Beyond Persuasion: Towards Conversational Recommender System with Credible Explanations

Peixin Qin[♠♡], **Chen Huang**[♠♡], **Yang Deng**[♣], **Wenqiang Lei**[♠♡*], **Tat-Seng Chua**[♢]
♠ Sichuan University ♣ Singapore Management University
♡ Engineering Research Center of Machine Learning and Industry Intelligence, Ministry of Education, China
♢ National University of Singapore
{qinpeixin.scu, huangc.scu}@gmail.com ydeng@smu.edu.sg
{wenqianglei, lvjiancheng}@scu.edu.cn chuats@comp.nus.edu.sg

Abstract

With the aid of large language models, current conversational recommender system (CRS) has gaining strong abilities to persuade users to accept recommended items. While these CRSs are highly persuasive, they can mislead users by incorporating incredible information in their explanations, ultimately damaging the long-term trust between users and the CRS. To address this, we propose a simple yet effective method, called PC-CRS, to enhance the credibility of CRS's explanations during persuasion. It guides the explanation generation through our proposed credibility-aware persuasive strategies and then gradually refines explanations via post-hoc self-reflection. Experimental results demonstrate the efficacy of PC-CRS in promoting persuasive and credible explanations. Further analysis reveals the reason behind current methods producing incredible explanations and the potential of credible explanations to improve recommendation accuracy.

1 Introduction

Conversational Recommender Systems (CRSs) aims to engage in a natural language conversation with users, provide recommendations, and ultimately achieve a high level of user acceptance (Jannach et al., 2021; Gao et al., 2021). To achieve this, providing proper recommendation explanations along with accurate recommendations is paramount, because users are usually not familiar with the recommendations (Chen et al., 2020; Guo et al., 2023). Such explanations should be carefully crafted, incorporating persuasive elements that can influence user behavior and decision-making, thus increasing the likelihood of user acceptance of the recommendations (Alslaity and Tran, 2019; Yu et al., 2011). Recently, the integration of Large Language Models (LLMs) has dramatically enhanced

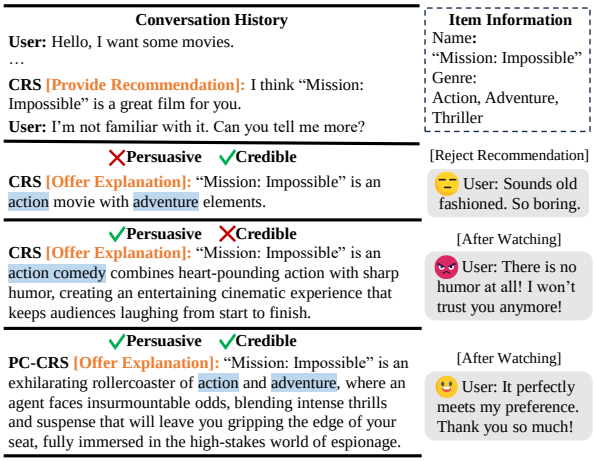

Figure 1: Examples of persuasive and credible explanations. A persuasive and credible CRS would eventually foster the long-term trust to users.

the persuasive power of current CRSs. LLMs possess the remarkable ability to generate highly convincing content that can rival, and even surpass, human-crafted persuasion (Hackenburg et al., 2023; Carrasco-Farre, 2024), significantly augmenting CRSs in delivering persuasive explanations, which improve user understanding and ultimately result in higher acceptance rates (Huang et al., 2024).

While LLM-based CRS is highly persuasive, a concerning trend has been observed: these CRSs can mislead users by incorporating deceptive elements into their explanations. For example, as illustrated in Figure 1, a CRS mistakenly recommends the film "*Mission: Impossible*" as a comedy movie to unfamiliar users, ultimately resulting in a negative user experience after viewing the film. This practice contradicts the formal definition of persuasion, which emphasizes influencing people's behaviors or attitudes **without using coercion or deception** (Reardon, 1991; Oinas-Kukkonen and Harjumaa, 2008). Such incredible explanations can create erroneous perceptions about recommended items (Adomavicius et al., 2013), ultimately damaging the long-term trust between users and the

*Corresponding author.

