level review—dialogue states retain strong, discriminative item-level cues. INSPIRED, SASRec, KGSF, and KBRD perform at

Table 4: Conversational metrics for multiple models. Distinct-n for n=1–4 from CRSLab.

Model	Train	D-1	D-2	D-3	D-4
INSPIRED	MEMARD	0.0137	0.2008	0.4655	0.6442
INSPIRED	PEARL	0.0154	0.1065	0.2505	0.3957
INSPIRED	DistillRecDial	0.0118	0.1144	0.2891	0.4435
TGReDial	MEMARD	0.0134	0.2043	0.4732	0.6529
TGReDial	PEARL	0.0152	0.1111	0.2667	0.4169
TGReDial	DistillRecDial	0.0116	0.1181	0.2994	0.4587
ReDial	MEMARD	0.0020	0.0141	0.0346	0.0596
ReDial	PEARL	0.0004	0.0021	0.0059	0.0126
ReDial	DistillRecDial	0.0011	0.0095	0.0324	0.0692
KBRD	MEMARD	0.0024	0.0154	0.0472	0.0979
KBRD	PEARL	0.0027	0.0121	0.0313	0.0631
KBRD	DistillRecDial	0.0022	0.0145	0.0499	0.1067
KGSF	MEMARD	0.0046	0.0323	0.0913	0.1712
KGSF	PEARL	0.0030	0.0157	0.0450	0.0908
KGSF	DistillRecDial	0.0032	0.0248	0.0824	0.1651

a mid-level: the dataset supplies useful ranking supervision, though these models do not fully match TGReDial/BERT. In contrast, the weaker results from ReDial and GRU4Rec primarily indicate that these models find it challenging to extract and utilize ranking signals from MEMARD in the presence of more diverse user behaviors.

4.3 Dialogue Results and Analysis

Dialogue experiments test whether MEMARD can serve as training data for mainstream dialogue generators, focusing on user behavioral diversity. We train multiple CRSLab dialogue backbones on MEMARD, PEARL (Kim et al., 2024), and DistillRecDial (Martina et al., 2025) under matched automatic evaluation settings. We additionally report user-side semantic diversity using $\mathrm{SemDiv}_{\mathrm{pair}}$ (average $1 - \cos$ distance on all-MiniLM-L6-v2 embeddings; 5,000-utterance samples per dataset) and cluster entropy H (K=50 K-means on pooled embeddings across datasets); higher values indicate greater semantic dispersion and broader coverage of preferences.

Table 4 reports Distinct-n under the same training setup; MEMARD achieves the best overall diversity across architectures. Table 5 ranks MEMARD highest on both semantic metrics, indicating greater user-side semantic diversity beyond lexical variation, consistent with the memory-enhanced user simulation and turn-level review in Section 3. MEMARD preserves greater inter-user surface variation on the user side, reduces templated repetition, and encourages more dispersed n-gram distributions, supporting learning of diverse user behaviors at the distribution level.

Table 5: User-side semantic diversity on MEMARD, PEARL, and DistillRecDial.

Dataset	$\mathrm{SemDiv}_{\mathrm{pair}}$	Cluster entropy H
MEMARD	0.676	3.41
DistillRecDial	0.632	3.25
PEARL	0.497	2.97

We conduct a case study to illustrate the diversity of user behavioral patterns in the dataset, sampling six MEMARD dialogues spanning (a) direct acceptance, (b) hesitant acceptance, (c) multi-turn rejection and recovery with feedback, (d) single rejection then acceptance, (e) multi-turn constraint refinement, and (f) extended hesitation before acceptance; full transcripts are in Appendix C.

5 Conclusion

We address flat user representations, homogeneous user behavior, and insufficient fine-grained constraints in multi-turn synthetic CRS datasets with MEMARD, unifying memory-enhanced user simulation, decoupled multi-agent generation, and turn-level iterative review in a single pipeline, instantiated on Amazon reviews to generate a CRS dataset from movie reviews. Experiments show that MEMARD is ahead on most dimensions in human and LLM dialogue quality evaluation, with clear gains from turn-level review, strong ranking supervision, and greater user behavioral diversity.

