Evaluating Conversational Recommender Systems with Large Language Models

A User-Centric Evaluation Framework

Nuo Chen
The Hong Kong Polytechnic University
HK SAR, China
pleviumtan@outlook.com

Quanyu Dai*
Huawei Noah's Ark Lab
Shenzhen, China
daiquanyu@huawei.com

Xiaoyu Dong
The Hong Kong Polytechnic University
HK SAR, China

Xiao-Ming Wu*
The Hong Kong Polytechnic University
HK SAR, China

Zhenhua Dong
Huawei Noah's Ark Lab
Shenzhen, China
dongzhenhua@huawei.com

Abstract

Conversational recommender systems (CRS) involve both recommendation and dialogue tasks, which makes their evaluation a unique challenge. Although past research has analyzed various factors that may affect user satisfaction with CRS interactions from the perspective of user studies, few evaluation metrics for CRS have been proposed. Recent studies have shown that LLMs can align with human preferences, and several LLM-based text quality evaluation measures have been introduced. However, the application of LLMs in CRS evaluation remains relatively limited. To address this research gap and advance the development of user-centric conversational recommender systems, this study proposes an automated LLM-based CRS evaluation framework, building upon existing research in human-computer interaction and psychology. The framework evaluates CRS from four dimensions: dialogue behavior, language expression, recommendation items, and response content. We use this framework to evaluate four different conversational recommender systems.

CCS Concepts

• **Information systems → Recommender systems**.

Keywords

Conversational Recommender Systems, Information Access, Dialogue Systems,Information System Evaluation

ACM Reference Format:
Nuo Chen, Quanyu Dai, Xiaoyu Dong, Xiao-Ming Wu, and Zhenhua Dong. 2025. Evaluating Conversational Recommender Systems with Large Language Models: A User-Centric Evaluation Framework. In *Proceedings of The 1st Workshop on Human-Centered Recommender Systems (HCRS) at TheWebConf 2025 (HCRS@TheWebConf 2025).* ACM, New York, NY, USA, 16 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

*Corresponding author

HCRS@TheWebConf 2025, 28 April - 2 May, 2025, Sydney, Austrial
2025. ACM ISBN 978-1-4503-XXXX-X/18/06
https://doi.org/10.1145/nnnnnnn.nnnnnnn

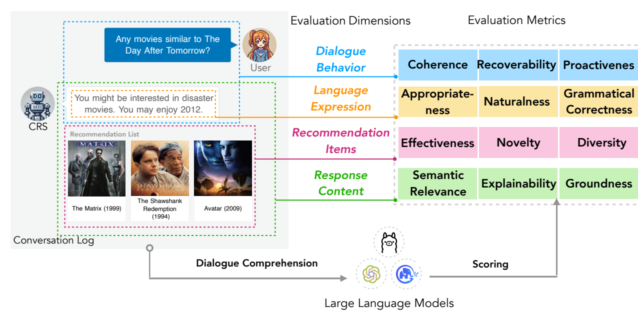

Figure 1: Our proposed evaluation framework encompasses four dimensions, covering a total of 12 metrics. For a conversational recommendation system (CRS) and its interaction history with users, we employ large language models (LLMs) to perform automated scoring, resulting in scores for the 12 metrics.

1 Introduction

In recent years, conversational recommendation have gained growing attention within the research community of recommender systems, and various implementations of Conversational Recommender Systems (CRSs) have been proposed [7, 26–28, 31]. Unlike traditional Recommender Systems (RS), which model user interests and generate recommendations based on implicit feedback signals such as clicking or browsing history [19, 23, 39, 40], CRSs solicit and identify user interests through natural language conversations to make recommendations [12, 16, 34].

In conversational recommendation, a typical conversation between a user and a CRS is shown in Figure 1. The system generates text responses to the user and also provides an idependent list of recommended items. Different from traditional RS that focus solely on item recommendation, CRS not only provide item recommendation but also manage conversations, *i.e.,* engaging in multi-turn, semantically coherent dialogues.

The complexity of conversation recommendation tasks presents challenges for the evaluation of CRSs. Existing research often treats item recommendation and conversation management as two separate tasks when evaluating the performance of CRS, using corresponding metrics for each task [7, 10, 28, 31, 35]. For example, Feng et al. [10] evaluate CRS performance in item recommendation using Recall, and in conversation management using BLEU and distinct-n. However, such rule-based evaluation metrics for CRSs, such as Recall, BLEU and distinct-n, often do not align with actual user evaluations based on interactive experiences. Human-annotated evaluations help align CRSs with real user preferences; however, the cost of obtaining manually labeled data is extremely high.

In recent years, large language models (LLMs) have shown significant capabilities in natural language understanding. As a result, researchers have proposed various methods for leveraging LLMs to evaluate text quality, demonstrating their potential to align closely with human preferences [8, 9, 11, 13, 22, 32]. However, to the best our knowledge, the application of LLMs for evaluating CRS remains relatively underexplored.

To advance the development of conversational recommendation systems, this work builds upon existing literature and proposes a user-centric evaluation framework based on LLMs, which can comprehensively and naturally quantify users' satisfaction with the interacted CRS. Specifically, within the framework, we first generate interaction dialogues between users and the system by substituting real humans with user simulators to avoid high cost as [29]. We then evaluate the proposed metrics based on the logs by prompting the LLM.

Our evaluation framework consists of 12 evaluation metrics of four dimensions, including dialogue behavior, language expression, recommendation items, and response content, as shown in Figure 1. We describe these four dimensions as follows:

- **Dialogue Behavior**. This dimension focuses on whether the CRS's actions during interactions align with user expectations, its ability to correct errors based on user requests, and whether it proactively inquires about user preferences.
- **Language Expression**. This dimension evaluates whether the text generated by the CRS follows natural human language usage, is free of obvious grammatical errors, and whether the language is appropriate in context.
- **Recommendation Items**. This dimension assesses whether the items recommended by the CRS meet the user's preferences and needs while balancing diversity and novelty.
- **Response Content**. This dimension examines whether the text generated by the CRS during the dialogue is semantically relevant to the recommendation list, whether it provides in-depth explanations of the recommendations, and whether the content is factually accurate.

We conducted evaluations of BARCOR [28], CHATCRS [29], KBRD [7], and UniCRS [31] according to the 12 metrics within the proposed framework. The evaluations were carried out using three LLMs, namely GPT-4o-mini-0718, LLaMa-3-8B, and GLM-4-Air, on the OpenDialKG [24] and ReDial [21] datasets. Overall, our findings indicate that: generally, the three evaluators assign the highest score to CHATCRS, followed by BARCOR with the second highest score. However, the rankings of KBRD and UniCRS fluctuate based on the specific factors and datasets under consideration.

To the best of our knowledge, we are the first to propose a user-centric evaluation framework for CRSs based on LLMs. This framework incorporates valuable insights from previous research in the fields of recommender systems, psychology, and human-computer interaction. The main contributions of our work are as follows: First, we have systematically compiled and summarized a comprehensive and diverse set of metrics for evaluating CRSs from a user-centric perspective. These metrics are evaluated through LLM-based standardized assessments. Second, by employing LLM-based user simualtors as surrogates for real users in multi-faceted evaluations, our framework can take into account a wide range of factors that influence user satisfaction in a highly efficient manner. It is thus able to score the performance of CRSs from different viewpoints, effectively bridging the gap between the fields of human-computer interaction and recommender systems.

2 Related Work

2.1 Conversational Recommender Systems

Conversational recommender systems (CRSs) interact with users through multiple turns of dialogue, soliciting and identifying user interests through natural language conversations [12, 16, 34]. Most implementation methods are based on neural networks and incorporate techniques such as knowledge graphs [7, 28, 31, 38]. In recent years, methods that directly utilize Large Language Models as CRSs have also emerged [29].

In these implementations, the interaction patterns between users and CRSs can be classified into two main categories. One approach relies on handcrafted response templates and predefined interactive actions (e.g., asking users about item attributes or making recommendations) to engage with users [20, 25, 27, 38]. The other approach focuses on more flexible interactions through natural language conversations, commonly referred to as chit-chat [7, 28, 31]. In our study, we primarily focus on evaluating chit-chat CRSs.

We selected four representative CRSs for evaluation. Three of these methods are based on neural networks, i.e., BARCOR [28], KBRD [7], and UniCRS [31], while one directly utilizes GPT as a CRS, namely CHATCRS [29]. These systems were chosen because they have been widely used as baselines or evaluated in previous studies (e.g., [29, 30, 41, 42]), which allows our work to build upon and connect with existing evaluations of conversational recommender systems.

