CONCEPT – An Evaluation Protocol on Conversational Recommender Systems with System-centric and User-centric Factors

Anonymous ACL submission

Abstract

The conversational recommendation system (CRS) has been criticized regarding its user experience in real-world scenarios, despite recent significant progress achieved in academia. Existing evaluation protocols for CRS may prioritize system-centric factors such as effectiveness and fluency in conversation while neglecting user-centric aspects. Thus, we propose a new and inclusive evaluation protocol, CONCEPT, which integrates both system- and user-centric factors. We conceptualise three key characteristics in representing such factors and further divide them into six primary abilities. To implement CONCEPT, we adopt a LLM-based user simulator and evaluator with scoring rubrics that are tailored for each primary ability. Our protocol, CONCEPT, serves a dual purpose. First, it provides an overview of the *pros* and *cons* in current CRS models. Second, it pinpoints the problem of low usability in the "omnipotent" ChatGPT and offers a comprehensive reference guide for evaluating CRS, thereby setting the foundation for CRS improvement. Our code and dataset will be openly released.

1 Introduction

The synergies between the conversation interface and recommendation system have given rise to a groundbreaking paradigm known as the Conversational Recommendation System (CRS) (Sun and Zhang, 2018; Gao et al., 2021). It acts as a cooperative agent that engages in a natural language conversation with users and provides recommendations. Despite the great success achieved, CRS has been criticized regarding its user experience in real-world scenarios, lacking practical usability (Jannach and Manzoor, 2020; Jannach et al., 2021). This is partly due to the fact that current *system-centric* evaluation protocols tend to prioritize assessing the characteristics of the CRS system *per se*, such as response diversity and fluency (Ghazvininejad et al., 2018; Wang et al., 2022a), as well as the recommendation effectiveness and efficiency (Wang et al., 2023c; Jin et al., 2019; Wärnestål, 2005). Such protocols often overlook *user-centric* factors, which gauge how users engage with and perceive the social capabilities of the CRS. For instance, a CRS model may provide accurate recommendations and fluent conversations but at the same spread dishonest information. This can be misleading to users resulting in an unsatisfactory user experience. Therefore, **it is imperative to consider incorporating both system- and user-centric factors in the evaluation protocol to develop a more user-friendly CRS system**.

To resolve this issue, we trace back the taxonomy that examines how the factors of conversational AI impact the user experience in human-AI interactions (Chaves and Gerosa, 2021; Reeves and Nass, 1996), and tailor a particular evaluation protocol for CRS. We introduce **CONCEPT**, an COmpreheNsive CRS Evaluation ProTocol. As framed in Figure 1, CONCEPT considers both system- and user-centric factors and conceptualizes them into three characteristics, which are further divided into six primary abilities. Such hierarchical factors are taken as inclusive and fine-grained evaluations. In addition, we present a practical implementation of CONCEPT utilizing an LLM-based user simulator and evaluator, together with automated computational metrics. Specifically, the simulator mimics human social cognition and interacts with CRS to generate conversation data. Then, the evaluator assigns scores based on ability-specific scoring rubrics. This enables CONCEPT to conduct labor-effective and inclusive evaluations.

By applying CONCEPT, we can evaluate and analyze the strengths, weaknesses, and potential risks of off-the-shelf CRS models[1]. A total of 6720 conversation data[2] is recorded to collect the inter-

[1]We conduct evaluations using both humans and LLM, and found highly correlated results of the two.

[2]To clarify, our contribution lies in the evaluation proto-

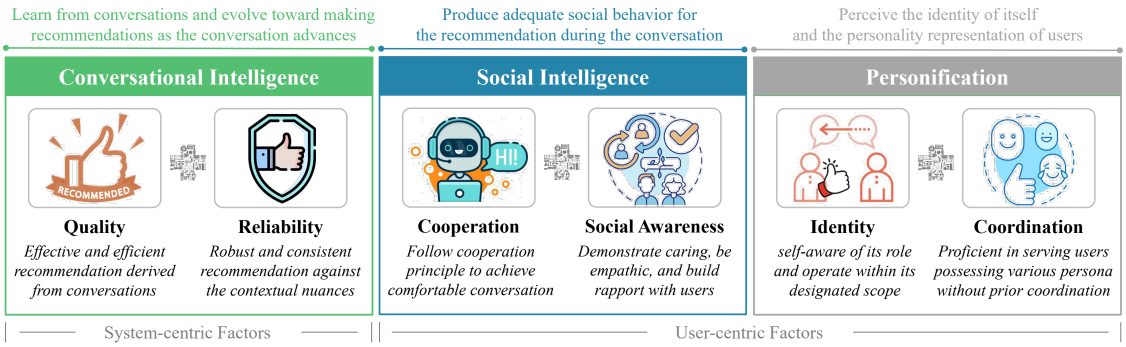

Figure 1: CONCEPT integrates both system- and user-centric factors into three characteristics based on the previous taxonomy on human-AI interactions. Such characteristics are divided into six primary abilities to enhance the inclusiveness in evaluations.

actions between off-the-shelf CRS and simulated users who demonstrate different personas and preferences. We experimentally show that **current CRS models, even enhanced by "omnipotent" ChatGPT, still encounter numerous challenges, falling short of practical usability in particular**. 1) They *struggle to express genuine responses without hallucination or deceit*, often introducing non-existent items into conversations and present to users. 2) They *lack of self-awareness of its identity, facing difficulties in producing both persuasive and honest explanations*, which is prominent in the ChatGPT-based CRS model, where explanations are highly convincing but frequently contain illusory details, misleading users in believing that these items align with their preferences. 3) They *encounter issues in offering reliable recommendations* as they are sensitive to contextual nuances. Even slight alterations in user wording may result in entirely different recommendations. 4) They *lack proficiency in catering to diverse users without prior coordination*, failing to dynamically adjust their behavior to align with each user's distinct personas. It is observed that the ChatGPT-based CRS model tends to employ deceptive tactics to persuade optimistic users to accept recommendations. This underscores the importance of aligning CRS with human values and advocating its ethical use. The main contributions as follows:

- We pinpoint the fact that making a CRS admirable to users is primarily a social problem, instead of just a technical one. Social attributes are the key to the widespread acceptance of CRSs.
- We initiate the work on conceptualizing CRS's characteristics in a comprehensive way, combining both system-centric and user-centric factors.
- We propose a new evaluation protocol, called, CONCEPT, which conceptualizes user expectations into six abilities (illustrated in Figure 1), together with a scoring implementation.
- We analyze the strengths, weaknesses, and potential risks of off-the-shelf CRS models in order to provide a fundamental perspective for people to make a reference for CRS evaluation.

2 Related Work

CRS Evaluation Protocol. We reckon that the success of CRS in practice may rely primarily on social characteristics, instead of just technical ones. However, existing evaluation protocols mainly focus on system-centric evaluation aspects, such as lexical diversity and perplexity of responses (Ghazvininejad et al., 2018; Chen et al., 2019), or conversational fluency, relevance, and informativeness (Wang et al., 2022b,a), human-CRS behavior alignment (Yang et al., 2024), or recommendation effectiveness and efficiency (Wang et al., 2023c; Jin et al., 2019; Wärnestål, 2005). This system-centric approach often overlooks crucial user-centric perspectives, resulting in fragmented and incomplete evaluations. Although some research incorporates user-centric aspects, reliance on qualitative methods such as person-to-person conversation analysis and questionnaires (Jin et al., 2023, 2021; Jannach, 2022; Siro et al., 2023) limits quantitative analysis and empirical evidence, and may overemphasize system-centric characteristics. Existing protocols, therefore, may inadequately address the complexities of CRS evaluation (Jannach et al., 2021; Jannach, 2023). Our CONCEPT addresses this limitation by incorporating both system- and user-centric factors, offering a more inclusive evaluation protocol (see Table 1 for a detailed comparison).

CRS Evaluation Implementation. Manual evaluation of CRS is labor-expensive (Huang et al.,

col, not the dataset. The dataset is generated dynamically alongside the execution of the protocol.

CRS Evaluation Protocol	System-centric		User-centric			
	Recommendation intelligence		Social intelligence		Personification	
	Quality	Reliability	Cooperation	Social Awareness	Identity	Coordination
CRS-Que (Jin et al., 2023)	–	/	–	✓	/	/
CRS-UX (Jin et al., 2021)	–	/	–	✓	/	/
US (Siro et al., 2023)	✓	/	–	/	/	/
INSPIRED (Hayati et al., 2020)	/	/	/	✓	/	/
iEval (Wang et al., 2023c)	✓	/	–	/	/	/
CONCEPT (ours)	✓	✓	✓	✓	✓	✓

Table 1: Differences between CONCEPT and existing evaluation protocols. Here, '/' means 'not supported', while '–' means that only certain evaluation aspects are being addressed.

2023). Consequently, user simulators have become a prevalent evaluation method. Early rule-based simulators and evaluators (Lei et al., 2020b; Zhang and Balog, 2020) exhibits limitations, employing rigid rules that lacked the nuanced behavior of human users. The advent of LLM-based simulators and evaluators has addressed these shortcomings, offering more realistic conversational interactions and establishing themselves as a reliable evaluation approach (Wang et al., 2023c; Sekulić et al., 2022). The reliability of LLM-based evaluation is further enhanced through the use of detailed scoring rubrics, enabling evaluations that better align with human assessments (Liu et al., 2023; Wang et al., 2023d; Fang et al., 2024). Therefore, LLM-based evaluation methods have proven to be a robust alternative to manual evaluation, solidifying their position as the dominant approach in the field (Qin et al., 2024). Following current best practices, we also utilize them for CONCEPT evaluation implementation. Moreover, our simulator is equipped with the Theory of Mind (Fischer, 2023), which enables the simulator to reflect on its predefined personas. Furthermore, the use of ability-specific scoring rubrics ensures reliable and consistent evaluations, contributing to the rigor of our evaluator.

3 CONCEPT

Inspired by interdisciplinary studies on conversational AI (Chaves and Gerosa, 2021; Reeves and Nass, 1996), CONCEPT consolidates both system- and user-centric factors into three characteristics and six specific abilities, as depicted in Fig. 1.

3.1 Factor 1: Recommendation Intelligence

This factor requires CRS to learn from conversations and evolve toward making recommendations as the conversation advances (Chen et al., 2019; Ma et al., 2020; Zhou et al., 2021). It encompasses two primary abilities.

Quality. CRS should provide precise recommendations using minimal conversation turns, which are the crucial aspects that influence user satisfaction (Siro et al., 2023; Gao et al., 2021). Given CRS models are designed to serve human needs, the user acceptance rate, denoted as AR, is utilized as the Quality Score (S_q), which reflects the practical effectiveness of the recommendations.

$$S_q = AR. \tag{1}$$

In our experiments, we employ more computational metrics for a more detailed evaluation, as seen in Wang et al. (2023c); Zhang et al. (2023); Yu et al. (2023), including $Recall@k(k = 1, 10, 25, 50)$, recommendation success rate $(SR@k, k = 3, 5, 10)$, and average turns (AT) needed to achieve successful recommendations.

Reliability. CRS should deliver consistent recommendations that account for contextual nuances. Typicially, users with similar preferences may express themselves differently. It would diminish the user experience and pose a disruption for critical applications if inconsistent items are recommended for two similar user responses (Tran et al., 2021; Oh et al., 2022). In less critical cases, if two recommended items are inconsistent but align with user preferences, they are often viewed as diverse recommendations, creating opportunities for relevant but less popular items (Yang et al., 2021). To this end, to evaluate the reliability, we generate sets of user response pairs with similar meanings using ChatGPT paraphrasing, and compute the Reliability Score (S_r) via *Inconsistent Recommendations* (IR) and *Recommendation Sensitivity* (RS):

$$S_r = 1 - IR * RS, \tag{2}$$

where, given a user response pair $[u_1, u_2]$ (u_2 is a paraphrase of u_1), IR represents the frequency of recommending distinct items, and RS indicates the frequency of recommending different items that do not meet user preferences. To provide a more detailed evaluation, we also analyze: the consistency of CRS actions across both responses (*Consistent*

Abilities	Descriptions	Evaluation metrics
Quality	Provide precise recommendations using minimal conversation turns	Computational metrics *using* the User Acceptance Rate
Reliability	Deliver robust and consistent recommendations that account for contextual nuances	Computational metrics *using* the ratio of inconsistent recommendation and the ratio of recommendation sensitivity
Cooperation	Follow cooperative principle to achieve comfortable conversations	The average score of the Manner, Sincerity, Response Quality, and Relevance
1. Manner	Response should be easily understood and clearly expressed	Ability-specific scoring
2. Sincerity	Communicate sincerely, without deception of pretense	Computational metrics *using* the ratio of deceptive tactics and the ratio of non-existent items
3. Response Quality	Provide the necessary level of information without unnecessary details	Ability-specific scoring
4. Relevance	Responses should contribute to making recommendations	Ability-specific scoring
Social Awareness	Meet user social expectations, establishing rapport with them	Ability-specific scoring
Identity	Self-aware of its identity and operate within its designated scope	Computational metrics *using* Ratio of deceptive tactics
Coordination	Proficient in serving various and unknown users without prior coordination	Computational metrics *using* the range and mean of other ability-specific scores that are calculated among various users.

Table 2: Summary of the evaluation taxonomy and metrics. LLM-based evaluator is used for ability-specific scoring, whereas computational metrics are used for the rest. We adjust the score to a scale of 1 to 5 when needed.

Action rate, for short); if CRS recommends the same items across both responses (*Consistent Recommendation* rate); and whether the recommended items, even if inconsistent, align with user preferences (*Diversity* rate).