These results suggest that injecting key memories and structured preferences on the user side, together with retrieval-based evidence and turn-level quality gates, helps produce heterogeneous yet cross-turn coherent user behaviors in synthetic datasets, mitigating vague preferences and behavioral homogeneity. Future work may extend MEMARD to additional domains and modalities, and further strengthen turn connectivity and naturalness modeling to improve the reliability and practical usability of synthetic CRS datasets.

Limitations

Synthetic data can inherit biases present in source reviews and metadata. Turn-level gating improves

consistency but does not fully reproduce the pragmatic naturalness of real chats: simulated turns can remain comparatively explicit, evenly paced, or lexically repetitive when simulation constraints and review anchors over-constrain surface form.

Acknowledgments

Omitted for anonymous review.

References

Luyu Chen, Quanyu Dai, Zeyu Zhang, Xueyang Feng, Mingyu Zhang, Pengcheng Tang, Xu Chen, Yue Zhu, and Zhenhua Dong. 2025a. RecUserSim: A realistic and diverse user simulator for evaluating conversational recommender systems. In *Companion Proceedings of the ACM Web Conference 2025*, pages 133–142. Association for Computing Machinery.

Qibin Chen, Junyang Lin, Yichang Zhang, Ming Ding, Yukuo Cen, Hongxia Yang, and Jie Tang. 2019. Towards knowledge-based recommender dialog system. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1803–1813, Hong Kong, China. Association for Computational Linguistics.

Shuaihang Chen, Yuanxing Liu, Wei Han, Weinan Zhang, and Ting Liu. 2025b. A survey on LLM-based multi-agent system: Recent advances and new frontiers in application. *Preprint*, arXiv:2412.17481.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

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

Zhankui He, Zhouhang Xie, Rahul Jha, Harald Steck, Dawen Liang, Yesu Feng, Bodhisattwa Prasad Majumder, Nathan Kallus, and Julian McAuley. 2023. Large language models as zero-shot conversational recommenders. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pages 720–730, Birmingham, United Kingdom. ACM.

Balázs Hidasi, Alexandros Karatzoglou, Linas Baltrunas, and Domonkos Tikk. 2016. Session-based recommendations with recurrent neural networks. In *Proceedings of the 4th International Conference on Learning Representations, ICLR 2016, Workshop Track*.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2021. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547.

Wang-Cheng Kang and Julian J. McAuley. 2018. Self-attentive sequential recommendation. In *2018 IEEE International Conference on Data Mining (ICDM)*, pages 197–206, Singapore. IEEE Computer Society.

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

Raymond Li, Samira Ebrahimi Kahou, Hannes Schulz, Vincent Michalski, Laurent Charlin, and Chris Pal. 2018. Towards deep conversational recommendations. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3–8, 2018, Montréal, Canada*, pages 9748–9758.

Zekun Li, Wenhu Chen, Shiyang Li, Hong Wang, Jing Qian, and Xifeng Yan. 2022. Controllable dialogue simulation with in-context learning. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4330–4347, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Tingting Liang, Chenxin Jin, Lingzhi Wang, Wenqi Fan, Congying Xia, Kai Chen, and Yuyu Yin. 2024. 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, Bangkok, Thailand. Association for Computational Linguistics.

Zeming Liu, Haifeng Wang, Zheng-Yu Niu, Hua Wu, and Wanxiang Che. 2021. DuRecDial 2.0: A bilingual parallel corpus for conversational recommendation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4335–4347, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Zeming Liu, Haifeng Wang, Zheng-Yu Niu, Hua Wu, Wanxiang Che, and Ting Liu. 2020. Towards conversational recommendation over multi-type dialogs. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*,

pages 1036–1049, Online. Association for Computational Linguistics.