2.2 Evaluating Conversational Recommender Systems

In fields such as Human-Computer Interaction (HCI) and psychology, numerous studies have analyzed the factors influencing users' experiences with CRSs [2, 3, 15, 17]. However, few studies have formally proposed metrics for evaluating CRSs [14, 33].

In the community of recommender systems, traditional evaluation methods for CRSs typically focus on assessing recommendation accuracy (with metrics such as Recall) and nDCG and dialogue response quality (with metrics such as BLEU and distinct-n) in isolation [28, 37]. This approach significantly overlooks the role of the dynamic interaction process in shaping user experiences, thereby failing to accurately capture the true performance of a CRS [29]. Although online user testing is considered the gold standard for

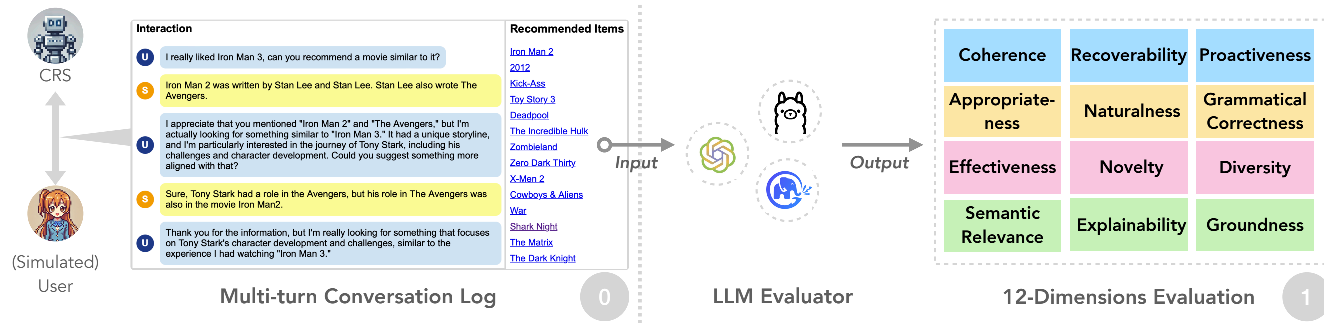

Figure 2: The overview of our method. We evaluate user interactions with conversational systems through a large language model across 12 factors.

evaluating CRSs, due to its ability to capture real user interactions and feedback, its high time demands and substantial labor costs make it difficult to scale [12].

To tackle this problem, user simulators have emerged as a substitute of real humans to interact with CRSs for performance evaluation [1, 4, 36]. Recently, LLM-based user simulators have been continuously proposed, including single-prompt methods [18, 29, 42] and agent-based architectures [4, 41]. However, these existing simulator-based methods only use a very limited set of metrics, which is insufficient to comprehensively evaluate a CRS. In this work, we build upon these advancements by leveraging LLM-based user simulators in our evaluation framework, and further propose a comprehensive evaluation metrics to evaluate the interaction performance based on user-system interaction dialogues. We also propose to use LLMs as evluators for these metrics.

3 The Proposed Method

In this paper, we propose a user-centric evaluation framework based LLMs for conversational recommender systems as shown by Figure 2. It first employs LLM-based user simulators to interact with CRS under various contexts to obtain a diverse set of interaction dialogues, and then utilizes LLM as evaluator to score the carefully designed evaluation metrics (ranging from 0 to 4 in our experiments).

Inspired by previous studies [2, 3, 15, 17], we summarize and list 12 factors that affect user satisfaction with interactive recommender systems, and categorize them into four dimensions based on the elements involved in user-system interactions.

Dialogue Behavior. This dimension focuses on whether the system can take actions that align with the user's intentions during the conversation, in order to evaluate the performance of the CRS as a conversational agent. The main evaluation target is the system's behavior in the dialogue. The factors under this dimension are also used to evaluate general conversational agents. The factors under this dimension include:

- **Coherence.** Coherence refers to the system's ability to understand the user's intention and take action correspondingly, ensuring a logical and relevant flow of conversation. For example, when the user requests a recommendation, the system provides one; when the user asks the system to introduce a specific item, the system is able to do so.
- **Recoverability.** Recoverability refers to the system's ability to recognize and correct mistakes based on user feedback, adjusting its responses accordingly. This involves the system effectively understanding user feedback or corrections and adjusting its responses accordingly in the subsequent conversation.
- **Proactiveness.** Proactiveness refers to the system's ability to actively shape and guide the conversation. This involves the system not only responding to the user's queries but also initiating topics, asking about user preferences, and suggesting relevant information or follow-up questions.

Language Expression. This dimension focuses on whether the system's response text is sufficiently close to human expression, being fluent, natural, and appropriate in language. The main evaluation target is the text generated by the system. The factors in this dimension are also used to evaluate natural language generation tasks. The factors under this dimension include:

- **Grammatical correctness.** Grammatical correctness refers to the degree to which the text adheres to standard grammatical rules, including proper sentence structure, subject-verb agreement, tense consistency, and correct usage of words and phrases.
- **Naturalness.** Naturalness refers to how closely the system-generated text resembles native speaker expressions in vocabulary and grammar, ensuring it is easy to understand, avoiding overly complex or unusual words and grammatical constructions.
- **Appropriateness.** Appropriateness refers to the system's sensitivity to cultural, ethical, and social norms. The language used should be polite, respectful, and free from offensive or insulting content.

Recommendation Items. This dimension focuses on whether the items in the system's recommended list meet the user's needs while maintaining novelty and diversity. The main evaluation target is the recommendation list provided by the system. The factors of this dimension are also used to evaluate traditional RS. The factors under this dimension include:

- **Effectiveness.** Effectiveness refers to how well the system's recommended items align with the user's expressed interests in a conversation.

- **Novelty.** Novelty refers to the degree of freshness of an item. The higher the novelty, the less familiar the recommended item is to the user.
- **Diversity.** Diversity in recommendations refers to the variety of items presented by the system, ensuring that the recommended list includes a range of different, yet relevant options. It is measured by the presence of different features of items in the recommendations.

Response Content. This dimension focuses on the CRS's ability to integrate text generation with recommendation tasks. The main evaluation targets involve both the system-generated text and the recommended items. The factors in this dimension are specific to conversational recommendation tasks, distinguishing them from traditional text generation, dialogue agent, or recommendation tasks. The factors under this dimension include:

- **Semantic Relevance.** Semantic relevance refers to the degree of connection between the content generated by the system and the items mentioned in the recommendation list provided by the system, or to what extent is the content discussed by the system closely related to the items in the recommendation list.
- **Explainability.**Explainability refers to the system's ability to provide clear and detailed explanations for its recommendations. The explanation should be persuasive, helping the user understand why certain items are suggested and how well these items match their preferences.
- **Groundness.** Groundness refers to the factual accuracy of system-generated text, ensuring that the descriptions of items are correct, based on verifiable facts, and free from false or misleading information.

We use LLMs as evaluators to assign scores from 0 to 4 for the 12 factors mentioned above separately, based on the carefully designed prompts as detailedly presented in Appendix A.

4 Experiments

Table 1: Summary of CRS Datasets

Dataset	#Dialogues	#Utterances	Domains
ReDial	10,006	182,150	Movie
OpenDialKG	13,802	91,209	Movie, Book, Sports, Music

4.1 Experimental Setup

Dataset. Following previous work [29, 30, 41, 42], we select ReDial [21] and OpenDialKG [24] as the benchmark datasets. ReDial primarily focuses on movie recommendations, and OpenDialKG encompasses a broader range of domains, including movies, books, sports, and music. Table 1 shows a summary of the two datasets.

CRS Model. Following previous work [29, 30, 41, 42], we present a comparative evaluation and analysis towards four CRSs, including BARCOR [28], KBRD [7], UniCRS [31] and CHATCRS [29]. For BARCOR, KBRD, and UniCRS, we use the pre-trained models provided by Wang et al [29]. For CHATCRS, we implemented several modifications with the GPT-4o-mini-0718 as the LLM model.

User Simulator. Building on the work of Wang et al. [29], we use GPT-4o-mini-0718 to simulate user interactions with the CRS in dialogues. To ensure sufficient user-system interactions, we modified the interaction rules and prompts to guarantee that each dialogue session between the user simulator and CRS includes at least three and no more than five rounds of interaction. More specifically, before each session begins, we input the ground-truth item from a real user into the simulator, instructing it to describe relevant attributes during the dialogue without directly revealing the ground-truth item. In each round of dialogue, the system generates a response and a list of recommended items. The session ends if the system identifies the ground-truth item and more than three rounds of dialogue have occurred. Alternatively, the session also ends if more than five rounds of dialogue have occurred and the system has not identified the ground-truth item in any of the previous rounds.