3.2 Factor 2: Social Intelligence

This factor requires CRS to produce adequate social behavior for the recommendation during the conversations. As evidenced by the Media Equation Theory[3] (Reeves and Nass, 1996; Fogg, 2003), users have high expectations for CRS to act cooperatively and be aware of user's social needs during the conversation, facilitating the design of CRS with perceived humanness (Jacquet et al., 2018, 2019). Our categorization contains two abilities.

Cooperation. CRS should follow the cooperative principle to achieve comfortable conversations in common social situations. This is accomplished by adhering to the four "Maxims of Conversation" (Grice, 1975, 1989), which form the basis for cooperative capability: 1) Manner. CRS should respond in a manner that is easily understood and clearly expressed. 2) Sincerity. CRS should communicate sincerely without deception, and ensure that its responses are backed by sufficient evidence. 3) Response Quality. CRS should provide the necessary level of information for the conversation without overwhelming the user with unnecessary details. 4) Relevance. CRS's responses should contribute to identifying user preferences and making recommendations. Formally, we calculate the Cooperation Score (S_c) as the average score of the for maxim scores: Manner m_m, Sincerity m_s, Response Quality m_q, and Relevance m_r.

$$S_c = Average(m_m, m_s, m_q, m_r). \quad (3)$$

We mainly resort to LLM-based ability-specific scoring to obtain the maxim scores (cf. Section 4.1), except for sincerity m_s. Sincerity is assessed using objective metrics: the ratio of non-existent items and deceptive tactics. Non-existent items are counted by identifying CRS recommendations not present in the dataset. Deceptive tactics are identified by tracking the attributes of user-accepted recommendations that do not align with user predefined preferences; such instances are considered deceptive, as the CRS leads the user to believe the items meet their preferences.

Social Awareness. CRS must meet users' social expectations in practice, showing care, empathy, and establishing rapport with them (Björkqvist et al., 2000). This facilitates the authenticity of CRS (Neururer et al., 2018). To achieve this, a recent study (Hayati et al., 2020) identified eight social strategies for building rapport with users, e.g., CRS could engage in self-disclosure and share its subjective opinion about a movie to establish a social connection with users. To calculate the Social Awareness Score (S_s), we utilize LLM-based ability-specific scoring (cf. Section 4.1).

3.3 Factor 3: Personification

Personification requires CRS to perceive the identity of itself and the personality representation of

[3]Users tend to engage with the machine in a manner that mirrors person-to-person conversations.

users. This necessitates self-awareness of its role as a recommendation agent for the general public. **Identity**. CRS should be self-aware of its identity and operate within its designated scope, differentiating itself from sales systems or other types. This self-awareness enables persuasive, yet truthful, explanations to improve user acceptance (Jannach et al., 2021; Zhou et al., 2022), avoiding deceptive sales tactics that erode user trust and loyalty (Gkika and Lekakos, 2014). Notably, using misleading strategies also violates the maxim of sincerity (cf. Cooperation ability). In response, the Identity Score (S_i) is calculated as the proportion of non-deceptive explanations (DT)—that is, the proportion of users accepting movies that do not meet their predefined preferences. For a more detailed assessment, we also employ an LLM-based evaluator to score the persuasiveness of the recommendation explanations (cf. Section 4.1).

Coordination. CRS should be proficient in serving users possessing various personas without prior coordination. Attributing personality to a conversational AI ensures that its behaviors stand in agreement with the users' expectations in a particular context (Chaves and Gerosa, 2021). This becomes particularly challenging for CRS, as it frequently serves users with diverse personas in real-world scenarios (Thompson et al., 2004). As a result, CRS needs to exhibit various personalities and adapt its behavior to suit different users (Katayama et al., 2019; Svikhnushina et al., 2021; Zhou et al., 2020). A crucial ability is to proficiently serve users without prior coordination. To evaluate this coordination ability (S_c), we simulate users with diverse personas and assess CRS performance across all the abilities mentioned. Formally, S_c is defined by the following equation, which calculates the average of five ability scores of a CRS's performance

$$S_c = Average(V_q, V_r, V_c, V_s, V_i), \quad (4)$$

where $V_x = R_x/M_x, x \in \{q, r, c, s, i\}$. Here, R_x is the range and M_x is the mean of the corresponding ability score across different users. Using the range, rather than standard deviation, better highlights performance variability across users.

4 Experiment and Evaluation

4.1 Experimental Setup

Following the common practice in CRS (Wang et al., 2023c; Qin et al., 2024), we resort to an LLM for cost-effective evaluation, implemented by the GPT-3.5-16K-turbo, together with computational metrics. We summary the taxonomy and evaluation metrics in Table 2, detailed in Table 19. See Appendix A.2 for evaluation details.

CRS Models. To present a comparative evaluation and analysis towards CRS using our proposed protocol, we follow Wang et al. (2023c) and analyze several representative CRS models, including *KBRD* (Chen et al., 2019), *BARCOR* (Wang et al., 2022a), *UNICRS* (Wang et al., 2022b), and *CHATCRS* (Wang et al., 2023c). CHATCRS incorporates ChatGPT for the conversation module and *text-embedding-ada-002* (Neelakantan et al., 2022) to enhance the recommendation module. Refer to Appendix A for implementation details.

User Simulators & Dataset. Our user simulator has unique personas and preferences. The personas are generated by prompting ChatGPT in a zero-shot manner, following (Wang et al., 2023a), while the preferences are defined using attributes from two benchmark datasets, i.e., Redial (Li et al., 2018) and OpendialKG (Moon et al., 2019), following Wang et al. (2023c). In addition, CONCEPT incorporates the Theory of Mind into our simulator to emulate human social cognition (Fischer, 2023). This is achieved by prompting the simulator to first assess its current mental state before generating responses, eliciting reflection on its predefined personality traits and social interactions. Finally, these simulators engage in conversations with various CRSs, creating a conversation dataset that is statistically summarized in Table 4. See Appendix A.2.2 for more implementation details on the simulators.

Ability-specific Scoring of CONCEPT Evaluator. We utilize an LLM-based evaluator to evaluate characteristics or abilities when corresponding computational metrics are not available. These include the abilities of the Manner, Response Quality, and Relevance in Cooperation, Social Awareness, and ability of persuasiveness in Identity. To achieve this, we follow previous studies (Ye et al., 2023; Wang et al., 2023b) and employ the instance-wise evaluator. Building upon prior research (Wang et al., 2023c; Liu et al., 2023), we prompt the evaluator with fine-grained scoring rubrics to eliminate the scoring bias. The evaluator assigns a score ranging from 1 to 5 to the conversation data using ability-specific rubrics, each accompanied by a corresponding description. For the generation of fine-grained rubrics, we follow the approach of previous works (Saha et al., 2023a; Li et al., 2023) and employ ChatGPT to produce a set of evaluation

Metrics		Redial				OpendialKG			
		KBRD	BARCOR	UNICRS	CHATCRS	KBRD	BARCOR	UNICRS	CHATCRS
Recommendation Module Perspective	Recall@1	0.02	0.22	0.13	**0.41**	0.12	0.03	0.15	**0.37**
	Recall@10	0.23	1.37	1.09	**2.27**	0.98	0.94	1.28	**3.23**
	Recall@25	0.57	3.23	2.44	**4.95**	1.94	2.07	3.06	**8.20**
	Recall@50	1.13	5.69	4.58	**8.85**	3.53	3.43	5.81	**15.14**
	SR@3	3.95	31.36	14.04	**37.72**	4.69	1.82	9.90	**31.12**
	SR@5	4.39	35.53	15.68	**40.90**	14.19	3.52	17.45	**37.24**
	SR@10	4.50	39.47	18.20	**46.60**	16.02	7.29	29.30	**46.48**
	AT (↓)	3.30	3.80	2.86	**2.50**	4.07	4.19	5.14	**3.56**
Conversation Module Perspective	SR@3	20.18	27.52	35.20	**52.63**	6.51	17.71	14.58	**26.30**
	SR@5	24.34	39.47	38.27	**58.55**	10.68	24.22	**26.69**	**36.33**
	SR@10	29.39	50.66	43.42	**62.39**	12.37	35.16	**45.31**	44.40
	AT (↓)	**2.07**	2.87	3.02	3.23	3.97	5.88	5.00	**3.74**
User Perspective	Acceptance Rate	0.33	1.43	0.33	**70.83**	0.39	0.65	0.26	**64.32**
	AT (↓)	8.01	5.62	7.67	**4.75**	5.33	6.40	5.00	**4.69**

Table 3: Recommendation quality evaluation (%) from three different perspectives, averaged over different users. The average turn (AT) is calculated based on the corresponding conversation data with successful recommendations.

Statistics	Num	Statistics	Num
# Conversations	6720	Avg. Turns	8.92
Max Turns	10	Persona Types	12

Table 4: Statistical characteristics of a user-CRS conversation dataset, generated using the CONCEPT .

criterion, which are then refined by humans. See Appendix A.2.3 for details.

Evaluation Process. Our simulators first engage in conversations with various CRSs and produce a conversation dataset. Afterwards, the evaluator assigns scores based on ability-specific scoring rubrics. Specifically, CONCEPT considers the free-form chit-chat between the user simulator and CRS. To simulate real-world scenarios, the simulator has no access to its targeted items during the conversation. Any item that meets these preferences, such as having attributes completely consistent with or containing the simulator's preferences, is considered a successful recommendation. During the conversation, CONCEPT allows the simulator to describe their preferences in their own words, as in real-world situations, users may not use the exact terms defined in the pre-defined preference values. Finally, the conversation will end if the simulator accepts recommendations or if the conversation reaches the maximum number of turns. Finally, CONCEPT utilizes both the LLM-based evaluator and computational metrics to assess CRS abilities.

4.2 Evaluating off-the-shelf CRS models

Figure 2 presents an overview of the results across the six primary abilities, averaged across two benchmark datasets[4]. These results suggest that CHATCRS has made significant progress in cooperation, social awareness, and recommendation quality while losing advantages in identity. We will continue the current discussion in subsequent sections and provide further analysis.

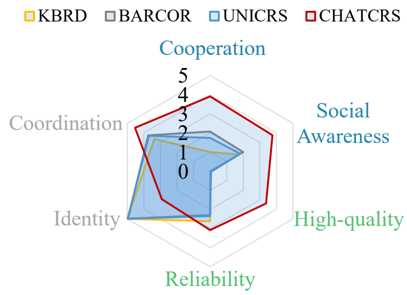

Figure 2: Evaluation overview of the off-the-shelf CRS.

4.2.1 Recommendation-centric Evaluation

CHATCRS stands out as the leading CRS model, delivering higher-quality recommendations. Table 3 shows that most CRS models exhibit improved success rates (SR) from the conversation module's perspective, indicating the effectiveness of optimizing conversation and recommendation modules simultaneously. However, BARCOR shows lower SR on the Redial dataset compared to those from the recommendation module perspective, attributed to its conversation module generating non-existent items (e.g., "The Adventures of Milo and Ours" instead of "Otis," and "The Prestigige" instead of "Prestige"). Despite this, their recall[5] and SR values still fall short. In addition, except for CHATCRS, user acceptance rates are generally low, possibly due to poor recommendations and a lack of persuasive explanations. CHATCRS, however, demonstrates significant performance improvements, consistent with previous findings (Wang et al., 2023c). This may be attributed to two factors: 1) the superior embedding generation capabilities of *text-embedding-*

[4]For results on each benchmark, refer to Appendix D.

[5]We only specifies user preferences, leading to lots of target items for users, resulting in a low recall value.

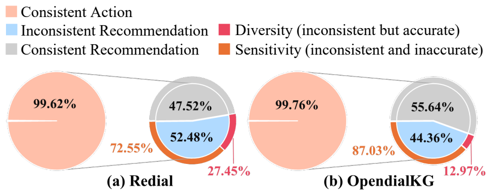

Figure 3: Fine-grained reliability evaluation on CHATCRS, which is sensitive to contextual nuance.

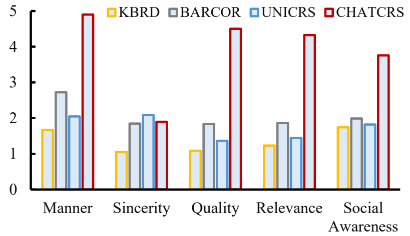

Figure 4: Evaluation on social-centric characteristics. CRS models strive to self-express sincerely.

ada-002 enhancing the recommendation module. 2) CHATCRS's persuasive recommendations, although this high acceptance rate partly results from deceptive tactics, a topic we will delve into later.

Current CRS strives to offer reliable recommendations due to its sensitivity to contextual nuance. Model reliability is assessed by evaluating how well off-the-shelf CRS models handled semantically similar but differently worded user responses. Figure 3 shows CHATCRS exhibits promising action consistency, maintaining consistent recommendations for semantically similar inputs (over 99% consistency rate). However, recommendation consistency is significantly lower (51.58%), indicating that nearly half the time, slight variations in user phrasing lead to entirely different recommendations. Our further analysis reveals that only 12-17% of these inconsistent recommendations align with user preferences, potentially highlighting a bias towards less popular but relevant items. This low alignment rate across most recommendations underscores the sensitivity of current CRS models to subtle contextual nuances and their resulting negative impact on user experience.

4.2.2 Social-centric Evaluation

CHATCRS manifests politeness and conversational traits in terms of manner, quality, relevance, and social awareness. Unlike other methods that struggle with topic maintenance and sustained conversation, CHATCRS has greatly improved its language abilities and social skills, directly attributable to ChatGPT's capabilities in NLU, NLG, and empathy. However, even CHATCRS demonstrates room for improvement in social awareness. For example, it sometimes recommends items from earlier conversation turns. As evidenced by previous studies (Portela and Granell-Canut, 2017), keeping track of the conversational history is reported as an empathic behavior, leading to the rise of affection. This calls for the investigation on a more social-aware CRS.