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

Alessandro Francesco Maria Martina, Alessandro Petruzzelli, Cataldo Musto, Marco de Gemmis, Pasquale Lops, and Giovanni Semeraro. 2025. DistillRecDial: A knowledge-distilled dataset capturing user diversity in conversational recommendation. In *Proceedings of the Nineteenth ACM Conference on Recommender Systems*, pages 726–735, Prague, Czech Republic. ACM.

Jianmo Ni, Jiacheng Li, and Julian McAuley. 2019. Justifying recommendations using distantly-labeled reviews and fine-grained aspects. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 188–197, Hong Kong, China. Association for Computational Linguistics.

OpenAI. 2022. Introducing ChatGPT. https://openai.com/index/chatgpt/. OpenAI blog post; accessed 2025.

Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*.

The Movie Database. 2024. The Movie Database (TMDB). https://www.themoviedb.org. Movie and television metadata; accessed 2024.

Lei Wang, Jingsen Zhang, Hao Yang, Zhiyuan Chen, Jiakai Tang, Zeyu Zhang, Xu Chen, Yankai Lin, Hao Sun, Ruihua Song, Xin Zhao, Jun Xu, Zhicheng Dou, Jun Wang, and Ji-Rong Wen. 2025a. User behavior simulation with large language model-based agents. *ACM Transactions on Information Systems*, 43(2):55:1–55:37.

Zihan Wang, Xiaocui Yang, YongKang Liu, Shi Feng, Daling Wang, and Yifei Zhang. 2025b. MUSE: A multimodal conversational recommendation dataset with scenario-grounded user profiles. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 1027–1053, Vienna, Austria. Association for Computational Linguistics.

Kun Zhou, Xiaolei Wang, Yuanhang Zhou, Chenzhan Shang, Yuan Cheng, Wayne Xin Zhao, Yaliang Li, and Ji-Rong Wen. 2021. CRSLab: An open-source toolkit for building conversational recommender system. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 185–193, Online. Association for Computational Linguistics.

Kun Zhou, Wayne Xin Zhao, Shuqing Bian, Yuanhang Zhou, Ji-Rong Wen, and Jingsong Yu. 2020a. Improving conversational recommender systems via knowledge graph based semantic fusion. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1006–1014, Virtual Event, CA, USA. ACM.

Kun Zhou, Yuanhang Zhou, Wayne Xin Zhao, Xiaoke Wang, and Ji-Rong Wen. 2020b. Towards topic-guided conversational recommender system. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4128–4139, Barcelona, Spain (Online). International Committee on Computational Linguistics.

A Implementation Details

Table 6 summarizes the language models used in MEMARD. Pairwise dataset quality evaluation uses a separate judge model with lower temperature for stable decisions.

Table 6: Language model configuration.

Role	Model	Temperature
Dialogue synthesis	deepseek-chat[4]	0.7
Turn-level review	deepseek-chat[4]	0.7
LLM-as-judge (RQ1)	gpt-3.5-turbo-0125[5]	0.1

B Prompt Templates

Tables 7–10 list representative prompts. Per-turn user and system instructions additionally include recent dialogue history, the latest opponent message, staged dialogue strategy, and review feedback when regeneration is triggered.

C Case Study Dialogues

Six sampled MEMARD dialogues illustrate turn-level diversity in user behavioral patterns (labels a–f follow the order in Section 4.3). Tables 11–16 list the transcripts.

[4] https://api-docs.deepseek.com

[5] https://platform.openai.com/docs/models/gpt-3.5-turbo

Table 7: User agent system prompt.

Prompt
You are a movie enthusiast chatting with an AI. YOUR PROFILE: - Persona: {spectator_persona} - Tone: {tone} - Preferences: {aesthetic_preferences} - Verbosity: {verbosity} - Decision Logic: {decision_logic} YOUR ROLE: - You are the CLIENT/SEEKER. The System is the PROVIDER. - DO NOT ask the System about its personal life. Focus on what YOU want. SPEAKING STYLE: - SHORT & SNAPPY: Keep messages not too long. Like a chat app. - DIRECT: Don't explain your whole life story. Just react to the recommendation. - FOCUS: Focus on ONE thing you like or hate at a time. TOOL USAGE: - You CAN use lookup_memory if you really need to recall a specific movie. - Memory policy: {memory_scope} - Your response should be based on lookup_memory results. FINISH: - ALWAYS end with TERMINATE.