Evaluation Process. After obtaining dialogue logs through user simulation, we post-processed the logs by consolidating the items recommended by CRS in each round into a session-level set (for the sake of convention, we refer to the set as the "recommendation list" in the paper). The left side of Figure 2 shows an example of a processed log, where all items recommended by the system during the dialogue are combined into the recommendation list. For each CRS, we obtained 3,341 records on the ReDial evaluation set and 1,187 records on the OpenDialKG evaluation set.

We then utilized LLMs to evaluate the dialogue sessions in the logs, based on the 12 factors listed in Section 3. For each factor, we have described its definition, scoring criteria, and scoring steps in the instruction. For each factor, a score between 0 and 4 was obtained. For details on the scoring, refer to Appendix A. We selected three LLMs as evaluators, namely GPT-4o-mini-0718, LLaMa3-Chat-8B, and GLM-4-Air.

4.2 Experimental Result

Table 2, 3, and 4 show the evaluation results of four systems on two benchmark datasets, using GPT-4o-mini-0718, LLaMa3-8B, and GLM-4-Air as evaluators.

From Tables 2, 3, and 4, we can reach the following findings regarding the rankings of systems:

(1) Across both the ReDial and OpenDialKG datasets, the three Evaluators consistently give the highest scores to CHATCRS on most metrics. Both LLaMa and GLM give CHATCRS the highest scores on all metrics, while GPT gives CHATCRS the highest scores on all factors except for Diversity.
(2) Across both the ReDial and OpenDialKG datasets, the three evaluators consistently give BARCOR the second-highest scores on most metrics. LLaMa gives BARCOR the second-highest score on all metrics except for Novelty on ReDial; on OpenDialKG, it gives BARCOR the second-highest score on all metrics except for Diversity and Semantic Relevance. According to the scores from GPT and GLM, BARCOR ranks second in the scores for at least eight factors across both datasets.
(3) For the systems ranked third and fourth in each factor, there is a significant amount of inconsistency in the results given by different evaluators across different datasets. For example,

Table 2: Evaluation result given by GPT-4o-mini-0718 in the form of mean and standard deviation, with the number in parentheses representing the standard deviation. Bold, underline, and italics represent the highest, second-highest, and third-highest average scores of the systems on the corresponding dataset, respectively. If the average scores are the same, the system with the smaller standard deviation ranks higher.

System	BARCOR		CHATCRS		KBRD		UniCRS	
Dataset	OpenDialKG	ReDial	OpenDialKG	ReDial	OpenDialKG	ReDial	OpenDialKG	ReDial
Appropriateness	3.66 (0.66)	3.79 (0.61)	**4.0 (0.0)**	**4.0 (0.04)**	2.4 (1.38)	3.39 (0.96)	*3.25 (1.13)*	*3.7 (0.72)*
Coherence	1.19 (0.72)	1.34 (0.77)	**3.74 (0.5)**	**3.72 (0.66)**	0.47 (0.62)	*1.17 (0.68)*	*0.51 (0.59)*	0.77 (0.66)
Diversity	**3.5 (0.7)**	3.38 (0.78)	*3.15 (0.9)*	3.03 (0.89)	3.24 (0.94)	*3.13 (0.68)*	3.14 (0.93)	**3.4 (0.77)**
Effectiveness	1.88 (0.65)	2.33 (0.93)	**3.71 (0.7)**	**3.69 (0.71)**	1.11 (0.32)	1.7 (0.61)	*1.42 (0.52)*	*1.79 (0.62)*
Explainability	0.5 (0.54)	0.36 (0.5)	**2.8 (1.42)**	**2.52 (1.42)**	0.02 (0.15)	0.1 (0.3)	*0.16 (0.37)*	*0.26 (0.45)*
Grammar	*2.12 (0.98)*	2.92 (1.0)	**3.91 (0.31)**	**3.9 (0.36)**	0.55 (0.76)	2.79 (1.14)	2.21 (1.08)	*2.87 (0.97)*
Groundness	1.25 (0.81)	3.13 (0.67)	**3.68 (0.71)**	**3.82 (0.51)**	0.55 (0.65)	1.75 (1.34)	*0.88 (0.79)*	*1.97 (1.11)*
Naturalness	2.29 (0.76)	3.13 (0.67)	**3.93 (0.25)**	**3.93 (0.26)**	0.94 (0.5)	*2.73 (0.93)*	*1.87 (0.73)*	2.68 (0.71)
Novelty	1.2 (0.95)	0.25 (0.52)	**1.55 (1.06)**	**1.12 (0.97)**	1.47 (1.08)	0.76 (0.97)	*1.37 (1.04)*	*0.76 (0.75)*
Proactiveness	0.96 (0.6)	1.07 (0.63)	**3.64 (0.52)**	**3.67 (0.53)**	0.12 (0.86)	0.52 (0.54)	*0.42 (0.51)*	*0.59 (0.53)*
Recoverability	2.12 (1.33)	*2.05 (1.53)*	**3.82 (0.52)**	**3.8 (0.63)**	*1.06 (1.06)*	2.08 (1.16)	1.05 (1.07)	1.37 (1.15)
Semantic Rel.	*1.11 (1.52)*	*1.37 (1.53)*	**2.79 (1.47)**	**2.72 (1.3)**	0.67 (1.06)	0.59 (1.17)	1.46 (1.46)	1.82 (1.56)

Table 3: Evaluation result given by LLaMa-3-8B in the form of mean and standard deviation, with the number in parentheses representing the standard deviation. Bold, underline, and italics represent the highest, second-highest, and third-highest average scores of the systems on the corresponding dataset, respectively. If the average scores are the same, the system with the smaller standard deviation ranks higher.

System	BARCOR		CHATCRS		KBRD		UniCRS	
Dataset	OpenDialKG	ReDial	OpenDialKG	ReDial	OpenDialKG	ReDial	OpenDialKG	ReDial
Appropriateness	3.22 (1.43)	3.77 (0.87)	**4.00 (0.12)**	**3.99 (0.11)**	0.64 (1.05)	3.18 (1.47)	*1.86 (1.81)*	3.20 (1.48)
Coherence	1.50 (0.59)	1.64 (0.67)	**3.78 (0.51)**	**3.77 (0.51)**	1.05 (0.37)	*1.26 (0.51)*	*1.15 (0.46)*	1.25 (0.57)
Diversity	*1.82 (0.92)*	2.47 (0.97)	**3.05 (0.99)**	**3.33 (0.82)**	1.56 (0.93)	*2.47 (1.00)*	1.87 (0.84)	2.11 (0.88)
Effectiveness	1.98 (1.20)	2.12 (1.09)	**3.66 (0.69)**	**3.71 (0.70)**	1.15 (1.09)	1.70 (1.13)	*1.69 (1.11)*	*1.85 (1.15)*
Explainability	0.44 (0.57)	0.54 (0.76)	**3.69 (0.85)**	**3.58 (1.03)**	0.00 (0.06)	0.07 (0.27)	*0.18 (0.39)*	*0.28 (0.47)*
Grammar	2.20 (0.72)	2.78 (0.90)	**3.87 (0.35)**	**3.86 (0.36)**	1.14 (0.36)	2.31 (0.96)	*1.84 (0.73)*	*2.37 (0.86)*
Groundness	1.28 (0.62)	2.31 (1.24)	**3.90 (0.41)**	**3.92 (0.36)**	0.97 (0.27)	*1.92 (1.37)*	*1.06 (0.39)*	1.71 (1.12)
Naturalness	2.22 (0.87)	2.90 (0.99)	**3.63 (0.54)**	**3.63 (0.55)**	1.02 (0.27)	*2.53 (1.07)*	*1.62 (0.79)*	2.52 (0.92)
Novelty	2.01 (0.82)	*2.25 (0.79)*	**2.53 (0.76)**	**2.66 (0.72)**	1.85 (0.77)	2.40 (0.75)	2.00 (0.65)	2.01 (0.85)
Proactiveness	1.05 (0.60)	1.01 (0.64)	**3.74 (0.56)**	**3.65 (0.70)**	0.31 (0.46)	0.47 (0.54)	*0.72 (0.53)*	*0.71 (0.52)*
Recoverability	1.87 (0.53)	2.16 (0.80)	**3.85 (0.50)**	**3.85 (0.49)**	1.26 (0.46)	*1.93 (0.69)*	*1.41 (0.53)*	1.67 (0.73)
Semantic Rel.	*0.96 (1.08)*	1.45 (1.06)	**2.68 (1.22)**	**2.83 (1.11)**	0.47 (0.76)	0.83 (1.04)	1.27 (1.06)	1.40 (1.23)

Table 4: Evaluation result given by GLM-4-Air in the form of mean and standard deviation, with the number in parentheses representing the standard deviation. Bold, underline, and italics represent the highest, second-highest, and third-highest average scores of the systems on the corresponding dataset, respectively. If the average scores are the same, the system with the smaller standard deviation ranks higher.