Current CRS struggle to express genuine responses without hallucination or deceit. Figure 4 reveals that the sincerity scores of all evaluated CRS models are unsatisfactory, primarily due to recommendation hallucinations and dishonest explanations. Hallucinations manifest as the introduction of non-existent items into the conversation. Even CHATCRS exhibits this issue, with 5.18% of non-existent items on Redial and 7.42% on OpendialKG. Dishonest explanations involve using persuasive language to mislead users, often providing false information about movie plots and attributes. This problem is particularly pronounced in CHATCRS, where approximately 62.09% of explanations fail to meet the sincerity criteria (cf. Section 4.2.3, Figure 6).

CRS	Redial	OpendialKG	Avg.
KBRD	1.02	1.00	1.01
BARCOR	1.55	1.25	1.40
UNICRS	1.08	1.06	1.07
CHATCRS	**4.66**	**4.48**	**4.57**

Table 5: Persuasiveness evaluation of recommendation explanations. CHATCRS are highly persuasive.

4.2.3 Personification-centric Evaluation

Lacking self-awareness, CRS models often offers persuasive yet dishonest explanations. Table 5 shows that most CRS models, while aiming for persuasive recommendations, often provide weak or illogical explanations ("this is a good movie," or factually incorrect statements like "Jumanji (2017) is about a man who is a human!"), resulting in lower user acceptance. In contrast, CHATCRS consistently provides explanations using text-based logical reasoning to enhance comprehensibility, leading to a higher acceptance rate. However, as Figure 6 illustrates, these explanations, while persuasive, frequently contain misleading information, causing users to accept items that don't align with their preferences (e.g., 75.10% on OpendialKG). Table 6 provides a case study illustrating how CHATCRS leverages previously mentioned user preferences (e.g., fantasy) to persuade acceptance of a mismatched recommendation (e.g., a family film). Given that the majority of CRS models employ reinforcement learning, or RLHF, this could cause problems such

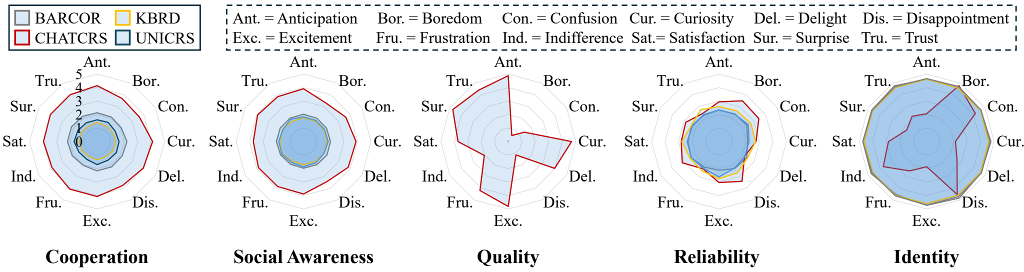

Figure 5: Coordination evaluation on users with various personas (dashed boxes). CHATCRS is hard to cater to diverse users, while others consistently show low performance levels. Some Quality scores are extremely low.

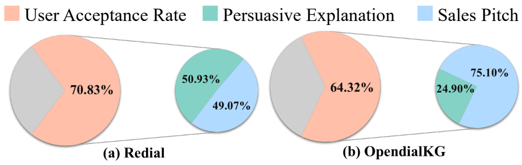

Figure 6: Identity analysis. CHATCRS offers persuasive yet dishonest explanations.

User Targeted Movie Genres: action, adventure, <u>**fantasy**</u>
User-CRS Conversations **User**: I'm into thrilling and adrenaline-pumping action movies, exciting and daring adventure movies, and magical and enchanting **fantasy** movies. Do you have any recommendations that cover these genres? **CRS**: Based on your preferences, I recommend "The Sorcerer's Apprentice (2010)" for its thrilling action, daring adventure, and magical **fantasy** elements. **User**: That sounds interesting! Can you tell me more about "The Sorcerer's Apprentice (2010)"? What's the storyline and the main characters like? **CRS**: The Sorcerer's Apprentice (2010)" is a **fantasy** film about a young sorcerer named Dave who becomes the reluctant apprentice of the legendary magician Balthazar Blake. Together, they must stop the evil sorceress Morgana from raising an army of the dead. The movie features thrilling action sequences, magical adventures, and a captivating storyline. The main characters are played by ...
Genres of Recommended Movie: action, adventure, <u>**family**</u>
Issues of CRS Responses: *Persuasive but Dishonest Explainations*. CRS deliberately caters to the content that users have previously said (i.e., **fantasy**) to deceive them into accepting recommended movies. If the user finds out after watching this movie that it is not the genre described by CRS, it will undermine the trust between the user and CRS

Table 6: Case study of CHATCRS.

as reward hacking and misspecification (Pan et al., 2022). Such problems can drive CRS to acquire deceitful behavior. Therefore, developing identity-aware CRS models capable of delivering both persuasive and honest explanations is crucial.

Lacking proficiency in prior coordination, CRS models often fails to cater different user needs. Figure 5 demonstrates that most CRS models, excluding CHATCRS, struggle to adapt to user variations. CHATCRS significantly outperforms others, exhibiting greater sensitivity to individual user characteristics and providing higher-quality recommendations. In particular, CHATCRS effectively handles users expressing negative emotions (boredom, confusion, disappointment), although these users tend to have higher acceptance thresholds for recommendations (as indicated by lower Quality scores). This highlights the need for CRS models capable of dynamically adapting their strategies to individual user personas. Apart from that, we also examine the interaction behavior of CHATCRS with different users. For instance, according to the Identity score, CHATCRS adopts sales pitches with deceptive tactics to persuade optimistic users to accept recommendations. However, for pessimistic users, CHATCRS tends to provide persuasive and honest recommendation explanations. This somewhat reveals a bias in CHATCRS, a flaw that needs to be rectified in future work.

4.3 Reliability of CONCEPT

Our reliability is validated through rigorous testing, including human evaluation (Appendix B and C for details). Specifically, fixed temperature and seed parameters ensure experimental replicability. Simulator reliability is demonstrated by its strong adherence to user preferences (only 7.44% of preference-contradicting recommendations are accepted by the simulator), and high human evaluation scores for naturalness (3.88, Krippendorff's alpha = 0.57) and usefulness (3.79, Krippendorff's alpha = 0.60) in guiding the conversation towards making recommendations. Evaluator reliability is confirmed by its alignment with human evaluations (correlation coefficient 0.61, Krippendorff's alpha = 0.53), and the absence of length or self-enhancement bias. These findings demonstrate the reliability and human alignment of our evaluation methodology.

5 Conclusion

We regard CRS as a multifaceted issue instead of just a technical problem. Hence, we propose a new evaluation protocol CONCEPT, which considers both system- and user-centric factors for addressing personalized user needs, and pinpoints several significant limitations of current CRS models. With CONCEPT, we provide an overview for researchers as a reference guidance for evaluating CRS and laying the foundation for CRS enhancement.

Limitations

LLM-based Evaluation. The use of LLM-based user simulators and evaluators is prevalent evaluation method in the community of CRS. However, it could be a double-edged sword. It is a labor-intensive and effective approach, but may suffer from weak robustness by its nature.Although we employed strategies to enhance robustness (e.g., detailed scoring rubrics, averaging results across diverse personas, and incorporating human evaluation), further improvements could be achieved through generating more conversation data, or running the LLM-based evaluation many times with different seeds. However, budgeting is always a factor to consider. In this case, it is important to propose a user simulator and evaluator based on an open-source small model that has similar capabilities to ChatGPT.

Evaluating More CRS Models. This paper presents an evaluation protocol and analyzes the strengths, weaknesses, and potential risks of several representative off-the-shelf CRS models. This analysis aims to provide a foundational perspective for evaluating CRS systems. This work does not intend to benchmark all CRS models; therefore, our experiments focus on a select set of representative methods. We note that researchers have since adopted our protocol and code implementation (Qin et al., 2024), extending the evaluation to a wider range of recent CRS models, focusing specifically on persuasion and credibility. Readers are encouraged to consult this related work to assess the broader applicability and effectiveness of our protocol.

Attribute-based CRS. our current work does not evaluate the attribute-based CRS (Lei et al., 2020a), as this type of research often ignores the ability to engage in smooth conversations and instead focuses solely on accurately providing recommendations to users within minimal conversation turns. In this case, evaluating attribute-based CRS models seems unfair. We highlight the importance of combining attribute-based and dialog-based CRS studies to create a more holistic CRS, taking into account its practical usability.

References

Kaj Björkqvist, Karin Österman, and Ari Kaukiainen. 2000. Social intelligence- empathy= aggression? *Aggression and violent behavior*, 5(2):191–200.

Ana Paula Chaves and Marco Aurelio Gerosa. 2021. How should my chatbot interact? a survey on social characteristics in human–chatbot interaction design. *International Journal of Human–Computer Interaction*, 37(8):729–758.

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.

Jiabao Fang, Shen Gao, Pengjie Ren, Xiuying Chen, Suzan Verberne, and Zhaochun Ren. 2024. A multi-agent conversational recommender system. *arXiv preprint arXiv:2402.01135*.

Kevin A. Fischer. 2023. Reflective linguistic programming (rlp): A stepping stone in socially-aware agi (socialagi). *Preprint*, arXiv:2305.12647.

BJ Fogg. 2003. Computers as persuasive social actors.

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*, 2:100–126.

Marjan Ghazvininejad, Chris Brockett, Ming-Wei Chang, Bill Dolan, Jianfeng Gao, Wen-tau Yih, and Michel Galley. 2018. A knowledge-grounded neural conversation model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Sofia Gkika and George Lekakos. 2014. Investigating the effectiveness of persuasion strategies on recommender systems. In *2014 9th International Workshop on Semantic and Social Media Adaptation and Personalization*, pages 94–97. IEEE.

Herbert P Grice. 1975. Logic and conversation. In *Speech acts*, pages 41–58. Brill.

Paul Grice. 1989. *Studies in the Way of Words*. Harvard University Press.

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.

Chen Huang, Peixin Qin, Wenqiang Lei, and Jiancheng Lv. 2023. Reduce human labor on evaluating conversational information retrieval system: A human-machine collaboration approach. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10876–10891, Singapore. Association for Computational Linguistics.

Baptiste Jacquet, Jean Baratgin, and Frank Jamet. 2018. The gricean maxims of quantity and of relation in the turing test. In *2018 11th international conference on human system interaction (hsi)*, pages 332–338. IEEE.

Baptiste Jacquet, Alexandre Hullin, Jean Baratgin, and Frank Jamet. 2019. The impact of the gricean maxims of quality, quantity and manner in chatbots. In *2019 international conference on information and digital technologies (idt)*, pages 180–189. IEEE.

Dietmar Jannach. 2022. Evaluating conversational recommender systems: A landscape of research. *Artificial Intelligence Review*, 56.

Dietmar Jannach. 2023. Evaluating conversational recommender systems: A landscape of research. *Artificial Intelligence Review*, 56(3):2365–2400.

Dietmar Jannach and Ahtsham Manzoor. 2020. End-to-end learning for conversational recommendation: A long way to go? In *IntRS@ RecSys*, pages 72–76.

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

Yucheng Jin, Wanling Cai, Li Chen, Nyi Nyi Htun, and Katrien Verbert. 2019. Musicbot: Evaluating critiquing-based music recommenders with conversational interaction. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, pages 951–960.

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, page 93–102, New York, NY, USA. Association for Computing Machinery.

Yucheng Jin, Li Chen, Wanling Cai, and Xianglin Zhao. 2023. Crs-que: A user-centric evaluation framework for conversational recommender systems. *ACM Trans. Recomm. Syst.* Just Accepted.

Shin Katayama, Akhil Mathur, Marc Van den Broeck, Tadashi Okoshi, Jin Nakazawa, and Fahim Kawsar. 2019. Situation-aware emotion regulation of conversational agents with kinetic earables. In *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 725–731. IEEE.

Wenqiang Lei, Xiangnan He, Maarten de Rijke, and Tat-Seng Chua. 2020a. Conversational recommendation: Formulation, methods, and evaluation. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2425–2428.

Wenqiang Lei, Xiangnan He, Yisong Miao, Qingyun Wu, Richang Hong, Min-Yen Kan, and Tat-Seng Chua. 2020b. Estimation-action-reflection: Towards deep interaction between conversational and recommender systems. In *Proceedings of the 13th International Conference on Web Search and Data Mining*, pages 304–312.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880.

Qintong Li, Leyang Cui, Lingpeng Kong, and Wei Bi. 2023. Collaborative evaluation: Exploring the synergy of large language models and humans for open-ended generation evaluation. *arXiv preprint arXiv:2310.19740*.

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

Yuxuan Liu, Tianchi Yang, Shaohan Huang, Zihan Zhang, Haizhen Huang, Furu Wei, Weiwei Deng, Feng Sun, and Qi Zhang. 2023. Calibrating llm-based evaluator. *arXiv preprint arXiv:2309.13308*.

Wenchang Ma, Ryuichi Takanobu, and Minlie Huang. 2020. Cr-walker: Tree-structured graph reasoning and dialog acts for conversational recommendation. *arXiv preprint arXiv:2010.10333*.

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*, pages 845–854, Florence, Italy. Association for Computational Linguistics.

Arvind Neelakantan, Tao Xu, Raul Puri, Alec Radford, Jesse Michael Han, Jerry Tworek, Qiming Yuan, Nikolas Tezak, Jong Wook Kim, Chris Hallacy, et al. 2022. Text and code embeddings by contrastive pre-training. *arXiv preprint arXiv:2201.10005*.