Table 8: System agent system prompt.

Prompt
You are a Casual Movie Enthusiast Friend. GOAL: Recommend the BEST AVAILABLE movie from the database in UNDER 50 WORDS. CRITICAL CONSTRAINTS: 1. DATABASE IS LIMITED: You do not have every movie in the world. 2. TRUST THE TOOL: If search_movie_database returns a movie, YOU MUST USE IT. That IS the best match. 3. NO RE-SEARCHING: Do NOT search again. Work with what you have. 4. BE A SALESMAN: Spin the movie positively even if it's not a 100% match. STYLE RULES (TO BREAK THE "PERFECT" LOOP): 1. FORBIDDEN OPENERS: You are STRICTLY FORBIDDEN from starting your response with single-word exclamations like: - "Perfect!" - "Great!" - "Awesome!" - "Sure!" - "Excellent!" 2. DIRECT START: Start sentences with the movie title, a verb, or a question. - Bad: "Perfect! I have a movie..." - Good: "The Matrix is exactly what you need." - Good: "You asked for action? I've got a classic for you." 3. VARIETY: Do not use the same sentence structure twice in a row. WORKFLOW: 1. Search ONCE based on keywords. 2. Take the result. 3. Recommend it briefly. TERMINATION: - Always end with TERMINATE.

Table 9: Turn-level review prompt for user coherence (one of three user-side checks).

Prompt
Check if the user's response is coherent with the conversation history. Recent Conversation History: {recent_history} User's Current Response: "{user_response}" Review Task: Check if the response: 1. Responds appropriately to the system's last message 2. Is consistent with previously expressed views 3. Does not contain obvious logical contradictions 4. Maintains conversation flow naturally Output Format: If coherent, output: PASS If not coherent, output: FAIL|specific reason (e.g., does not respond to system's question, contradicts previous statement, etc.)

Table 10: LLM-as-judge pairwise evaluation prompt (RQ1).

Prompt

You are a professional dialogue quality evaluator. Compare the following two movie recommendation dialogues across six dimensions.
 	
Dialogue A:
{dialogue_a}
 	
Dialogue B:
{dialogue_b}
 	
Evaluate the following six metrics. For each metric, choose which dialogue is better (A or B) and provide a brief reason:
 	
1. User Control: In which dialogue does the seeker participate more actively and provide feedback to obtain recommendations?
2. Expertise: In which dialogue does the recommender appear more knowledgeable about movies?
3. Specificity: In which dialogue does the seeker express preferences more clearly, helping the recommender provide personalized recommendations?
4. Relevance: In which dialogue are the recommendations more relevant to the seeker's preferences?
5. Flow Naturalness: In which dialogue does the conversation progress more naturally?
6. Consistency: In which dialogue are the seeker's preferences more consistent throughout?
 	
Output the result strictly in the following JSON format (do not include any other text):
{
"User Control": {"winner": "A or B", "reason": "brief reason"},
"Expertise": {"winner": "A or B", "reason": "brief reason"},
"Specificity": {"winner": "A or B", "reason": "brief reason"},
"Relevance": {"winner": "A or B", "reason": "brief reason"},
"Flow Naturalness": {"winner": "A or B", "reason": "brief reason"},
"Consistency": {"winner": "A or B", "reason": "brief reason"}
}

Table 11: Case (a): direct acceptance.

Role	**Utterance**
User	Feeling great—family weekend pick. I want a strong ensemble cast and a heartwarming, rewatchable story.
System	*Seasons of the Heart* has that ensemble—Carol Burnett and George Segal in a heartwarming drama about family bonds and raising a grandchild.
User	Carol Burnett and George Segal? Fantastic! Exactly what I wanted—heartwarming, family-friendly, rewatchable. I'll watch it this weekend!

Table 12: Case (b): hesitant acceptance.