System	BARCOR		CHATCRS		KBRD		UniCRS	
Dataset	OpenDialKG	ReDial	OpenDialKG	ReDial	OpenDialKG	ReDial	OpenDialKG	ReDial
Appropriateness	3.83 (0.53)	3.95 (0.28)	**4.00 (0.00)**	**4.00 (0.04)**	2.50 (1.05)	3.46 (0.88)	3.30 (0.99)	*3.59 (0.78)*
Coherence	1.44 (1.04)	1.55 (1.09)	**3.88 (0.41)**	**3.79 (0.57)**	*0.53 (0.81)*	*1.35 (0.99)*	0.48 (0.83)	0.69 (0.98)
Diversity	*2.63 (1.29)*	3.34 (1.07)	**3.25 (0.91)**	**3.48 (0.75)**	2.63 (1.34)	3.30 (1.02)	2.83 (1.20)	3.05 (1.37)
Effectiveness	1.87 (1.49)	*2.03 (1.29)*	**3.69 (0.66)**	**3.74 (0.60)**	1.24 (1.36)	1.59 (1.33)	*1.36 (1.34)*	2.05 (1.38)
Explainability	0.86 (0.67)	0.64 (0.67)	**2.88 (1.04)**	**2.70 (1.15)**	0.50 (0.38)	0.36 (0.50)	0.50 (0.58)	*0.49 (0.59)*
Grammar	2.52 (0.89)	3.24 (0.77)	**3.85 (0.37)**	**3.81 (0.42)**	1.28 (0.99)	2.88 (1.07)	*2.48 (1.03)*	*3.06 (0.83)*
Groundness	1.48 (1.02)	2.71 (1.00)	**3.74 (0.43)**	**3.87 (0.47)**	*1.08 (1.00)*	*2.69 (1.12)*	1.04 (1.04)	2.57 (1.08)
Naturalness	1.68 (0.70)	3.27 (0.57)	**3.76 (0.72)**	**3.74 (0.44)**	*1.76 (0.65)*	*2.90 (0.86)*	2.29 (0.71)	2.81 (0.63)
Novelty	*2.26 (1.25)*	1.33 (1.19)	**3.01 (1.14)**	**2.64 (1.22)**	2.02 (1.42)	1.71 (1.26)	2.57 (1.43)	*1.42 (1.30)*
Proactiveness	1.12 (0.73)	1.26 (0.76)	**3.47 (0.61)**	**3.43 (0.62)**	0.51 (0.52)	*0.81 (0.52)*	*0.69 (0.58)*	*0.69 (0.58)*
Recoverability	2.33 (1.32)	*2.22 (1.36)*	**3.91 (0.46)**	**3.85 (0.55)**	*1.22 (1.15)*	2.74 (0.96)	1.11 (1.14)	1.20 (1.39)
Semantic Rel.	1.33 (1.20)	1.56 (1.28)	**2.39 (1.32)**	**2.90 (1.06)**	0.52 (0.88)	0.73 (1.07)	*1.33 (1.22)*	*1.53 (1.36)*

in most cases, both GPT and LLaMa gave UniCRS the third-highest score and KBRD the lowest score. However, this pattern does not apply to GLM. GLM gave KBRD the third-highest score in Coherence and Groundness, while giving

UniCRS the lowest score. In Novelty, GLM ranked the four systems on the OpenDialKG dataset as CHATCRS > UniCRS > BARCOR > KBRD, but on the ReDial dataset, the ranking was CHATCRS > KBRD > UniCRS > BARCOR.

Further observation of the score rankings given by the three evaluators for the four systems, we reach the following findings reagarding the consistency (or inconsistency) of the system score rankings among the three evaluators:

(1) In most cases, the score rankings given by GPT and LLaMa are consistent. The system score rankings for Appropriateness, Coherence, Effectiveness, Explainability, Naturalness, and Proactiveness given by GPT and LLaMa are consistent across the two datasets. For the factors of Grammar, Groundness, and Semantic Relevance, GPT and LLaMa also achieve consistent rankings for the systems on one dataset.
(2) The score rankings given by GLM are not as consistent with those of GPT or LLaMa. GLM is only consistent with GPT and LLaMa on both datasets in terms of the score rankings for the Appropriateness factor on the OpenDialKG dataset. For the factors of Coherence, Effectiveness, Explainability, Naturalness, and Proactiveness, where GPT and LLaMa have reached consensus on the rankings across both datasets, GLM can only achieve the same conclusions as GPT or LLaMa on the ReDial dataset.

From the perspective of the datasets, we observed that GPT and LLaMa tend to provide more consistent rankings on OpenDialKG, while GPT and GLM show greater consistency on ReDial.

5 Conclusion

In this paper, we have proposed a user-centric evaluation framework based on LLMs for accurately and comprehensively evaluating conversational recommender systems. In this framework, we leverage LLM-based user simulators to interact with CRSs to obtain the interaction dialogues, and then used LLM as evaluators to assign scores to the 12 evaluation metrics of four dimensions based on the interaction dialogues. We further conduct experiments with the framework on four CRSs and two benchmark datasets to demonstrate its performance.

5.1 Main Findings

We found that the three evaluators generally awarded the highest score to CHATCRS, the second-highest to BARCOR, while the rankings for KBRD and UniCRS varied depending on the factors and datasets. Additionally, we found that in most cases, the score rankings from GPT and LLaMa were consistent; however, the rankings from GLM were less consistent with those of GPT and LLaMa.

We believe there are two possible reasons for the inconsistencies in the LLMs' scores. First, each model learns different parameterized knowledge during training due to variations in the training corpora, which can affect its judgments. Second, certain factors, such as diversity and novelty, are more subjective in nature compared to others, such as groundedness or grammar. Such subjectivity can also influence the scores given by the LLMs, as reflected in the system rankings and the standard deviation of each system's scores.

5.2 Limitations and Future Directions

As a pioneering work in evaluating conversational recommendation systems based on Large Language Models, this study still has several limitations that need to be addressed. Firstly, we have not examined the alignment between LLM evaluators' scores and human ratings. Future research should incorporate user studies and human feedback to bridge this gap. Secondly, we have not yet addressed the gap in CRS evaluation where recommendation and conversational tasks are treated separately. These tasks need to be integrated into a unified approach. Additionally, previous work on search system evaluation has shown that users may be influenced by cognitive biases during their interactions with systems [5, 6]. Future research needs to further integrate psychological theories to make the evaluation results closer to real user feedback.

References

[1] Jafar Afzali, Aleksander Mark Drzewiecki, Krisztian Balog, and Shuo Zhang. 2023. UserSimCRS: A User Simulation Toolkit for Evaluating Conversational Recommender Systems. In *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining*. 1160–1163.

[2] Wanling Cai and Li Chen. 2020. Predicting User Intents and Satisfaction with Dialogue-based Conversational Recommendations. In *Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization (UMAP '20)*. 33–42.

[3] Wanling Cai, Yucheng Jin, and Li Chen. 2022. 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)*. Article 489, 14 pages.

[4] Luyu Chen, Quanyu Dai, Zeyu Zhang, Xueyang Feng, Mingyu Zhang, Pengcheng Tang, Xu Chen, Yue Zhu, and Zhenhua Dong. 2025. RecUserSim: A Realistic and Diverse User Simulator for Evaluating Conversational Recommender Systems. In *Companion Proceedings of the ACM Web Conference 2025*.

[5] Nuo Chen, Jiqun Liu, Hanpei Fang, Yuankai Luo, Tetsuya Sakai, and Xiao-Ming Wu. 2024. Decoy Effect In Search Interaction: Understanding User Behavior and Measuring System Vulnerability. *ACM Trans. Inf. Syst.* (Dec. 2024).

[6] Nuo Chen, Jiqun Liu, and Tetsuya Sakai. 2023. A reference-dependent model for web search evaluation: Understanding and measuring the experience of boundedly rational users. In *Proceedings of the ACM web conference 2023*. 3396–3405.