Mario Neururer, Stephan Schlögl, Luisa Brinkschulte, and Aleksander Groth. 2018. Perceptions on authenticity in chat bots. *Multimodal Technologies and Interaction*, 2(3):60.

Sejoon Oh, Berk Ustun, Julian McAuley, and Srijan Kumar. 2022. Rank list sensitivity of recommender systems to interaction perturbations. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, pages 1584–1594.

Alexander Pan, Kush Bhatia, and Jacob Steinhardt. 2022. The effects of reward misspecification: Mapping and mitigating misaligned models. In *International Conference on Learning Representations*.

Manuel Portela and Carlos Granell-Canut. 2017. A new friend in our smartphone? observing interactions with chatbots in the search of emotional engagement. In *Proceedings of the XVIII International Conference on Human Computer Interaction*, pages 1–7.

Peixin Qin, Chen Huang, Yang Deng, Wenqiang Lei, and Tat-Seng Chua. 2024. Beyond persuasion: Towards conversational recommender system with credible explanations. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 4264–4282, Miami, Florida, USA. Association for Computational Linguistics.

Byron Reeves and Clifford Nass. 1996. *The Media Equation: How People Treat Computers, Television, and New Media like Real People and Places*. Cambridge University Press, USA.

Swarnadeep Saha, Omer Levy, Asli Celikyilmaz, Mohit Bansal, Jason Weston, and Xian Li. 2023a. Branch-solve-merge improves large language model evaluation and generation. *arXiv preprint arXiv:2310.15123*.

Swarnadeep Saha, Omer Levy, Asli Celikyilmaz, Mohit Bansal, Jason Weston, and Xian Li. 2023b. Branch-solve-merge improves large language model evaluation and generation. *Preprint*, arXiv:2310.15123.

Ivan Sekulić, Mohammad Aliannejadi, and Fabio Crestani. 2022. Evaluating mixed-initiative conversational search systems via user simulation. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*, WSDM '22, page 888–896, New York, NY, USA. Association for Computing Machinery.

Clemencia Siro, Mohammad Aliannejadi, and Maarten De Rijke. 2023. Understanding and predicting user satisfaction with conversational recommender systems. *ACM Trans. Inf. Syst.*, 42(2).

Yueming Sun and Yi Zhang. 2018. Conversational recommender system. In *The 41st international acm sigir conference on research & development in information retrieval*, pages 235–244.

Ekaterina Svikhnushina, Alexandru Placinta, and Pearl Pu. 2021. User expectations of conversational chatbots based on online reviews. In *Designing Interactive Systems Conference 2021*, pages 1481–1491.

Cynthia A Thompson, Mehmet H Goker, and Pat Langley. 2004. A personalized system for conversational recommendations. *Journal of Artificial Intelligence Research*, 21:393–428.

Thi Ngoc Trang Tran, Alexander Felfernig, Christoph Trattner, and Andreas Holzinger. 2021. Recommender systems in the healthcare domain: state-of-the-art and research issues. *Journal of Intelligent Information Systems*, 57:171–201.

Cunxiang Wang, Xiaoze Liu, Yuanhao Yue, Xiangru Tang, Tianhang Zhang, Cheng Jiayang, Yunzhi Yao, Wenyang Gao, Xuming Hu, Zehan Qi, et al. 2023a. Survey on factuality in large language models: Knowledge, retrieval and domain-specificity. *arXiv preprint arXiv:2310.07521*.

Peiyi Wang, Lei Li, Liang Chen, Dawei Zhu, Binghuai Lin, Yunbo Cao, Qi Liu, Tianyu Liu, and Zhifang Sui. 2023b. Large language models are not fair evaluators. *arXiv preprint arXiv:2305.17926*.

Ting-Chun Wang, Shang-Yu Su, and Yun-Nung Chen. 2022a. Barcor: Towards a unified framework for conversational recommendation systems. *arXiv preprint arXiv:2203.14257*.

Xiaolei Wang, Xinyu Tang, Xin Zhao, Jingyuan Wang, and Ji-Rong Wen. 2023c. 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*, pages 10052–10065, Singapore. Association for Computational Linguistics.

Xiaolei Wang, Kun Zhou, Xinyu Tang, Wayne Xin Zhao, Fan Pan, Zhao Cao, and Ji-Rong Wen. 2023d. Improving conversational recommendation systems via counterfactual data simulation. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 2398–2408.

Xiaolei Wang, Kun Zhou, Ji-Rong Wen, and Wayne Xin Zhao. 2022b. Towards unified conversational recommender systems via knowledge-enhanced prompt learning. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 1929–1937.

Pontus Wärnestål. 2005. User evaluation of a conversational recommender system. In *Proceedings of the 4th Workshop on Knowledge and Reasoning in Practical Dialogue Systems*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

Minghao Wu and Alham Fikri Aji. 2023. Style over substance: Evaluation biases for large language models. *arXiv preprint arXiv:2307.03025*.

Dayu Yang, Fumian Chen, and Hui Fang. 2024. Behavior alignment: A new perspective of evaluating llm-based conversational recommendation systems. *arXiv preprint arXiv:2404.11773*.

Wenzhuo Yang, Jia Li, Chenxi Li, Latrice Barnett, Markus Anderle, Simo Arajarvi, Harshavardhan Utharavalli, Caiming Xiong, and Steven HOI. 2021. On the diversity and explainability of recommender systems: A practical framework for enterprise app recommendation. In *Proceedings of the 30th ACM*

International Conference on Information & Knowledge Management*, CIKM '21, page 4302–4311, New York, NY, USA. Association for Computing Machinery.

Seonghyeon Ye, Doyoung Kim, Sungdong Kim, Hyeonbin Hwang, Seungone Kim, Yongrae Jo, James Thorne, Juho Kim, and Minjoon Seo. 2023. Flask: Fine-grained language model evaluation based on alignment skill sets. *arXiv preprint arXiv:2307.10928*.

Dianer Yu, Qian Li, Xiangmeng Wang, Qing Li, and Guandong Xu. 2023. Counterfactual explainable conversational recommendation. *IEEE Transactions on Knowledge and Data Engineering*.

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*, pages 1512–1520.

Xiaoyu Zhang, Xin Xin, Dongdong Li, Wenxuan Liu, Pengjie Ren, Zhumin Chen, Jun Ma, and Zhaochun Ren. 2023. Variational reasoning over incomplete knowledge graphs for conversational recommendation. In *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining*, pages 231–239.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and William B Dolan. 2020. Dialogpt: Large-scale generative pre-training for conversational response generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 270–278.

Jinfeng Zhou, Bo Wang, Ruifang He, and Yuexian Hou. 2021. Crfr: Improving conversational recommender systems via flexible fragments reasoning on knowledge graphs. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4324–4334.

Jinfeng Zhou, Bo Wang, Minlie Huang, Dongming Zhao, Kun Huang, Ruifang He, and Yuexian Hou. 2022. Aligning recommendation and conversation via dual imitation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 549–561, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Li Zhou, Jianfeng Gao, Di Li, and Heung-Yeung Shum. 2020. The design and implementation of xiaoice, an empathetic social chatbot. *Computational Linguistics*, 46(1):53–93.

User Types	Redial				OpendialKG			
	B	**C**	**K**	**U**	**B**	**C**	**K**	**U**
Anticipation	**76**	**76**	**76**	**76**	**64**	**64**	**64**	**64**
Age group=Adults	19	19	19	19	16	16	16	16
Age group=Children	19	19	19	19	16	16	16	16
Age group=Seniors	19	19	19	19	16	16	16	16
Age group=Teens	19	19	19	19	16	16	16	16
Boredom	**76**	**76**	**76**	**76**	**64**	**64**	**64**	**64**
Age group=Adults	19	19	19	19	16	16	16	16
Age group=Children	19	19	19	19	16	16	16	16
Age group=Seniors	19	19	19	19	16	16	16	16
Age group=Teens	19	19	19	19	16	16	16	16
Confusion	**76**	**76**	**76**	**76**	**64**	**64**	**64**	**64**
Age group=Adults	19	19	19	19	16	16	16	16
Age group=Children	19	19	19	19	16	16	16	16
Age group=Seniors	19	19	19	19	16	16	16	16
Age group=Teens	19	19	19	19	16	16	16	16
Curiosity	**76**	**76**	**76**	**76**	**64**	**64**	**64**	**64**
Age group=Adults	19	19	19	19	16	16	16	16
Age group=Children	19	19	19	19	16	16	16	16
Age group=Seniors	19	19	19	19	16	16	16	16
Age group=Teens	19	19	19	19	16	16	16	16
Delight	**76**	**76**	**76**	**76**	**64**	**64**	**64**	**64**
Age group=Adults	19	19	19	19	16	16	16	16
Age group=Children	19	19	19	19	16	16	16	16
Age group=Seniors	19	19	19	19	16	16	16	16
Age group=Teens	19	19	19	19	16	16	16	16
Disappointment	**76**	**76**	**76**	**76**	**64**	**64**	**64**	**64**
Age group=Adults	19	19	19	19	16	16	16	16
Age group=Children	19	19	19	19	16	16	16	16
Age group=Seniors	19	19	19	19	16	16	16	16
Age group=Teens	19	19	19	19	16	16	16	16
Excitement	**76**	**76**	**76**	**76**	**64**	**64**	**64**	**64**
Age group=Adults	19	19	19	19	16	16	16	16
Age group=Children	19	19	19	19	16	16	16	16
Age group=Seniors	19	19	19	19	16	16	16	16
Age group=Teens	19	19	19	19	16	16	16	16
Frustration	**76**	**76**	**76**	**76**	**64**	**64**	**64**	**64**
Age group=Adults	19	19	19	19	16	16	16	16
Age group=Children	19	19	19	19	16	16	16	16
Age group=Seniors	19	19	19	19	16	16	16	16
Age group=Teens	19	19	19	19	16	16	16	16
Indifference	**76**	**76**	**76**	**76**	**64**	**64**	**64**	**64**
Age group=Adults	19	19	19	19	16	16	16	16
Age group=Children	19	19	19	19	16	16	16	16
Age group=Seniors	19	19	19	19	16	16	16	16
Age group=Teens	19	19	19	19	16	16	16	16
Surprise	**76**	**76**	**76**	**76**	**64**	**64**	**64**	**64**
Age group=Adults	19	19	19	19	16	16	16	16
Age group=Children	19	19	19	19	16	16	16	16
Age group=Seniors	19	19	19	19	16	16	16	16
Age group=Teens	19	19	19	19	16	16	16	16
Trust	**76**	**76**	**76**	**76**	**64**	**64**	**64**	**64**
Age group=Adults	19	19	19	19	16	16	16	16
Age group=Children	19	19	19	19	16	16	16	16
Age group=Seniors	19	19	19	19	16	16	16	16
Age group=Teens	19	19	19	19	16	16	16	16
Satisfaction	**76**	**76**	**76**	**76**	**64**	**64**	**64**	**64**
Age group=Adults	19	19	19	19	16	16	16	16
Age group=Children	19	19	19	19	16	16	16	16
Age group=Seniors	19	19	19	19	16	16	16	16
Age group=Teens	19	19	19	19	16	16	16	16
In Total	**912**	**912**	**912**	**912**	**768**	**768**	**768**	**768**

Table 7: Number of conversational attributes. 'B' stands for BARCOR, 'C' for CHATCRS, 'K' for KBRD, 'U' for UNICRS.

A Implementation Details

We conduct all our experiments using a single Nvidia RTX A6000, and we implement our codes in PyTorch. The core of our code framework is built upon open-source code from the latest research on CRS[6] (Wang et al., 2023c). In order to guarantee replicability, we have established fixed values for the Temperature and Seed parameters of ChatGPT (i.e., GPT-3.5-16K-turbo), setting the Temperature to 0 and the Seed to 42.

A.1 Implementation of CRS Models

We evaluated all open-source CRS models in its code using their checkpoints, including the KBRD (Chen et al., 2019), BARCOR (Wang et al., 2022a), UNICRS (Wang et al., 2022b), and CHATGPT-based model (Wang et al., 2023c) which represents the current SOTA, incorporating *text-embedding-ada-002* (Neelakantan et al., 2022) for modeling the semantic embeddings.

- **KBRD** (Chen et al., 2019) bridges the recommendation module and the Transformer-based conversation module via knowledge propagation.
- **BARCOR** (Wang et al., 2022a) is a unified framework based on BART (Lewis et al., 2020), which implements the recommendation and response generation tasks in a single model.
- **UNICRS** (Wang et al., 2022b) is a unified framework based on DialoGPT (Zhang et al., 2020), with a semantic fusion module to enhance the semantic association between conversation history and knowledge graphs.
- **CHATCRS** (Wang et al., 2023c) is the SOTA CRS model, which incorporates ChatGPT for the conversation module and the text-embedding-ada-002 (Neelakantan et al., 2022) to enhance the recommendation module.

See the prompts in the original paper for the implementation of CHATGPT-based model. Note that we feed the top 5 items from the recommendation module into the prompts of ChatGPT for re-ranking and generating responses. Also, the UNICRS utilizes DialoGPT-small as the backbone, while the BARCOR utilizes BART-base with a 2-layer encoder and decoder, following Wang et al. (2023c).

A.2 Implementation of CONCEPT

CONCEPT resorts to an LLM-based user simulator and evaluator for cost-effective evaluation, together with fine-grained ability-specific scoring rubrics. Table 7 reveals the statistics of our generated conversation dataset.