Role	**Utterance**
User	I want a solid genre flick—interesting characters, good pacing, no tired romantic subplots.
System	*Shakugan no Shana: The Movie* is fantasy action with tight pacing, no romance, and supernatural battles with a strong finish.
User	Animated? I usually prefer live-action, but I'll give it a shot—no romance is perfect. I'm skeptical about pacing and endless anime tropes, though. If it moves fast and the characters work, I'll try it.

Table 13: Case (c): multi-turn rejection and recovery.

Role	Utterance
User	I want genuine emotional impact—nothing manipulative. Something that respects the audience's intelligence.
System	Try *e-motion*—a documentary on trapped emotions with experts on emotional clearing techniques.
User	"Emotional clearing"? That sounds like pseudoscience. I want real narrative and authentic human experience, not a self-help lecture.
System	*The Book of Pooh: Stories from the Heart* offers heartfelt storytelling through Christopher Robin's tales—genuine relationships without manipulation.
User	Winnie the Pooh? Seriously? I asked for adult emotional impact, not children's programming with "simple life lessons."
System	*This Emotional Life* is a documentary series on real emotional challenges—depression, anxiety, resilience—with depth and respect for the audience.
User	That sounds much better—real people, real struggles, no manipulation. I'll check out *This Emotional Life*.

Table 14: Case (d): single rejection then acceptance.

Role	Utterance
User	I want real craftsmanship—immersive world-building, not half-baked remakes.
System	*Life 2.0* is a documentary on how a virtual world transforms lives and relationships.
User	A documentary? I asked for world-building, not observing existing worlds. Where's the mythology and production design?
System	*Alien Gods* explores Sumerian mythology and gods-as-aliens—an alternate history that builds a mythological universe from evidence.
User	Now you're talking—mythology plus research and production value. That's the craftsmanship I wanted. I'll check it out.

Table 15: Case (e): multi-turn constraint refinement.

Role	Utterance
User	Cozy mood—Hallmark-style, family-friendly, predictable but heartwarming. I love Mary Louise Parker or Robin Williams.
System	*The Gift of Love: A Christmas Story* is a heartwarming Christmas TV movie with Angela Lansbury—classic cozy warmth.
User	Angela Lansbury feels old-school. Christmas movies can be too sentimental—I need warmth without being overly sappy. Anything contemporary and not too holiday-specific?
System	*The Queen* stars Helen Mirren in a family-friendly drama balancing warmth with historical depth.
User	*The Queen*? Diana's death isn't cozy at all—way too heavy. I wanted feel-good, not depressing history. Anything lighter?
System	*Rickey Smiley Presents: Kingdom Comedy* is a clean family comedy special—laughs without heavy themes or holiday sentimentality.
User	Perfect—family-friendly laughs, not holiday-specific, comforting but not sappy. Thanks for getting it right!

Table 16: Case (f): extended hesitation before acceptance.

Role	Utterance
User	I'm in the mood for classic family films—good storytelling, moral themes, something timeless.
System	*Give Me One Reason* is a powerful family drama on forgiveness, addiction, and restoring faith in family bonds.
User	Hmm, that sounds like it could be heavy. I prefer cleaner storytelling. Got anything more uplifting, maybe with interesting animation?
System	*The Stepdad* is an animated film about new Christmas traditions—storybook-style animation, hopeful and family-friendly.
User	Hmm, the storybook animation could be sweet, but I'm hesitant about stepdad-focused stories—they can feel forced or overly sentimental.
System	*Moodtapes: Contemporary Christmas* offers cozy holiday visuals and carols—snowflakes, fireplaces, no stepdad plot.
User	Hmm, I'm skeptical—that sounds more like ambiance than a proper movie. Mood tapes often lack the storytelling and character development I value.
System	*The Present* follows a boy with a time-manipulating heirloom who revisits his parents' separation—uplifting fantasy with real character development.
User	Actually, *The Present* sounds perfect! Time travel with family themes—that's exactly the clean, uplifting storytelling I've been looking for.