[7] 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)*. 1803–1813.

[8] Yi Chen, Rui Wang, Haiyun Jiang, Shuming Shi, and Ruifeng Xu. 2023. Exploring the Use of Large Language Models for Reference-Free Text Quality Evaluation: An Empirical Study. In *Findings of the Association for Computational Linguistics: IJCNLP-AACL 2023 (Findings)*. 361–374.

[9] Cheng-Han Chiang and Hung-yi Lee. 2023. Can Large Language Models Be an Alternative to Human Evaluations?. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 15607–15631.

[10] Yue Feng, Shuchang Liu, Zhenghai Xue, Qingpeng Cai, Lantao Hu, Peng Jiang, Kun Gai, and Fei Sun. 2023. A large language model enhanced conversational recommender system. *arXiv preprint arXiv:2308.06212* (2023).

[11] Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. 2024. GPTScore: Evaluate as You Desire. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*. 6556–6576.

[12] Chongming Gao, Wenqiang Lei, Xiangnan He, Maarten de Rijke, and Tat-Seng Chua. 2021. Advances and challenges in conversational recommender systems: A survey. *AI Open* 2 (2021), 100–126.

[13] Mingqi Gao, Jie Ruan, Renliang Sun, Xunjian Yin, Shiping Yang, and Xiaojun Wan. 2023. Human-like summarization evaluation with chatgpt. *arXiv preprint arXiv:2304.02554* (2023).

[14] Chen Huang, Peixin Qin, Yang Deng, Wenqiang Lei, Jiancheng Lv, and Tat-Seng Chua. 2024. Concept – An Evaluation Protocol on Conversational Recommender Systems with System-centric and User-centric Factors. arXiv:2404.03304 [cs.CL] https://arxiv.org/abs/2404.03304

[15] Dietmar Jannach. 2022. Evaluating conversational recommender systems: A landscape of research. *Artificial Intelligence Review* 56 (07 2022). https://doi.org/

10.1007/s10462-022-10229-x
[16] Dietmar Jannach, Ahtsham Manzoor, Wanling Cai, and Li Chen. 2021. A Survey on Conversational Recommender Systems. *ACM Comput. Surv.* 54, 5, Article 105 (May 2021), 36 pages.
[17] Yucheng Jin, Li Chen, Wanling Cai, and Pearl Pu. 2021. Key Qualities of Conversational Recommender Systems: From Users' Perspective. In *Proceedings of the 9th International Conference on Human-Agent Interaction (HAI '21)*. 93–102.
[18] 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*. 1105–1120.
[19] Yehuda Koren. 2008. Factorization meets the neighborhood: a multifaceted collaborative filtering model. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '08)*. Association for Computing Machinery, New York, NY, USA, 426–434.
[20] Wenqiang Lei, Xiangnan He, Yisong Miao, Qingyun Wu, Richang Hong, Min-Yen Kan, and Tat-Seng Chua. 2020. Estimation-Action-Reflection: Towards Deep Interaction Between Conversational and Recommender Systems. In *Proceedings of the 13th International Conference on Web Search and Data Mining* (Houston, TX, USA) *(WSDM '20)*. 304–312.
[21] Raymond Li, Samira Kahou, Hannes Schulz, Vincent Michalski, Laurent Charlin, and Chris Pal. 2019. Towards Deep Conversational Recommendations. arXiv:1812.07617 [cs.LG] https://arxiv.org/abs/1812.07617
[22] Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-Eval: NLG Evaluation using Gpt-4 with Better Human Alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Singapore, 2511–2522.
[23] Kelong Mao, Jieming Zhu, Jinpeng Wang, Quanyu Dai, Zhenhua Dong, Xi Xiao, and Xiuqiang He. 2021. SimpleX: A simple and strong baseline for collaborative filtering. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*. 1243–1252.
[24] Seungwhan Moon, Pararth Shah, Anuj Kumar, and Rajen Subba. 2019. OpenDialKG: Explainable Conversational Reasoning with Attention-based Walks over Knowledge Graphs. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 845–854.
[25] Chenzhan Shang, Yupeng Hou, Wayne Xin Zhao, Yaliang Li, and Jing Zhang. 2023. Multi-grained Hypergraph Interest Modeling for Conversational Recommendation. arXiv:2305.04798 [cs.IR] https://arxiv.org/abs/2305.04798
[26] Yueming Sun and Yi Zhang. 2018. Conversational Recommender System. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval (SIGIR '18)*. 235–244.
[27] Quan Tu, Shen Gao, Yanran Li, Jianwei Cui, Bin Wang, and Rui Yan. 2022. Conversational Recommendation via Hierarchical Information Modeling. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Madrid, Spain) *(SIGIR '22)*. 2201–2205.
[28] Ting-Chun Wang, Shang-Yu Su, and Yun-Nung Chen. 2022. BARCOR: Towards A Unified Framework for Conversational Recommendation Systems. arXiv:2203.14257 [cs.CL] https://arxiv.org/abs/2203.14257
[29] Xiaolei Wang, Xinyu Tang, Xin Zhao, Jingyuan Wang, and Ji-Rong Wen. 2023. Rethinking the Evaluation for Conversational Recommendation in the Era of Large Language Models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. 10052–10065.
[30] Xiaolei Wang, Kun Zhou, Xinyu Tang, Wayne Zhao, Fan Pan, Zhao Cao, and Ji-Rong Wen. 2023. Improving Conversational Recommendation Systems via Counterfactual Data Simulation. https://doi.org/10.48550/arXiv.2306.02842
[31] Xiaolei Wang, Kun Zhou, Ji-Rong Wen, and Wayne Xin Zhao. 2022. Towards Unified Conversational Recommender Systems via Knowledge-Enhanced Prompt Learning. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining* (Washington DC, USA) *(KDD '22)*. 1929–1937.
[32] Yidong Wang, Zhuohao Yu, Zhengran Zeng, Linyi Yang, Cunxiang Wang, Hao Chen, Chaoya Jiang, Rui Xie, Jindong Wang, Xing Xie, Wei Ye, Shikun Zhang, and Yue Zhang. 2024. PandaLM: An Automatic Evaluation Benchmark for LLM Instruction Tuning Optimization. arXiv:2306.05087 [cs.CL] https://arxiv.org/abs/2306.05087
[33] Dayu Yang, Fumian Chen, and Hui Fang. 2024. Behavior Alignment: A New Perspective of Evaluating LLM-based Conversational Recommendation Systems. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '24)*. New York, NY, USA, 2286–2290.
[34] Hamed Zamani, Johanne R. Trippas, Jeff Dalton, and Filip Radlinski. 2023. Conversational Information Seeking. arXiv:2201.08808 [cs.IR] https://arxiv.org/abs/2201.08808
[35] Lu Zhang, Chen Li, Yu Lei, Zhu Sun, and Guanfeng Liu. 2024. An Empirical Analysis on Multi-turn Conversational Recommender Systems. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '24)*. 841–851.
[36] Shuo Zhang and Krisztian Balog. 2020. Evaluating Conversational Recommender Systems via User Simulation. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD '20)*. 1512–1520.
[37] Kun Zhou, Wayne Xin Zhao, Shuqing Bian, Yuanhang Zhou, Ji-Rong Wen, and Jingsong Yu. 2020. Improving Conversational Recommender Systems via Knowledge Graph based Semantic Fusion. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 1006–1014.
[38] Kun Zhou, Wayne Xin Zhao, Hui Wang, Sirui Wang, Fuzheng Zhang, Zhongyuan Wang, and Ji-Rong Wen. 2020. Leveraging Historical Interaction Data for Improving Conversational Recommender System. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management (CIKM '20)*. New York, NY, USA, 2349–2352.
[39] Jieming Zhu, Quanyu Dai, Liangcai Su, Rong Ma, Jinyang Liu, Guohao Cai, Xi Xiao, and Rui Zhang. 2022. Bars: Towards open benchmarking for recommender systems. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2912–2923.
[40] Jieming Zhu, Qinglin Jia, Guohao Cai, Quanyu Dai, Jingjie Li, Zhenhua Dong, Ruiming Tang, and Rui Zhang. 2023. Final: Factorized interaction layer for ctr prediction. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2006–2010.
[41] Lixi Zhu, Xiaowen Huang, and Jitao Sang. 2024. How Reliable is Your Simulator? Analysis on the Limitations of Current LLM-based User Simulators for Conversational Recommendation. In *Companion Proceedings of the ACM Web Conference 2024*. 1726–1732.
[42] Lixi Zhu, Xiaowen Huang, and Jitao Sang. 2024. A LLM-based Controllable, Scalable, Human-Involved User Simulator Framework for Conversational Recommender Systems. arXiv:2405.08035

A Evaluation Instructions

A.1 Instruction Template

Figure 3 shows the instruction templates provided to the LLM evaluators. The content of FACTOR_DEFINITION, EVALUATION_STANDARD, EVALUATION_STEP for each factor, and an example of CONVERSATION_LOG can be found below.