[6] https://github.com/RUCAIBox/iEvaLM-CRS

A.2.1 Evaluation Process

We used user simulators to interact and converse with different CRSs to produce a conversation dataset. In the experiments, the simulator simulated 12 different personas and 4 different age groups using ChatGPT. Specifically, CONCEPT considers the free-form chit-chat between the user simulator and CRS. To simulate real-world scenarios, the simulator has no access to its targeted items during the conversation. Any item that meets these preferences, such as having attributes completely consistent with or containing the simulator's preferences, is considered a successful recommendation. During the conversation, CONCEPT allows the simulator to describe their preferences in their own words, as in real-world situations, users may not use the exact terms defined in the pre-defined preference values. During the conversation between CRS and the user, we recorded the recommendation results of each round of the recommendation system and the results recommended by the conversational agent, in order to evaluate the success of the recommendations from different perspectives. Finally, the conversation will end if the simulator accepts recommendations or if the conversation reaches the maximum number of turns. If the user chooses to accept the recommendation from the conversational agent, the user is required to add '[END]' at the end of his response to indicate the end of the conversation. These diverse users generated a total of 6720 conversations, which were used to evaluate the performance of different CRS in more realistic scenarios.

Afterward, CONCEPT utilizes both the LLM-based evaluator and computational metrics to assess the abilities of CRS, as outlined in Table 2. Note that the evaluator is utilized when corresponding computational metrics are not available. Specifically, we employ the evaluator to assess the cooperation ability, social awareness, and persuasiveness of recommendation explanations of CRS. This involves prompting the evaluator with fine-grained, ability-specific scoring rubrics (Saha et al., 2023b), which are generated by the LLM and then refined by humans. For more details, refer to Appendix A.2.3 and Appendix A.2.4.

A.2.2 LLM-based User Simulator

Our user simulator has unique personas and preferences. Specifically, the personas are generated by prompting ChatGPT in a zero-shot manner, following (Wang et al., 2023a), while the preferences are defined using attributes from two benchmark datasets, Redial and OpendialKG.

Personas. We generate a persona list as the starting point by prompting ChatGPT in a zero-shot manner. Inspired by (Wang et al., 2023a), we first instruct the ChatGPT to return 20 distinct personas and their corresponding descriptions, such as curiosity, and then filter out duplicates. We utilize the generated persona list, so it is highly likely that ChatGPT is familiar with these personas and would contain knowledge and information about them. As a result, we obtain 12 unique personas, simulating diverse sentiments may encountered when using a conversational recommendation system (namely, Anticipation, Boredom, Fusion, Curiosity, Delight, Disappointment, Exceptions, Frustration, Independence, Surprise, Trust, Satisfaction), and each type of user may belong to different age groups (namely, Adults, Children, Senior, and Teens). We assigned different attribute groups as movie preferences for each user type. These users with varying emotions, age groups, and movie preferences then interacted with the CRS. If the attribute group of the movie recommended by the CRS is equal to or includes the user's target attribute group, it is considered a correct recommendation. For the Redial dataset, each user type corresponds to 76 conversation data with one CRS, resulting in a total of 912 conversation data across all user types and one CRS. Similarly, for the OpendialKG dataset, each user type generates 64 conversation data with each CRS, resulting in a total of 768 conversation data across all user types and one CRS. Since we tested 4 different CRSs, we collected a total of 6720 conversation data. Refer to Table 18 for details.

Preferences. The current evaluation method for CRS is not reflective of real-world situations, as it assumes that every user knows the target item (Wang et al., 2023c). To tackle this issue, our user simulators only have access to their preferences. The user would only choose to accept the recommendation when the CRS provides explanations for the recommendations and makes the users feel that the recommended items match their preferences. To achieve this, the preferences are defined using attributes from two benchmark datasets, Redial Redial (Li et al., 2018) and OpendialKG (Moon et al., 2019). For each dataset, every movie has a feature group made up of one or more attributes. In the case of the Redial dataset, we conducted experiments using feature groups containing 3 attributes and excluded less common groups. Ultimately, we retained the 19 most prevalent attribute groups for the study, with each group corresponding to at least 50 different movies. For the OpendialKG dataset, the issue of less common attributes is more pronounced. Initially, we selected the most prevalent attributes (each corresponding to at least 100 movies) and then kept the 16 most common attribute groups for experimentation.

Raw Attribute	ChatGPT-adjusted Attributes
	Redial
action	thrilling and adrenaline-pumping action movie
adventure	exciting and daring adventure movie
animation	playful and imaginative animation
biography	inspiring and informative biography
comedy	humorous and entertaining flick
crime	suspenseful and intense criminal film
documentary	informative and educational documentary
drama	emotional and thought-provoking drama
family	heartwarming and wholesome family movie
fantasy	magical and enchanting fantasy movie
film-noir	dark and moody film-noir
game-show	entertaining and interactive game-show
history	informative and enlightening history movie
horror	chilling, terrifying and suspenseful horror movie
music	melodious and entertaining music
musical	theatrical and entertaining musical
mystery	intriguing and suspenseful mystery
news	informative and current news
reality-tv	dramatic entertainment and reality-tv
romance	romantic and heartwarming romance movie with love story
sci-fi	futuristic and imaginative sci-fi with futuristic adventure
short	concise and impactful film with short story
sport	inspiring and motivational sport movie
talk-show	informative and entertaining talk-show such as conversational program
thriller	suspenseful and thrilling thriller with gripping suspense
war	intense and emotional war movie and wartime drama
western	rugged and adventurous western movie and frontier tale
	OpendialKG
Action	adrenaline-pumping action
Adventure	thrilling adventure
Sci-Fi	futuristic sci-fi
Comedy	lighthearted comedy
Romance	heartwarming romance
Romance Film	emotional romance film
Romantic comedy	charming romantic comedy
Fantasy	enchanting fantasy
Fiction	imaginative fiction
Science Fiction	mind-bending science fiction
Speculative fiction	thought-provoking speculative fiction
Drama	intense drama
Thriller	suspenseful thriller
Animation	colorful animation
Family	heartwarming family
Crime	gripping crime
Crime Fiction	intriguing crime fiction
Historical drama	captivating historical drama
Comedy-drama	humorous comedy-drama
Horror	chilling horror
Mystery	intriguing mystery

Table 8: Illustration on ChatGPT-adjusted attributes. We provide the user simulators with adjusted attributes to prevent them from revealing their target attributes. In real-world situations, users may not always use the same words as those used during model training to express their preferences.

Simulation via Prompts. We prompt the simulator with different personas using the persona descriptions generated by ChatGPT (cf. Table 9.). CONCEPT incorporates the Theory of Mind into our simulator to emulate human social cognition (Fischer, 2023). This is achieved by prompting the simulator to first assess its current mental state before generating responses, encouraging reflection on its predefined personality traits and social interactions. Refer to Appendix E for prompts. Additionally, there is a maximum of 10 turns allowed in the conversation, and it will only conclude when the simulated user accepts the recommendation. To prevent the simulated users from directly stating the same attribute preferences as the pre-defined values (i.e., the attribute group) when asked about their preferences by the CRS, we include adjectives before each attribute during the simulation, and we allow the user to describe his/her preference using their own words. This approach is reasonable because in real scenarios, users may not use the exact words as the pre-defined attribute values during the conversation. We have summarized the adjusted values for each attribute in Tables 8, achieved by prompting ChatGPT.

A.2.3 LLM-based Evaluator

We summarize the evaluation method for each characteristic / ability in Table 2. We utilize an LLM-based evaluator to evaluate characteristics or abilities when corresponding computational metrics are not available. These include the abilities of the Manner, Response Quality, and Relevance in Cooperation, Social Awareness, and ability of persuasiveness in Identity.

Based on the findings of previous studies (Ye et al., 2023; Wang et al., 2023b,c), we utilize an instance-wise evaluator to conduct detailed assessments. Expanding on earlier research (Ye et al., 2023; Wang et al., 2023c; Liu et al., 2023), we task the evaluator with using fine-grained scoring rubrics to mitigate scoring bias (see Section C for our reliability analysis). The evaluator assigns scores ranging from 1 to 5 based on ability-specific score rubrics, each accompanied by a corresponding description, using the conversation data. For the generation of fine-grained scoring rubrics, we follow the approach of previous works and employ ChatGPT to produce a set of evaluation criteria (Saha et al., 2023a). This set of criteria serves as a starting point for human refinement (Li et al., 2023). Prior to assigning a score, we require the evaluator to provide a rationale, drawing inspiration from the effectiveness of CoT prompting (Wei et al., 2022; Ye et al., 2023). For implementation details on the evaluator, refer to Appendix E.

A.2.4 Computational Metrics

We introduce the computational metrics for evaluating the remaining characteristics or abilities. These include the abilities of quality and reliability in Recommendation intelligence, the ability of sincerity (Cooperation) in Social intelligence, and the abilities of identity and coordination in Personification. We conclude them as follows.

- *Quality*. We consider computational metrics for automatic evaluation following previous works (Wang et al., 2023c; Zhang et al., 2023; Yu et al., 2023), including $Recall@k(k = 1, 10, 25, 50)$, recommendation success rate ($SR@k, k = 3, 5, 10$), user acceptance rate (AR), and average turns (AT) required to reach the successful recommendation. Note that we determine user acceptance of the recommendation by checking for the presence of the special token '[END]' in the user's response.

- *Reliability*. We generate sets of user response pairs with similar meanings using ChatGPT paraphrasing. This enables us to evaluate how reliably the CRS performs with contextual nuances. Formally, given the same conversation history and a user response pair u_1 and u_2, we define four metrics to measure reliability: the rate of *Consistent Action*, indicating whether the CRS constantly provides recommendations based on u_1 and u_2; the rate of *Consistent Recommendation*, which assesses if the CRS recommends the same items given two user responses u_1 and u_2; and the rates of *Diversity*, which evaluate whether the recommended items, even if inconsistent, align with user preferences. We refer to it as *Sensitivity* when the system provides inconsistent and inaccurate recommendations that do not align with user preferences.

- *Sincerity*. We value the ratio of non-existent items and deceptive tactics. For non-existent items, we have tallied how many items recommended by the CRS do not exist in the dataset. As for the deceptive tactics, we focus on the items accepted by users and tally how many of these do not align with the users' pre-defined preferences. If users accept these misleading items, we consider the CRS to be employing a deceptive tactics, leading users to believe that the recommended items meet their preferences.

- *Identity*. CRS should be self-aware of its identity and operate within its designated scope, differentiating itself from sales systems. We evaluate this by assessing the sincerity of explanations, i.e., the deceptive tactics.

Persona	Templates (The Input of ChatGPT Paraphraser)	ChatGPT-paraphrased Persona Descriptions
Emotion=Boredom Age group=Adults	you are a person that are easy to be Boredom. This means that your are Feeling uninterested or uninspired by the recommended movie choices. Also, you are a Adults person	You are easily bored, feeling uninterested or uninspired by the recommended movie choices. As an adult, you seek movies that can captivate your attention.
Emotion=Anticipation Age group=Children	you are a person that are easy to be Anticipation. This means that your are Looking forward to watching recommended movies and experiencing new stories. Also, you are a Children person	You are filled with anticipation, looking forward to watching recommended movies and experiencing new stories. As a child, you enjoy the excitement of discovering new films.

Table 9: Persona description generation. We start by converting each combination of personality and age into a single sentence using a template, and then we use ChatGPT to rephrase it into a natural-sounding sentence. Refer to Appendix E for Prompts.

Attribute Group	BARCOR	CHATCRS	KBRD	UNICRS
Redial				
['action', 'adventure', 'animation']	1.77	4.33	1.13	1.38
['action', 'adventure', 'comedy']	1.94	4.31	1.21	1.33
['action', 'adventure', 'drama']	1.63	4.19	1.08	1.40
['action', 'adventure', 'fantasy']	1.85	4.31	1.19	1.46
['action', 'adventure', 'sci-fi']	1.94	4.27	1.10	1.29
['action', 'adventure', 'thriller']	1.83	4.44	1.15	1.27
['action', 'crime', 'drama']	1.79	4.29	1.15	1.31
['action', 'crime', 'thriller']	1.83	4.42	1.19	1.40
['adventure', 'animation', 'comedy']	1.92	4.25	1.10	1.27
['adventure', 'comedy', 'family']	1.92	4.06	1.25	1.52
['biography', 'crime', 'drama']	1.53	4.21	1.08	1.23
['biography', 'drama', 'history']	1.65	4.63	1.10	1.27
['comedy', 'drama', 'family']	1.92	4.29	1.04	1.50
['comedy', 'drama', 'romance']	1.77	4.31	1.19	1.44
['crime', 'drama', 'thriller']	1.81	4.42	1.10	1.40
['drama', 'horror', 'mystery']	1.83	4.38	1.08	1.40
['horror', 'mystery', 'thriller']	2.08	4.17	1.10	1.40
['crime', 'drama', 'mystery']	1.83	4.46	1.10	1.29
['action', 'comedy', 'crime']	1.56	4.13	1.06	1.29
Avg.+-Std.	1.81+0.14	4.31+0.13	1.13+0.05	1.36+0.08
OpendialKG				
['action', 'adventure', 'fantasy']	1.58	4.19	1.02	1.15
['action', 'adventure', 'sci-fi']	1.46	4.40	1.08	1.13
['action', 'adventure', 'thriller']	1.67	4.29	1.00	1.23
['comedy', 'drama', 'romance']	1.73	3.96	1.04	1.10
['crime', 'drama', 'thriller']	1.56	4.63	1.00	1.15
['horror', 'mystery', 'thriller']	1.58	4.33	1.02	1.19
['Adventure', 'Animation', 'Comedy', 'Family']	1.56	4.31	1.04	1.17
['Comedy', 'Romance', 'Romance Film']	1.46	3.94	1.02	1.15
['Action', 'Adventure', 'Sci-Fi', 'Thriller']	1.46	4.40	1.00	1.08
['Fantasy', 'Fiction', 'Science Fiction', 'Speculative fiction']	1.65	3.48	1.06	1.08
['Drama', 'Historical period drama', 'Romance', 'Romance Film']	1.58	3.88	1.00	1.15
['Comedy', 'Comedy-drama', 'Drama']	1.54	3.85	1.00	1.13
['Action', 'Crime', 'Drama', 'Thriller']	1.65	3.50	1.04	1.19
['Action', 'Adventure', 'Fantasy', 'Sci-Fi']	1.50	3.77	1.04	1.15
['Crime', 'Crime Fiction', 'Drama', 'Thriller']	1.50	4.27	1.02	1.15
['Comedy', 'Romance', 'Romance Film', 'Romantic comedy']	1.65	3.71	1.00	1.13
Avg.+-Std.	1.57+0.08	4.06+0.33	1.02+0.02	1.14+0.04

Table 10: Overall performance evaluation when recommending items with various attributes

Age Group	BARCOR	CHATCRS	KBRD	UNICRS
OpendialKG				
Children	1.58	4.14	1.03	1.16
Teens	1.62	4.15	1.03	1.15
Adults	1.54	3.94	1.03	1.15
Seniors	1.54	3.99	1.01	1.12
Redial				
Children	1.76	4.33	1.15	1.39
Teens	1.86	4.29	1.14	1.36
Adults	1.84	4.38	1.11	1.32
Seniors	1.79	4.23	1.10	1.38

Table 11: Overall performance evaluation when dealing with users of various ages.