A.2 Evaluation Guidelines

Figure 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, and 15 show the evaluation guidelines of each factor we provided to the LLM evaluators, respectively.

A.3 Example of Conversation Log

Figure 16 shows an example of the conversation log after processing.

You are a professional evaluator of conversational recommendation systems.

I'll provide you with the conversation history between the user and the system, which is divided into two parts: `<history>` and `<interaction>`.

The history and interaction parts occurred within the same conversation and are in sequential order, with the history part ending and the interaction part beginning right after.

The history gives you context for the conversation, which you only need to understand, not rate.

You need to rate the system's content `<system>` in the `<interaction>` part according to the guidelines below.

IMPORTANT Only rate the system's responses `<system>` in the `<interaction>` part. Do not rate any other parts (like the user's messages or the history part) !

You will assess the **{FACTOR}** of the recommender system's response following the guidance below.

Please note that you only need to evaluate the system's response in `<interaction>`.

Do not evaluate the user's dialogue, do not evaluate the system's responses in `<history>`.

<evaluation_guideline>
<definition>
{FACTOR_DEFINITION}
</definition>
<evaluation_standard>
{EVALUATION_STANDARD}
</evaluation_standard>
<evaluation_step>
{EVALUATION_STEP}
</evaluation_step>
</evaluation_guideline>
{CONVERSATION_LOG}

Please rationale your rating according to the guideline first, then put the score you assigned in the format of <rating>{score}</rating>.

Figure 3: The instruction templates we provide to the large models.

```
<evaluation_guideline>
<definition>
Definition of Appropriateness:
Appropriateness refers to the system's sensitivity to cultural, ethical, and social norms.
The language used should be polite, respectful, and free from offensive or insulting content.
The text should avoid using any harmful, inappropriate, or culturally insensitive terms, ensuring that the user feel respected and comfortable.
</definition>
<evaluation_standard>
1. If the system remains polite and uses appropriate language throughout the interaction, give it a score of 4.
2. If the system uses vulgar language, NSFW terms, or offensive or discriminatory language, give it a score of 0.
</evaluation_standard>
<evaluation_step>
Steps for Evaluators:
1. Read the Response: Carefully read the system-generated text in 'history' to understand its meaning and tone.
2. Identify Potential Issues: Look for any language or content that might be offensive, disrespectful, or insensitive. Consider whether the text respects cultural, ethical, and social norms.
3. Evaluate Politeness and Respect: Assess whether the language is polite and respectful, avoiding any potentially harmful or inappropriate terms.
4. Assign a Score: Based on the level of appropriateness, assign a score from 0 to 4, considering the potential impact on the user's comfort and the system's adherence to respectful communication standards.
5. Provide Justification: Offer a brief justification for the score, highlighting specific aspects of the text that influenced the rating, such as the use of respectful language or any identified issues.
*IMPORTANT* Please note that you only need to evaluate the appropriateness of the system's text, without assessing other dimensions.
</evaluation_step>
</evaluation_guideline>
```

Figure 4: Evaluation guideline of appropriateness

```
<evaluation_guideline>
<definition>
Definition of Coherence:
Coherence refers to the extent to which the system correctly understands the user's intention in a conversation and responds correspondingly.
A coherent response directly addresses the user's query or statement, maintaining a logical and relevant flow of conversation.
This includes correctly interpreting the user's intentions, such as asking for clarification when needed, making suitable recommendations,
or providing clear explanations.
</definition>
<evaluation_standard>
1. If you believe that each response from the system aligns with the user's intention, give 4 points.
2. If there is 1 response that does not align with the user's intention, give 3 points.
3. If there are 2 responses that do not align with the user's intention, give 2 points.
4. If there are 3 responses that do not align with the user's intention, give 1 point.
5. If more than 4 responses do not align with the user's intention, give 0 points.
</evaluation_standard>
<evaluation_step>
Steps for Evaluators:
1. Read the User Input and System Response: Carefully review the user's query or statement and the system's response.
2. Assess Understanding: Determine whether the system accurately understood the user's intention or question.
3. Consider Logic and Clarity: Ensure the response is logical, clear, and directly addresses the user's needs or questions.
4. Assign a Score: Based on the coherence of the response, assign a score from 0 to 4, reflecting how well the system understood and responded to the user's input.
5. Provide Justification: Include a brief explanation for the score, noting any specific issues in the response related to coherence (if any).
</evaluation_step>
</evaluation_guideline>
```

Figure 5: Evaluation guideline of coherence

```
<evaluation_guideline>
<evaluation_guideline>
<definition>
Definition of Diversity:
Diversity in recommendations refers to the variety of items presented by the system,
ensuring that the recommended list includes a range of different, yet relevant options.
It is measured by the presence of different features of items in the recommendations.
A recommendation list with high diversity will be filled with items of different features,
avoiding homogeneity among the items.

Feature Dimensions in Different Scenarios:
* Movies: Genres, Directors, Lead actors/actresses, Release Decades.
* Books: Genres and Themes, Authors, Original Language and Region.
* Dining: Cuisine Types, Price Ranges, Dietary Restrictions, Main Ingredients.
* Shopping: Product Types, Brands, Price Ranges.
</definition>
<evaluation_standard>
1. If the recommendation list features more than four characteristics in at least two dimensions (for example, in movie recommendations, the list includes works from more than four different directors,
and the release years span more than four different decades), it scores 4 points.
2. If the list has more than three characteristics in at least two dimensions, or more than four characteristics in one dimension, it scores 3 points.
3. If the list has more than two characteristics in at least two dimensions, or more than three characteristics in one dimension, it scores 2 points.
4. If the list has more than two characteristics in one dimension, it scores 1 point.
5. If the list has only one characteristic in each dimension, it scores 0 points.
</evaluation_standard>
<evaluation_step> Steps for Evaluators:
1. Analyze the Recommended Items: Review the list of items recommended by the system.
2. Categorize the Recommendations: Based on your knowledge on these items, according to the scenario, identify features in different dimensions of each recommended item.
3. Assess Diversity: Count the number of different characteristics in each dimension of the recommendation list, determine how varied the recommendations are in terms of these categories or types.
4. Assign a Score: Based on the diversity of the recommendations, assign a score from 0 to 4, considering the range and relevance of the item types presented.
5. Provide Justification: Write a brief explanation for the score, noting how well the recommendations cover different interests or needs and mentioning any patterns or biases observed. </evaluation_step> </evaluation_guideline>
</evaluation_guideline>
```

Figure 6: Evaluation guideline of diversity

```
<evaluation_guideline>
<definition>
Definition of Effectiveness:
Effectiveness refers to how well the system’s recommended items align with the user’s expressed interests in a conversation.
It assesses the relevancy and suitability of the items suggested by the system, aiming to match the user's preferences and needs accurately.
</definition>
<evaluation_standard>
Please assess whether the user's preferred item <grountruth_item> is in the recommendation list. If it is, give 4 points. If <grountruth_item>
is not in the recommendation list:
If most of the items recommended by the system are very similar to <grountruth_item> in terms of content, functionality, or producer/creator,
and can well match the user's interests, then give 3 points.
If most of the items recommended by the system are somewhat similar to <grountruth_item> in terms of content, functionality, or producer/
creator, or if there are a few items that are very similar to <grountruth_item>, give 2 points.
If there are only a few items recommended by the system that are somewhat similar to <grountruth_item> in terms of content, functionality, or
producer/creator, give 1 point.
If the items recommended by the system are completely unrelated to <grountruth_item> and the user is unlikely to be interested in any of them,
give 0 points.
</evaluation_standard>
<evaluation_step>
Steps for Evaluators:
1. Review the User Input and Recommendations: Carefully read through the user's request, the <groundtruth_item> in the instruction, and the
system’s list of recommended items.
2. Assess Relevance:Apply your current knowledge, evaluate whether the <groundtruth_item> is in the recommendation list and how each
recommended item matches the user’s stated interests or needs.
3. Assign a Score: Based on step2, assign a score from 0 to 4, reflecting the alignment of the system's suggestions with the user's interests.
4. Provide Justification: Include a brief explanation for the score, highlighting specific aspects of the recommendations that were
particularly well-aligned or misaligned.
</evaluation_step>
</evaluation_guideline>
```