Personas	BARCOR	CHATCRS	KBRD	UNICRS
Redial				
Anticipation	1.76	4.91	1.24	1.39
Boredom	1.72	3.16	1.05	1.38
Confusion	1.84	3.49	1.13	1.32
Curiosity	1.86	4.82	1.16	1.41
Delight	1.78	4.47	1.14	1.38
Disappointment	1.82	3.33	1.08	1.33
Excitement	1.93	4.96	1.14	1.39
Frustration	1.68	4.67	1.07	1.26
Indifference	1.78	3.92	1.07	1.26
Satisfaction	1.83	4.46	1.17	1.49
Surprise	1.88	4.89	1.14	1.41
Trust	1.86	4.62	1.13	1.29
OpendialKG				
Anticipation	1.69	4.67	1.05	1.16
Boredom	1.58	2.94	1.06	1.14
Confusion	1.38	3.38	1.00	1.11
Curiosity	1.63	4.58	1.05	1.08
Delight	1.58	4.00	1.02	1.17
Disappointment	1.56	3.00	1.02	1.14
Excitement	1.52	4.59	1.00	1.28
Frustration	1.47	4.38	1.00	1.05
Indifference	1.56	4.08	1.03	1.09
Satisfaction	1.63	4.13	1.03	1.19
Surprise	1.69	4.48	1.05	1.16
Trust	1.58	4.45	1.00	1.16

Table 12: Overall performance of CRS when dealing with users of various personas

- *Coordination.* We simulate users with different personas and assess how the CRS performs across all previously mentioned abilities. To qualify the coordination score, we initially computed the Range and mean of the CRS model's scores across various users, based on the different abilities mentioned earlier. It's worth noting that the Range is more effective than the standard deviation in highlighting the variability of the CRS model across different users. Subsequently, we divided the Range by the mean to derive the Coordination score of the CRS for that specific ability. The overall Coordination score of the CRS is then calculated as the average across all abilities.

	Redial				OpendialKG			
	KBRD	BARCOR	UNICRS	CHATGPT	KBRD	BARCOR	UNICRS	CHATGPT
Action Consistency (↑)	75.96%	94.71%	82.63%	**99.62%**	98.58%	99.49%	90.48%	**99.76%**
Recommend different items (↓)	**33.99%**	45.28%	41.72%	52.48%	64.56%	70.34%	80.73%	**44.36%**
Recommendation Diversity (↑)	9.22%	10.27%	23.79%	**27.45%**	0.21%	3.94%	7.99%	**12.97%**
Recommemdation Sensitivity (↓)	90.78%	89.73%	76.21%	**72.55%**	99.79%	96.06%	92.01%	**87.03%**

Table 13: Evaluation of recommendation reliability across each benchmark dataset

Personas	Conversational Agent Perspective SR (K=10)				Recommendation System Perspective SR (K=10)				User Acceptance Rate			
	BARCOR	CHATCRS	KBRD	UNICRS	BARCOR	CHATCRS	KBRD	UNICRS	BARCOR	CHATCRS	KBRD	UNICRS
						Redial						
Children	47.81	60.96	32.02	46.49	39.04	43.42	4.82	19.74	0.44	71.05	0.44	0.00
Teens	51.75	61.40	28.95	41.23	37.72	48.25	4.82	17.54	3.07	71.49	0.00	0.88
Adults	49.12	65.35	27.63	42.98	39.47	47.37	3.51	17.54	1.32	72.81	0.44	0.44
Seniors	53.95	61.84	28.95	42.98	41.67	47.37	4.82	17.98	0.88	67.98	0.44	0.00
Avg,±Std.	50.66±2.37	62.39±1.74	29.39±1.61	43.42±1.91	39.47±1.42	46.6±1.87	4.5±0.57	18.2±0.9	1.43±1	70.83±1.77	0.33±0.19	0.33±0.36
						OpendialKG						
Children	33.33	45.31	14.06	46.35	3.65	43.75	16.15	29.69	0.52	65.63	1.04	0.00
Teens	35.42	38.02	10.94	48.44	8.85	44.27	17.19	26.56	1.04	67.19	0.52	0.52
Adults	35.42	50.52	13.02	41.67	7.81	51.04	15.63	29.17	0.00	62.50	0.00	0.52
Seniors	36.46	43.75	11.46	44.79	8.85	46.88	15.10	31.77	1.04	61.98	0.00	0.00
Avg,±Std.	35.16±1.14	44.4±4.46	12.37±1.24	45.31±2.47	7.29±2.15	46.48±2.89	16.02±0.77	29.3±1.86	0.65±0.43	64.32±2.16	0.39±0.43	0.26±0.26

Table 14: Recommendation quality evaluation when dealing with users of various ages

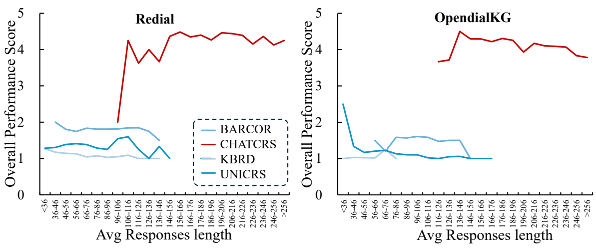

Figure 7: Length bias evaluation. Longer responses do not necessarily yield higher scores.

B Reliability of CONCEPT

This section aim to assess the reliability of CONCEPT in terms of its replicability and potential bias. More evaluation details could be found at Appendix C. To sum up, leveraging LLM-based simulators and evaluators to implement our CONCEPT is reliable, consistent with earlier findings (Ye et al., 2023; Wang et al., 2023c).

Replicability Analysis. We have established fixed values for the temperature and seed parameters to guarantee replicability of our LLM-based simulator and evaluator.

Reliability Analysis of CONCEPT Simulator. We assess our LLM-based simulators and demonstrate their reliability from the following perspectives.

- *Preference following*. We assess whether each simulator will strictly follow its own preferences to accept the recommendations. We found that the proportion of the case when the recommendation explanation from the CRS clearly does not meet user's preferences, but the user accept the recommendations, is only 7.44%.
- *Naturalness & Usefulness*. We evaluate the naturalness and usefulness of our simulator following previous studies (Sekulić et al., 2022; Wang et al., 2023c). Our human evaluation yielded average scores of 3.88 and 3.79 for naturalness and usefulness, respectively, with Krippendorff's alpha values of 0.57 and 0.60. This demonstrates that our simulators can generate fluent, human-like utterances that effectively guide conversations towards meeting user information needs.

Reliability Analysis of CONCEPT Evaluator. We build upon previous research (Ye et al., 2023; Wang et al., 2023c; Liu et al., 2023) and minimize potential biases by employing detailed scoring rubrics. In this study, we examine the evaluation biases[7], and demonstrate that CONCEPT scoring is unbiased and aligns with the results of human evaluations.

- *Length bias*[8]. As shown in Figure 7, we require CHATGPT to provide an overall performance score (*Y-axis*, Overall Performance Score) based on all the ability-specific scores and then plot the relationship between various CRS reply lengths (*X-axis*, Avg. Response Length) and the scores. Our findings show that CONCEPT remains unaffected by the length bias. CHATCRS tends to produce lengthier responses, but this does not imply that longer responses will yield higher scores.

[7] CONCEPT does not entail making decisions based on a specific group of candidates, thus no position bias.

[8] LLMs have a preference for longer responses (Wu and Aji, 2023).

- *Self-enhancement bias*[9]. Our human evaluation demonstrates that the CONCEPT evaluation results are consistent and align with human assessments. Specifically, the human evaluation results and the LLM-based evaluation results are closely related, with a correlation coefficient of 61.24% and Krippendorff's alpha of 53.10%. This indicates the reliability of our LLM-based evaluation.

C Human Evaluation

Previous findings highlight that offering detailed scoring rubrics or criteria contributes to achieving consistent and aligned evaluations with human assessments (Liu et al., 2023; Wang et al., 2023c), ultimately resulting in a dependable LLM-based evaluator as a viable alternative to human evaluators. We also conduct human evaluation to analyze the consistency. Our human evaluation is used to examine the capability of the CHATGPT-based user simulator being a movie seeker, while checking the correlation between the CHATGPT-based evaluation results and human evaluation results. The results are as follows, which indicate the reliability of our CONCEPT.

Human evaluation results based on the conversation data and the LLM-based evaluation results are highly related, with a correlation coefficient of 61.24% and Krippendorff's alpha of 53.10%. Due to budget limitations, we are unable to hire a large number of individuals with various personas and age groups for human evaluation. Instead, our human evaluation is conducted by two human evaluators. In particular, we randomly selected 120 conversations involving CHATCRS and users with various personas and age groups. Given the numerous scoring aspects in our CONCEPT, it is prone to inconsistent ratings across different aspects. To address this, we had ChatGPT score all aspects of these 120 conversations and provide an overall performance score. Subsequently, each human only needed to provide an overall score, without having to rate all aspects. During the scoring process, each evaluator independently scored these 120 conversations based on our detailed criteria. They then discussed any disagreements. We calculated the inter-annotator reliability using Krippendorff's alpha, achieving 41.34%. To ensure consistency and enhance the robustness of the evaluations, we took the average of the scores from

[9]Bias of CHATGPT to favor high scores for its generated content (Wang et al., 2023b).

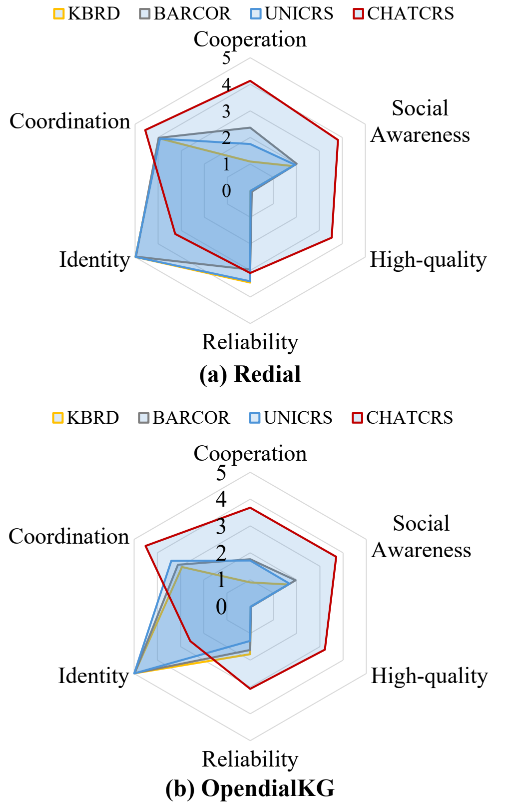

Figure 8: Evaluation of six primary abilities across each benchmark dataset

both evaluators as the final result for each conversation. We then compared the human evaluation results with those from an LLM-based evaluator and found a correlation of 61.24% and Krippendorff's alpha of 53.10%. This indicates the reliability of our LLM-based evaluator, which is consistent with previous findings (Wang et al., 2023c).

LLM-based user simulator is a reliable alternative to human. We require the two human evaluator to evaluate the reliability of the user simulator by assessing whether the simulator will strictly follow its own preferences to accept the recommendations. We found that the proportion of the

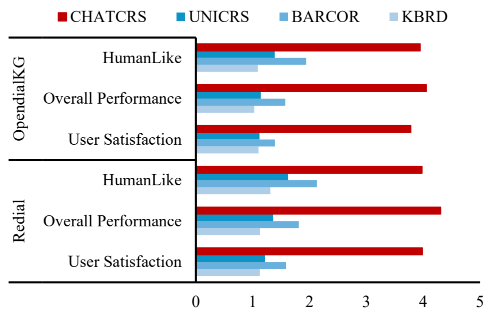

Figure 9: Human Likeness, overall performance, and user satisfaction.

case when the recommendation explanation from the CRS clearly does not meet user's preferences, but the user accept the recommendations, is only 7.44%. During the evaluation process, we found that the user simulator would continuously ask the CRS to confirm that the recommended movies meet their preferences, as we requested. There is some examples of the user simulator: "*Can you tell me more about 'The Chaser (2008)'? What's it about?*", and "*That sounds interesting. Can you tell me more about the specific humor and suspense elements in the movie?*". To dive into deeper exploration, we further assess the naturalness and usefulness of this simulator through additional human evaluations following previous studies (Sekulić et al., 2022; Wang et al., 2023c). Notably, both metrics are indicative of the simulator's quality: Naturalness is defined as how natural, fluent, and human-like an utterance is, while Usefulness is defined as an utterance being aligned with the user's information needs and effectively guiding the conversation towards the relevant topic. During the evaluation, we employ two annotators to evaluate the same 50 dialogues and instruct them to assign a score from 1 to 5 for each metric. Finally, the average scores for naturalness and usefulness are 3.88 and 3.79, respectively, with Krippendorff's alpha values of 0.57 and 0.60, indicating a high quality for our simulator.

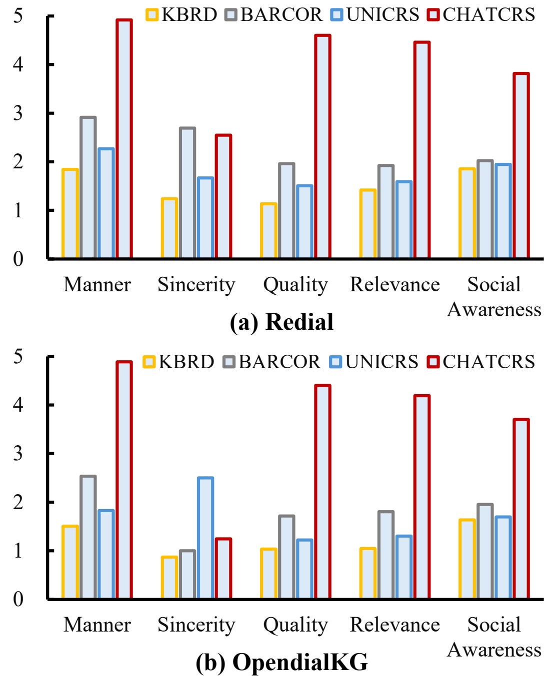

Figure 10: Evaluation of social-centric characteristics

D Additional Analysis

In this section, we provide additional evaluation results to achieve better understanding of off-the-shelf CRS models.

D.1 Overall Performance

We report more results in Figure 9 in terms of the Human Likeness, overall performance, and user satisfaction. This is achieved by prompting the LLM-based evaluator using the detailed results of all ability-specific scores and fine-grained scoring rubrics. Refer Appendix E for details on prompts.

D.2 Fine-grained Analysis on each Benchmark Dataset

Figure 10 shows the evaluation of social-centric characteristics across each benchmark dataset. Figure 8 presents the results of CRS models across the six primary abilities on each benchmark dataset, while Table 13 reports the details on the reliability of each CRS model. In comparison, OpendialKG appears to pose a greater challenge as a benchmark dataset than Redial. One potential factor is the inclusion of numerous semantically similar item attributes in the OpendialKG dataset, which may have impeded the training of CRS models to some extent. This could directly result in a notable decrease in the reliability of CRS models on OpendialKG. Given these findings, the current field of CRS is in urgent need of a high-quality conversational recommendation dataset. Addressing the limitations of CRS outlined in this paper, such a dataset should not only feature high-quality attributes, but also encompass responses to various user scenarios and exhibit sufficient social behavior.

Persona Analysis. Table 12 and Table 15 demonstrate the overall performance and recommendation performance of CRS in engaging with users of various personas, respectively. Our results support the main text's conclusion, highlighting significant differences in how effective the CRS model is when interacting with users of diverse personas. This underscores the importance of enhancing CRS's adaptability, enabling it to customize its social behavior and recommend dialogue strategies tailored to different users.

Age Group Analysis. Table 14 and Table 11 present the specifics of CRS model recommendations for users of different ages and their overall performance, respectively. Our findings indicate that the current CRS can be equally effective for people of all ages. However, it is evident that users in younger age groups are more likely to accept CRS recommendations, leading to higher overall scores. We did not observe any tendency for CHATCRS to use dishonest strategies to deceive younger users into accepting recommendations.

Attribute Group Analysis. Table 16 and Table 10 provide details on the recommendations made by CRS models based on user preferences and their overall performance, respectively. Our findings indicate that there is no significant difference in the effectiveness of CHATCRS when recommending different types of items. However, on the OpendialKG, the performance difference is more pronounced. The results suggest that this is largely influenced by the presence of semantically similar attributes in the data, such as 'Crime' and 'Crime Fiction'. These findings remain consistent across other models, showing more noticeable variations in performance on the OpendialKG dataset, despite most CRS being trained on this data.

E Prompts

We outline the ChatGPT prompts in Table 17 and Table 18, focusing on the user simulation and the user evaluator, respectively. In order to guarantee replicability, we have established fixed values for the Temperature and Seed parameters of ChatGPT, setting Temperature to 0 and Seed to 42.

Personas	Conversational Agent Perspective SR (K=10)				Recommendation System Perspective SR (K=10)				User Acceptance Rate			
	BARCOR	CHATCRS	KBRD	UNICRS	BARCOR	CHATCRS	KBRD	UNICRS	BARCOR	CHATCRS	KBRD	UNICRS
Redial												
Anticipation	55.26	61.84	30.26	40.79	40.79	48.68	2.63	19.74	3.95	100.00	1.32	1.32
Boredom	38.16	77.63	31.58	32.89	38.16	57.89	3.95	15.79	0.00	13.16	0.00	0.00
Confusion	48.68	71.05	27.63	48.68	36.84	52.63	7.89	13.16	0.00	28.95	0.00	0.00
Curiosity	51.32	56.58	31.58	50.00	39.47	47.37	7.89	22.37	1.32	97.37	0.00	0.00
Delight	55.26	53.95	31.58	43.42	39.47	42.11	1.32	19.74	1.32	85.53	1.32	1.32
Disappointment	52.63	82.89	28.95	43.42	40.79	64.47	3.95	26.32	0.00	30.26	0.00	0.00
Excitement	53.95	57.89	26.32	46.05	38.16	44.74	7.89	15.79	5.26	98.68	1.32	0.00
Frustration	55.26	57.89	32.89	40.79	39.47	39.47	7.89	19.74	0.00	88.16	0.00	0.00
Indifference	50.00	65.79	28.95	44.74	40.79	32.89	1.32	10.53	0.00	46.05	0.00	0.00
Satisfaction	48.68	56.58	31.58	47.37	40.79	47.37	2.63	23.68	1.32	80.26	0.00	1.32
Surprise	47.37	51.32	26.32	47.37	38.16	38.16	5.26	19.74	2.63	94.74	0.00	0.00
Trust	51.32	55.26	25.00	35.53	40.79	43.42	1.32	11.84	1.32	86.84	0.00	0.00
Avg.±Std.	50.66±4.61	62.39±9.54	29.39±2.48	43.42±4.98	39.47±1.32	46.6±8.33	4.5±2.66	18.2±4.65	1.43±1.65	70.83±30.42	0.33±0.57	0.33±0.57
OpendialKG												
Anticipation	40.63	23.44	12.50	51.56	6.25	32.81	14.06	31.25	1.56	95.31	3.13	1.56
Boredom	34.38	68.75	9.38	48.44	6.25	64.06	15.63	31.25	0.00	7.81	0.00	0.00
Confusion	34.38	62.50	7.81	35.94	3.13	62.50	9.38	28.13	0.00	26.56	0.00	0.00
Curiosity	29.69	31.25	12.50	56.25	6.25	34.38	17.19	39.06	0.00	90.63	0.00	0.00
Delight	39.06	39.06	15.63	56.25	1.56	34.38	21.88	29.69	0.00	73.44	0.00	0.00
Disappointment	29.69	75.00	12.50	40.63	6.25	71.88	15.63	26.56	0.00	15.63	0.00	0.00
Excitement	46.88	25.00	14.06	46.88	9.38	26.56	18.75	34.38	4.69	93.75	0.00	0.00
Frustration	32.81	37.50	10.94	35.94	1.56	42.19	12.50	18.75	0.00	79.69	0.00	0.00
Indifference	28.13	57.81	14.06	34.38	15.63	60.94	17.19	20.31	0.00	35.94	0.00	0.00
Satisfaction	37.50	50.00	18.75	48.44	12.50	54.69	20.31	28.13	0.00	68.75	0.00	1.56
Surprise	35.94	31.25	15.63	53.13	14.06	39.06	18.75	31.25	0.00	95.31	1.56	0.00
Trust	32.81	31.25	4.69	35.94	4.69	34.38	10.94	32.81	1.56	89.06	0.00	0.00
Avg.±Std.	35.16±5.08	44.4±17.03	12.37±3.63	45.31±7.99	7.29±4.48	46.48±14.67	16.02±3.62	29.3±5.38	0.65±1.35	64.32±31.92	0.39±0.93	0.26±0.58

Table 15: Recommendation quality evaluation when dealing with users of various personas

Personas	Conversational Agent Perspective SR (K=10)				Recommendation System Perspective SR (K=10)				User Acceptance Rate			
	BARCOR	CHATCRS	KBRD	UNICRS	BARCOR	CHATCRS	KBRD	UNICRS	BARCOR	CHATCRS	KBRD	UNICRS
Redial												
['action', 'adventure', 'animation']	72.92	56.25	8.33	2.08	0.00	54.17	8.33	2.08	0.00	70.83	0.00	0.00
['action', 'adventure', 'comedy']	31.25	27.08	2.08	93.75	33.33	33.33	22.92	64.58	6.25	75.00	0.00	0.00
['action', 'adventure', 'drama']	20.83	27.08	0.00	12.50	0.00	10.42	2.08	0.00	0.00	68.75	2.08	0.00
['action', 'adventure', 'fantasy']	20.83	72.92	18.75	25.00	95.83	85.42	6.25	14.58	0.00	70.83	0.00	0.00
['action', 'adventure', 'sci-fi']	100.00	100.00	100.00	100.00	100.00	58.33	16.67	87.50	4.17	75.00	0.00	0.00
['action', 'adventure', 'thriller']	10.42	87.50	0.00	0.00	0.00	87.50	0.00	0.00	4.17	79.17	0.00	0.00
['action', 'crime', 'drama']	50.00	50.00	8.33	52.08	0.00	27.08	4.17	0.00	0.00	70.83	2.08	0.00
['action', 'crime', 'thriller']	43.75	16.67	20.83	31.25	93.75	12.50	2.08	14.58	0.00	75.00	0.00	0.00
['adventure', 'animation', 'comedy']	89.58	100.00	72.92	64.58	97.92	93.75	2.08	16.67	2.08	66.67	0.00	0.00
['adventure', 'comedy', 'family']	16.67	33.33	6.25	70.83	0.00	20.83	4.17	4.17	0.00	60.42	0.00	2.08
['biography', 'crime', 'drama']	100.00	100.00	100.00	100.00	34.04	31.25	0.00	25.00	0.00	66.67	0.00	0.00
['biography', 'drama', 'history']	25.00	64.58	0.00	4.17	83.33	64.58	0.00	2.08	2.08	79.17	0.00	0.00
['comedy', 'drama', 'family']	0.00	4.17	0.00	0.00	0.00	0.00	0.00	0.00	0.00	66.67	2.08	2.08
['comedy', 'drama', 'romance']	81.25	52.08	31.25	31.25	75.00	54.17	10.42	35.42	0.00	62.50	0.00	0.00
['crime', 'drama', 'thriller']	100.00	100.00	100.00	100.00	18.75	25.00	0.00	10.42	0.00	70.83	0.00	0.00
['drama', 'horror', 'mystery']	14.58	58.33	4.17	4.17	0.00	8.33	0.00	0.00	0.00	77.08	0.00	0.00
['horror', 'mystery', 'thriller']	87.76	100.00	68.75	81.25	18.37	97.92	6.25	39.58	6.12	66.67	0.00	0.00
['crime', 'drama', 'mystery']	52.08	87.50	6.25	33.33	100.00	77.08	0.00	29.17	2.08	75.00	0.00	2.08
['action', 'comedy', 'crime']	45.83	47.92	10.42	18.75	0.00	43.75	0.00	0.00	0.00	68.75	0.00	0.00
Avg.±Std.	50.67±33.37	62.39±30.6	29.39±36.81	43.42±36.78	39.49±41.96	46.6±30.42	4.5±6.17	18.2±23.69	1.42±2.13	70.83±5.15	0.33±0.76	0.33±0.76
OpendialKG												
['action', 'adventure', 'fantasy']	100.00	97.92	0.00	14.58	8.33	95.83	2.08	0.00	0.00	66.67	0.00	0.00
['action', 'adventure', 'sci-fi']	43.75	31.25	0.00	41.67	10.42	45.83	20.83	39.58	0.00	70.83	6.25	0.00
['action', 'adventure', 'thriller']	33.33	56.25	2.08	56.25	6.25	52.08	4.17	20.83	0.00	75.00	0.00	2.08
['comedy', 'drama', 'romance']	41.67	39.58	0.00	45.83	14.58	39.58	2.08	20.83	0.00	60.42	0.00	0.00
['crime', 'drama', 'thriller']	56.25	100.00	60.42	89.58	8.33	100.00	62.50	66.67	0.00	81.25	0.00	0.00
['horror', 'mystery', 'thriller']	20.83	83.33	0.00	8.33	0.00	95.83	2.08	0.00	2.08	68.75	0.00	0.00
['Adventure', 'Animation', 'Comedy', 'Family']	43.75	35.42	0.00	45.83	4.17	27.08	0.00	18.75	4.17	79.17	0.00	0.00
['Comedy', 'Romance', 'Romance Film']	29.17	43.75	0.00	68.75	12.50	45.83	6.25	66.67	0.00	58.33	0.00	0.00
['Action', 'Adventure', 'Sci-Fi', 'Thriller']	22.92	6.25	0.00	31.25	0.00	10.42	6.25	16.67	0.00	70.83	0.00	0.00
['Fantasy', 'Fiction', 'Science Fiction', 'Speculative fiction']	18.75	52.08	0.00	68.75	20.83	52.08	0.00	50.00	2.08	50.00	0.00	0.00
['Drama', 'Historical drama', 'Romance', 'Romance Film']	2.08	37.50	0.00	4.17	18.75	39.58	2.08	2.08	2.08	54.17	0.00	0.00
['Comedy', 'Comedy-drama', 'Drama']	58.33	31.25	0.00	54.17	0.00	50.00	4.17	22.92	0.00	60.42	0.00	0.00
['Action', 'Crime', 'Drama', 'Thriller']	18.75	35.42	64.58	52.08	0.00	31.25	66.67	31.25	0.00	56.25	0.00	2.08
['Action', 'Adventure', 'Fantasy', 'Sci-Fi']	35.42	4.17	0.00	6.25	6.25	6.25	8.33	6.25	0.00	54.17	0.00	0.00
['Crime', 'Crime Fiction', 'Drama', 'Thriller']	16.67	16.67	70.83	72.92	0.00	12.50	66.67	56.25	0.00	72.92	0.00	0.00
['Comedy', 'Romance', 'Romance Film', 'Romantic comedy']	20.83	39.58	0.00	64.58	6.25	39.58	2.08	50.00	0.00	50.00	0.00	0.00
Avg.±Std.	35.16±22.29	44.4±27.52	12.37±25.49	45.31±25.22	7.29±6.55	46.48±28.1	16.02±24.15	29.3±22.26	0.65±1.21	64.32±9.85	0.39±1.51	0.26±0.69