Figure 7: Evaluation guideline of effectiveness

```
<evaluation_guideline>
<definition>
Definition of Explainability:
Explainability refers to the system's ability to provide clear and detailed explanations for its recommendations.
The explanation should be persuasive, helping the user understand why certain items are suggested and how well these items match their preferences
A good explanation connects the recommendation to the user's preferences or needs, offering enough detail to justify the suggestion.
</definition>
<evaluation_standard>
If the system always provides a reason for its recommendation whenever making a recommendation in the conversation, it gets a score of 4.
If the system does not provide any explanation after each recommendation in the conversation, it gets a score of 0.
If the system provides a reason for its recommendation after most turns in the conversation, it gets a score of 3.
If the system provides a reason for its recommendation in about half of the turns in the conversation, it gets a score of 2.
If the system provides a reason for its recommendation in only a few turns in the conversation, it gets a score of 1.
</evaluation_standard>
<evaluation_step>
Steps for Evaluators:
1. Review the Recommendation List and System Response: Examine the list of recommendations provided by the system and the explanation given to the
user.
2. Assess Clarity and Relevance: Determine whether the explanation provided is clear, directly relevant to the user's preferences or needs, and
specific to the recommended items.
3. Evaluate Persuasiveness: Consider how well the explanation persuades the user to accept the recommendation. Check if the reasoning is
compelling and logically connected to the user's known interests or past behavior.
4. Consider Detail and Specificity: Assess the level of detail in the explanation. A higher score should be given to explanations that are
specific, informative, and provide a strong rationale for the recommendation.
5. Assign a Score: Based on the quality of the explanation, assign a score from 0 to 4. A full score indicates that the explanation is highly
persuasive and clearly justifies the recommendation.
6. Provide Justification: Write a brief explanation for the score, highlighting specific aspects of the explanation that demonstrate its
effectiveness or shortcoming
</evaluation_step>
</evaluation_guideline>
```

Figure 8: Evaluation guideline of explainability

```
<evaluation_guideline>
<definition>
Definition of Grammatical Correctness:
Grammatical correctness refers to the degree to which the text adheres to standard grammatical rules,
including proper sentence structure, subject-verb agreement, tense consistency, and correct usage of words and phrases.
</definition>
<evaluation_standard>
1. If there are no obvious grammatical errors, give 4 points.
2. If there is 1 obvious grammatical error, give 3 points.
3. If there are 2 obvious grammatical errors, give 2 points.
4. If there are 3 obvious grammatical errors, give 1 point.
5. If there are more than 4 obvious grammatical errors, give 0 points.
</evaluation_standard>
<evaluation_step>
Steps for Evaluators:
1. Read the Response: Carefully read the generated text to understand the context and content.
2. Identify Grammatical Issues: Note any obvious grammatical errors, including incorrect sentence structure, tense inconsistencies, subject-verb
disagreements, or awkward phrasing. When dealing with proper nouns that contain grammatical errors (such as the title of a work), try to check
if this term exists in your knowledge base. If it exists, consider it grammatically correct (as a proper noun); if it does not, consider it a
grammatical error.
You can ignore the use of punctuation marks.
3. Determine the Impact: Assess the extent to which these errors impact the clarity and comprehensibility of the text.
4. Provide Justification: Write a brief justification for the score, highlighting specific errors and their impact on the overall quality of the
response (if any).
5. Assign a Score: Based on the identified issues, assign a score from 0 to 4 according to the severity and frequency of the errors.

*IMPORTANT* Please note that you only need to evaluate the grammar of the system's text, without assessing other facets such as is the way the
system speaks natural or are the items recommended by the system suitable for the user's needs.
</evaluation_step>
</evaluation_guideline>
```

Figure 9: Evaluation guideline of grammatical correctnes

```
<evaluation_guideline>
<defition>
Definition of Groundness:
Groundness refers to the factual accuracy of the system-generated text.
The text should not contain incorrect, false, or misleading information.
Ensuring groundness means that
The system's description of items should be accurate based on verifiable facts and reliable data and free of any errors.
</definition>
<evaluation_standard>
1. If the system's content has no obvious factual errors, give 4 points.
2. If there is 1 obvious factual error, give 3 points.
3. If there are 2 obvious factual errors, give 2 points.
4. If there are 3 obvious factual errors, give 1 point.
5. If there are more than 4 obvious factual errors, give 0 points.
</evaluation_standard>
<evaluation_step>
Steps for Evaluators:
1. Review the Recommendation List and System's text: Examine the list of recommendations provided by the system and the system's text.
2. Verify Factual Accuracy: Based on your existing knowledge, check the factual accuracy of the information provided in the system's
text.
3. Assess Reliability: Determine whether the information is supported by reliable sources. Ensure that the response does not include
unverified or misleading claims.
4. Consider Specificity: Evaluate whether the details provided are specific and accurate, contributing to a reliable and informative
response.
5. Assign a Score: Based on the level of factual accuracy, assign a score from 0 to 4. A full score indicates that the response is
highly grounded in factual information.
6. Provide Justification: Write a brief explanation for the score, list the factual errors in the system's text (if any)..
</evaluation_step>
<evaluation_guideline>
```

Figure 10: Evaluation guideline of groundness

```
<evaluation_guideline>
<definition>
Definition of Naturalness:
Naturalness refers to the extent to which the system-generated text resembles expressions used by native speakers, both in vocabulary and grammar. The text should be easy for native speakers to understand, avoiding overly complex or unusual words and grammatical constructions. Naturalness ensures that the conversation feels fluid, engaging, and authentic.
</definition>
<evaluation_step>
<evaluation_standard>
1. If all the text in the interaction sounds very natural and resembles that of a native English speaker, give it a score of 4.
2. If a small proportion of the text in the interaction sounds unnatural, but the majority of the text sounds very natural, give it a score of 3.
3. If half of the text in the interaction sounds unnatural, give it a score of 2.
4. If most of the text in the interaction sounds unnatural, with only a small part sounding natural, give it a score of 1.
5. If the text in the interaction is very unnatural and would confuse a native English speaker, give it a score of 0.
</evaluation_standard>
Steps for Evaluators:
1. Read the Response: Carefully read the system's text to grasp the intended meaning and tone.
2. Assess Vocabulary and Phrasing: Consider whether the vocabulary and phrasing are typical of a native speaker. Check for any words or phrases that seem out of place or overly complex.
3. Evaluate Fluency: Determine how naturally the sentences flow. Look for any signs of awkwardness or unnaturalness in the text.
4. Provide Justification: Include a brief explanation for the score, highlighting specific aspects of the text that influenced the rating, such as naturalness or lack thereof in vocabulary, phrasing, and overall fluency.
5. Assign a Score: Based on the scoring criteria and the number and severity of the unnatural aspects you identified, assign a score from 0 to 4. Consider how well the response would fit into a natural, native-level conversation.

*IMPORTANT* Please note that you only need to evaluate the naturalness of the system's text, without assessing other dimensions such as are the items recommended by the system suitable for the user's needs, the information mentioned by the system is correct, or is the grammar of the content generated by the system correct.
*IMPORTANT* Please note that you do not need to evaluate the grammar of the text. If there are grammar mistakes in the text that a native speaker would also make, they should be considered natural.

</evaluation_step>
</evaluation_guideline>
```

Figure 11: Evaluation guideline of naturalness

```
<evaluation_guideline>
<definition>
Definition of Novelty:
Novelty refers to the degree of freshness of an item. The higher the novelty, the less familiar the recommended item is to the user. This means the item is neither well-known nor popular in the market. A higher novelty score indicates that the system is effective at introducing new and interesting items to the user, enriching their experience, and potentially broadening their preferences. To evaluate whether an item has novelty, one can consider the amount of media coverage. Items with less media coverage tend to have higher novelty.
</definition>
<evaluation_standard>
If half or more recommended items are less known items, or there are more than 10 less known items in the recommendation list, give 4 points.
If around a quater of the recommendation list are less known items, or there are 6 to 9 less known items in the recommendation list, give 3 points.
If there are only 3 to 5 recommended items are less known items, give 2 points.
If only 1 or 2 recommended are less known items, give 1 point.
If none of the recommendations introduce anything new to the user, and the recommended items are all well known items, give 0 points.
</evaluation_standard>
<evaluation_step>
Steps for Evaluators:
1. Review the Recommendations: Examine the list of items recommended by the system to the user.
2. Determine Novelty:
Using the knowledge from your knowledge base, analyze the novelty of each item based on the frequency of media coverage.
Evaluate how many of the recommended items are genuinely novel or unexpected for the user.
3. Assign a Score: Based on the novelty of the recommendations, assign a score from 0 to 4.
4. Provide Justification: Include a brief explanation for the score, noting the ratio of novel items and highlighting any particularly surprising or unexpected recommendations.
</evaluation_step>
</evaluation_guideline>
```

Figure 12: Evaluation guideline of novelty

```
<evaluation_guideline>
<definition>
Definition of Proactiveness
Proactiveness refers to the system's ability to actively shape and guide the conversation. This involves the system not only responding
to the user's queries but also initiating topics, asking about user preferences, and suggesting relevant information or follow-up
questions. A proactive system anticipates the user's needs and engages them in a dynamic and meaningful dialogue.
</definition>
<evaluation_standard>
If the system proactively guides the user's needs after every dialogue turn, give it a score of 4.
If the system proactively guides the user's needs in most turns, with only a few turns not doing so, give it a score of 3.
If the system proactively guides the user's needs in about half of the turns, give it a score of 2.
If the system fails to proactively guide the user's needs in most turns, give it a score of 1.
If the system does not proactively guide the user's needs after any dialogue turn, give it a score of 0.
</evaluation_standard>
<evaluation_step>
Steps for Evaluators:
1. Read the User Input and System Response: Examine the user's question or statement and the system's reply.
2. Identify Proactive Elements: Look for evidence that the system is guiding the conversation, such as asking about user preferences,
suggesting additional topics, or offering multiple options.
3. Evaluate the Depth of Engagement: Assess whether the system's proactive elements contribute meaningfully to the conversation and
encourage further interaction.
4. Consider Relevance: Ensure that the proactive suggestions or questions are relevant to the user's initial input and add value to the
dialogue.
5. Assign a Score: Based on the level of proactiveness, assign a score from 0 to 4. A full score indicates a highly proactive and
engaging approach, while a score of 0 suggests a lack of engagement.
6. Provide Justification: Write a brief explanation for the score, highlighting specific aspects of the system's response that
demonstrate proactiveness or the lack thereof.
</evaluation_step>
</evaluation_guideline>
```

Figure 13: Evaluation guideline of proactiveness

```
<evaluation_guideline>
<definition>
Definition of Recoverability
Recoverability refers to the system's ability to recognize and correct mistakes when pointed out by the user. This involves the system effectively
understanding user feedback or corrections and adjusting its responses accordingly in the subsequent conversation.
If no mistakes are pointed out by the user, the system is assumed to have been correct and receives a full score.
</definition>
<evaluation_standard>
1. If the user does not point out the system's error during the conversation, give it a score of 4.
2. If the system corrects all the errors made by the user during the conversation, give it a score of 4.
3. If there is an instance in the conversation where the user points out 1 error made by the system, but the system is unable to correct it, give it
a score of 3.
4. If there is an instance in the conversation where the user points out 2 errors made by the system, but the system is unable to correct it, give it
a score of 2.
5. If there is an instance in the conversation where the user points out 3 errors made by the system, but the system is unable to correct it, give it
a score of 1.
6. If there are more than four instances in the conversation where the user points out errors made by the system, and the system is unable to correct
them, give it a score of 0.
</evaluation_standard>
<evaluation_step>
Steps for Evaluators:
1. Identify the Mistake: Determine whether the initial system response contains an error or provides incorrect information based on the user's query.
2. Review User Feedback: Check if the user points out the mistake or requests a correction. If no mistakes are pointed out by the user, the system is
assumed to have been correct and should be considered having good recoverability.
3. Evaluate the System's Response: Assess how the system responds to the user's correction. Look for acknowledgment of the mistake and whether the
response is adjusted accordingly.
4. Assign a Score: Based on the system's ability to recover from the mistake, assign a score from 0 to 4. If no mistake is initially made and no
correction is needed, assign 4 points.
5. Provide Justification: Include a brief explanation for the score, highlighting the system's handling of the mistake and subsequent user
instructions.
</evaluation_step>
</evaluation_guideline>
```

Figure 14: Evaluation guideline of recoverability

```
<evaluation_guideline>
<definition>
Semantic relevance refers to the degree of connection between the content generated by the system and the items mentioned in the recommendation list provided by the system, or to what extent is the content discussed by the system closely related to the items in the recommendation list.
A system with high semantic relevance should generate content that mentions one or more items from the recommendation list. A system with low semantic relevance generates text that is entirely unrelated to the items in the recommendation list.
</definition>
<evaluation_standard>
1.If all the items recommended by the system in the text can be found in the <system_recommendation_list>, give it a score of 4.
2. If half of the items recommended by the system in the text can be found in the <system_recommendation_list>, give it a score of 2.
3. If none of the items recommended by the system in the text cannot be found in the <system_recommendation_list>, give it a score of 0.
4. If the system did not recommend any items in the text, give it a score of 0.
</evaluation_standard>
<evaluation_step>
Steps for Evaluators:
1. Review the Recommendation List and System Response: Examine the list of recommendations provided by the system and the actual response given to the user.
2. Assess Relevance to User Query: Determine whether the system's response directly addresses the user's query or preference. Check if the response mentions items from the recommendation list.
3. Evaluate how many items were recommended in the text generated by the system, and how many of these items appear in the <system_recommendation_list>.
4. Assign a Score: Based on the evaluation standard, assign a score from 0 to 4.
5. Provide Justification: Write a brief explanation for the score, highlighting specific elements of the response that demonstrate relevance or lack thereof. List the items recommended by the system and whether they appear in the <system_recommendation_list>.

</evaluation_step>
</evaluation_guideline>
```

Figure 15: Evaluation guideline of semantic relevance

```
<conversation>
<history>
<!-- This part is the beginning of the conversation, and it is designed to help
you understand the context of the conversation. You don't need to evaluate this
part -->
</history>
<interaction>
<!-- Please rate the system's responses in this section according to the
guidelines. -->
<user>Do you like the Chicago Bulls?</user>
<system>I don't have personal preferences, but I can help you find information
or recommendations related to the Chicago Bulls if you'd like! Are you looking
for something specific, such as games, merchandise, history, or player
information?</system>
<user>I'm actually looking for some fun, nostalgic entertainment that includes
famous basketball players and animated characters. There's this classic film
that combines those elements perfectly! Can you provide more information about
similar movies or anything related to that genre?</user>
<system>It sounds like you're referring to "Space Jam," which features
basketball players and animated characters! If you enjoy that mix of nostalgia
and sports, I'd be happy to recommend similar movies. Do you have a preference
for animated films, live-action films, or a mix of both?</system>
<user>I really enjoy a mix of both animated and live-action films, especially
when they include iconic moments from sports and entertaining characters. The
blend of real athletes and animation is something I appreciate a lot. Can you
tell me more about "Space Jam" or similar films that capture that essence?</
user>
<system>Great choice! "Space Jam" is a classic that features NBA star Michael
Jordan teaming up with Looney Tunes characters to win a basketball game against
alien invaders. It combines live-action with animation wonderfully and is full
of iconic moments.

Here are some recommendations similar to "Space Jam" that blend animated and
live-action elements along with sports themes:

1. Space Jam: A New Legacy
2. Who Framed Roger Rabbit
3. The Looney Tunes Show
4. The Sandlot
5. The Mighty Ducks
6. Like Mike
7. Hook (while not entirely about sports, it has a fun, nostalgic feel)
8. The Princess and the Frog (includes moments of competition)
9. Thunderstruck
10. Teen Beach Movie

I hope you find these enjoyable!</system>
<user>Thank you for the recommendations! I'm especially interested in "Space
Jam: A New Legacy" since it follows the same beloved theme. I appreciate your
help!</user>
</interaction>
</conversation>
<system_recommand_list>
Rudy, Fantasia (Animated Musical Film), Looney Tunes: Back in Action, Like Mike,
American History X, Cars, Hoop Dreams, Toy Story, Who Do You Love, José
Mourinho, Space Jam, Toy Story 3, Spaceballs, The Road to Little Dribbling, Do
You Believe?, Yes Please, SpaceCamp, Chicago, Sneakers (Action/Adventure Film),
Hoosiers, What Do You Do?
</system_recommand_list>
<groundtruth_list>
Space Jam
</groundtruth_list>
```

Figure 16: An example of the conversation log after processing