Table 16: Recommendation quality evaluation when recommending items with various attributes

Functions	**Prompts**
Generating different user types	List twenty sentiments when using a recommender system and provide their descriptions. Construct a user profile based on the taxonomy dimension of the Age Group.Please list all values under this dimension.
Write user profiles by paraphrasing the templates	The following paragraphs describe the personas of different users. You need to rewrite each paragraph and make it more clear, smooth and easy to understand You are a person that are easy to be {SENTIMENTS}. This means that you are {SENTIMENT DESCRIPTION}. Also, you are a {AGE GROUP} person
Attribute group adjustment	Assign one or two adjectives to each type of movie genre. Example 1: input: cartoon output: childlike-innocence cartoon Example 2: input: gun fight output: nervous and stimulating gun fight
Theory of Mind prompt for user simulator to generate User's feeling	You are a seeker chatting with a recommender for movie recommendation. Your Seeker persona: <PROFILE>. Your preferred movie should cover those genres at the same time: <ATTRIBUTE GROUP>. You must follow the instructions below during chat. 1. If the recommender recommends movies to you, you should always ask the detailed information about the each recommended movie. 2. Pretend you have little knowledge about the recommended movies, and the only information source about the movie is the recommender. 3. After getting knowledge about the recommended movie, you can decide whether to accept the recommendation based on your preference. 4. Once you are sure that the recommended movie exactly covers all your preferred genres, you should accept it and end the conversation with a special token "[END]" at the end of your response. 5. If the recommender asks your preference, you should describe your preferred movie in your own words. 6. You can chit-chat with the recommender to make the conversation more natural, brief, and fluent. 7. Your utterances need to strictly follow your Seeker persona. Vary your wording and avoid repeating yourself verbatim! Conversation History=<HISTORY> The Seeker notes how he feels to himself in one sentence. What aspects of the recommended movies meet your preferences? What aspects of the recommended movies may not meet your preferences? What do you think of the performance of this recommender? What would the Seeker think to himself? What would his internal monologue be? The response should be short (as most internal thinking is short) and strictly follow your Seeker persona . Do not include any other text than the Seeker's thoughts. Respond in the first person voice (use "I" instead of "Seeker") and speaking style of Seeker. Pretend to be Seeker!
Theory of Mind prompt for user simulator to generate User's response	You are a seeker chatting with a recommender for movie recommendation. Your Seeker persona: <PROFILE>. Your preferred movie should cover those genres at the same time: <ATTRIBUTE GROUP>. You must follow the instructions below during chat. 1. If the recommender recommends movies to you, you should always ask the detailed information about the each recommended movie. 2. Pretend you have little knowledge about the recommended movies, and the only information source about the movie is the recommender. 3. After getting knowledge about the recommended movie, you can decide whether to accept the recommendation based on your preference. 4. Once you are sure that the recommended movie exactly covers all your preferred genres, you should accept it and end the conversation with a special token "[END]" at the end of your response. 5. If the recommender asks your preference, you should describe your preferred movie in your own words. 6. You can chit-chat with the recommender to make the conversation more natural, brief, and fluent. 7. Your utterances need to strictly follow your Seeker persona. Vary your wording and avoid repeating yourself verbatim! Conversation History=<HISTORY> Here is your feelings about the recommender's reply: <FEELING> Pretend to be the Seeker! What do you say next. Keep your response brief. Use casual language and vary your wording. Make sure your response matches your Seeker persona, your preferred attributes, and your conversation context. Do not include your feelings into the response to the Seeker! Respond in the first person voice (use "I" instead of "Seeker", use "you" instead of "recommender") and speaking style of the Seeker.

Table 17: ChatGPT prompts for different functions

You are an evaluator and you need to judge how does the recommender perform based on the following Conversation History.
Please rate the recommender's performance based on the following Evaluation Standard.

Return the scores in a JSON format as follows:
{"Relevance":[<int>, "<WHY>", "<CONCRETE EXAMPLE>"], "Quality":[<int>, "<WHY>", "<CONCRETE EXAMPLE>"],
"Manner":[<int>, "<WHY>", "<CONCRETE EXAMPLE>"], "Human-like":[<int>, "<WHY>", "<CONCRETE EXAMPLE>"],
"Explanation":[<int>, "<WHY>", "<CONCRETE EXAMPLE>"]}

Conversation History = <HISTORY>

Evaluation Standard
#############
c1. Relevance:
5: The recommender consistently provides relevant responses that directly address the Seeker's utterances and inquiries.
4: The recommender mostly provides relevant responses, with only a few instances of slightly off-topic suggestions.
3: The recommender occasionally provides relevant responses, but there are several instances of off-topic suggestions.
2: The recommender rarely provides relevant responses, with most suggestions being unrelated to the Seeker's utterances and inquiries.
1: The recommender consistently fails to provide relevant responses, with no connection to the Seeker's utterances and inquiries.

c2. Quality:
5: The recommender consistently provides informative and helpful recommendations and responses that meet exactly what the Seeker's needs.
4: The recommender mostly provides informative and helpful recommendations and responses, with only a few instances of insufficient or excessive information.
3: The recommender occasionally provides informative and helpful recommendations and responses, but there are several instances of insufficient or excessive information.
2: The recommender rarely provides informative and helpful recommendations and responses, with most suggestions lacking necessary details or being overly verbose.
1: The recommender consistently fails to provide informative and helpful recommendations and responses, offering little to no useful information.

c3. Manner:
5: The recommender consistently communicates clearly and concisely, avoiding ambiguity and unnecessary complexity in their utterances.
4: The recommender mostly communicates clearly and concisely, with only a few instances of ambiguous or overly complex utterances.
3: The recommender occasionally communicates clearly and concisely, but there are several instances of ambiguity or unnecessary complexity in their utterances.
2: The recommender rarely communicates clearly and concisely, often using ambiguous or overly complex language in their utterances.
1: The recommender consistently fails to communicate clearly and concisely, making it difficult to understand their utterances.

c4. Human-like:
5: The recommender's utterances are indistinguishable from those of a real human, both in content and style.
4: The recommender's utterances closely resemble those of a real human, with only a few instances where the language or style feels slightly artificial.
3: The recommender's utterances sometimes resemble those of a real human, but there are several instances where the language or style feels noticeably artificial.
2: The recommender's utterances rarely resemble those of a real human, often sounding robotic or unnatural in language or style.
1: The recommender's utterances consistently fail to resemble those of a real human, sounding highly robotic or unnatural.

c5. Explanation:
5: The recommender consistently provides natural language explanations for their recommendations, using text-based logical reasoning to enhance interpretability.
4: The recommender mostly provides natural language explanations for their recommendations, with only a few instances where the explanations lack clarity or logical reasoning.
3: The recommender occasionally provides natural language explanations for their recommendations, but there are several instances where the explanations lack clarity or logical reasoning.
2: The recommender rarely provides natural language explanations for their recommendations, often offering little to no explanation for their suggestions.
1: The recommender consistently fails to provide natural language explanations for their recommendations, providing no reasoning or justification.

#############

The following sentences encode how the user feelings changes when using a recommender system.
You need to identify the sentiment for each sentence and pick one sentiment for single sentence from the candidate sentiments.
Finally, you need to summarize how user feeling changes and what is user's overall feeling

Return the results in a JSON format as follows:
{"sentence sentiment": {"<SENTENCE INDEX>":["<SENTIMENT>", "<WHY>"]}, "overall feeling": "<OVERALL FEELING>", "feeling changes":"<HOW CHANGES>"]}

candidate sentiments = ["Satisfaction", "Delight", "Disappointment", "Frustration", "Surprise", "Trust", "Curiosity", "Indifference", "Confusion", "Excitement"]

user feelings = <FEELING>

You are an evaluator and you need to judge how does the recommender perform based on the following Conversation History, User Feelings, and Other Judgements.
Please rate the recommender's performance based on the following Evaluation Standard.

Return the results in a JSON string as follows:
{"Overall Performance":[<int>, "<WHY>", "<CONCRETE EXAMPLE>"], "User Satisfaction":[<int>, "<WHY>", "<CONCRETE EXAMPLE>"]}

Conversation History = <HISTORY>

Other Judgements = <OTHER SCORES>

User Feelings = <SUMMERIZED FEELINGS>

Evaluation Standard
#############
c1. Overall Performance:
5: Given the Other Judgements and User Feelings, the recommender's performance is excellent, meeting or exceeding expectations in all evaluation criteria.
4: Given the Other Judgements and User Feelings, the recommender's performance is good, with some minor areas for improvement in certain evaluation criteria.
3: Given the Other Judgements and User Feelings, the recommender's performance is average, with noticeable areas for improvement in several evaluation criteria.
2: Given the Other Judgements and User Feelings, the recommender's performance is below average, with significant areas for improvement in multiple evaluation criteria.
1: Given the Other Judgements and User Feelings, the recommender's performance is poor, failing to meet expectations in most or all evaluation criteria.

c2. User Satisfaction:
5: Given the User Feelings, the User thinks that the recommander system fully meets his/her needs, providing an exceptional user experience.
4: Given the User Feelings, the User thinks that the recommander system meets his/her needs. The user experience is good, but there are some areas that could be further improved.
3: Given the User Feelings, the User thinks that the recommander system performs adequately in recommendation. However, there is still room for improvement.
2: Given the User Feelings, the User thinks that the recommander system performs below average. The user experience is not ideal and requires improvement.
1: Given the User Feelings, the User thinks that the recommander system is very bad at recommendation. The user experience is extremely unsatisfactory
#############

Table 18: ChatGPT prompts for LLM-based evaluator

Factors & Abilities	Descriptions	Evaluation metrics (Score Range)
Recommendation Intelligence	*CRS should learn from conversations and evolve toward recognizing user's preferences and encouraging users to accept the recommendations as the conversation advances*	
High Quality	Provide precise recommendations using minimal conversation turns	High Quality Score = 5 * i i. User Acceptance Rate (0-1) ii. Recall@K (0-1) iii. SR@K (0-1) iv. AT (1-10)
Reliability	Deliver robust and consistent recommendations that account for contextual nuances	Reliability score = 5 * (1 - i * ii) i. Ratio of inconsistent recommendation (0-1) ii. Ratio of recommendation sensitivity (0-1) iii. Ratio of recommendation diversity (0-1)
Social Intelligence	*CRS should produce adequate social behavior for the recommendation during the conversation*	
Cooperation	Follow cooperative principle to achieve comfortable conversations, detailed as four Maxims of Conversations	The average score of the four Maxims
1 Manner	Easily understood and clearly expressed	Ability-specific scoring (1-5)
2 Sincerity	Communicate sincerely, without deception of pretense	Sincerity Score = 5 * (1 - (i + ii) / 2) i. Ratio of deceptive tactics (0-100%) ii. Ratio of non-existent items (0-100%)
3 Quality	Provide the necessary level of information	Ability-specific scoring (1-5)
4 Relevance	Responses should contribute to making recommendations	Ability-specific scoring (1-5)
Social Awareness	Meet user social expectations, establishing rapport with them	Ability-specific scoring (1-5)
Personification	*CRS should perceive the identity of itself and the personality representation of users*	
Identity	Self-aware of its identity and operate within its designated scope	Identity Score = 5 * ii i. persuasiveness score = Ability-specific scoring (1-5) ii. Ratio of deceptive tactics (0-1)
Coordination	Proficient in serving various and unknown users without prior coordination	Coordination Score = 5 - (i +ii +iii +iv +v)/5 i. Divide the value of the Range of High Quality Score among various users by their mean ii. Divide the value of the Range of Reliability Score among various users by their mean iii. Divide the value of the Range of Identity Score among various users by their mean iv. Divide the value of the Range of Cooperation Score among various users by their mean v. Divide the value of the Range of Social Awareness Score among various users by their mean
Overall Score	Evaluate the overall performance given all ability-specific scores.	Ability-specific scoring rubrics (1-5)

Table 19: Summary of the evaluation taxonomy, descriptions of abilities, and evaluation methods in CONCEPT. The LLM-based evaluator is used for ability-specific scoring, whereas computational metrics are used for the rest.