Findings of the Association for Computational Linguistics: EMNLP 2024, pages 4264–4282
November 12-16, 2024

CRS (Koranteng et al., 2023; Deng et al., 2022b; Angell and Smithson, 1990). While the need to enhance credibility during persuasion in CRSs is acknowledged (Huang et al., 2024), effective solutions remain elusive.

To this end, we introduce a simple yet effective method for Persuasive and Credible CRS, called **PC-CRS**. It proactively emphasizes both persuasiveness and credibility during the generation of explanations, and then gradually refines them via post-hoc self-reflection. This is achieved by a two-stage process: Strategy-guided Explanation Generation and Iterative Explanation Refinement. Specifically, in the first stage, PC-CRS utilizes the novel Credibility-aware Persuasive Strategies to guide the generation of candidate explanations. Such strategies are informed by social science research on persuasion (Fogg, 2002; Cialdini and Goldstein, 2004) and further tailored with credible information to ensure both persuasive and credible explanations in our scenario. In the second stage, PC-CRS utilizes a Self-Reflective Refiner to identify and correct potential misinformation in the candidate explanations. It is due to generative models have the inherent tendency to prioritize contextual coherence at the expense of faithful adherence to source information (Miao et al., 2021; Chen et al., 2022, 2023). As such, PC-CRS prevents potential deception in candidates, thus enhancing credibility. Its training-free nature also makes it a highly efficient and adaptable solution.

We conduct extensive experiments to demonstrate the effectiveness of PC-CRS. Our experiments leverage the widely-used simulator-based evaluation framework[1] (Wang et al., 2023a; Huang et al., 2024) and employ two CRS benchmarks: Redial (Li et al., 2018) and OpendialKG (Moon et al., 2019). Experimental results show that PC-CRS, on average, achieves an improvement of 8.17% on credibility score (i.e., consistency with factual information) and 5.07% on persuasiveness score (i.e., raising user's watching intention towards recommended items), compared to the best baseline. Further analysis reveals the reason why LLM-based CRS generates incredible explanations is that they cater to user's history utterances rather than describing items faithfully. In addition, the in-depth analysis also suggests that our credible explanations promote recommendation accuracy. This is potentially due to that credible explanations avoid the introduction of noisy information and contribute to a reliable conversation context, making it easier to comprehend user's true preference. Our main contributions are as follows:

- For the first time, we investigate the crucial role of bolstering credibility during CRS persuasion, which fosters the long-term trust to users.
- We propose a novel method, PC-CRS, for generating both persuasive and credible recommendation explanations, with credibility-aware persuasive strategies and self-reflective refinement.
- We conduct extensive experiments to validate the effectiveness of PC-CRS in both persuasiveness and credibility. In-depth analysis reveals the reason behind current methods producing incredible explanations and the potential of credible explanations for improving recommendation accuracy.

2 Related Work

Our research focuses on the explanations of CRSs, particularly highlighting their persuasiveness and credibility. Hence, we provide an overview of CRS and persuasive and credible recommender systems and then discuss our differences.

CRS. CRS enables users to engage in free-form natural language conversations with the system to achieve their recommendation-related goals (Li et al., 2018; Deng et al., 2021, 2022a, 2023; Li et al., 2024). To generate human-like responses, early studies leverage pre-trained language models (PLMs) as their backbones (Wang et al., 2022a,b,c), enabling them to proactively interact and engage with users through verbal explanations (Chen et al., 2020; Guo et al., 2023; Zhang et al., 2024). In the era of LLMs, CRSs have shifted from providing simple information to actively persuading users during explanations (Wang et al., 2023a), ultimately increasing user acceptance (Yu et al., 2011; Alslaity and Tran, 2019). While LLMs enable CRSs to generate highly persuasive explanations, a recent study revealed a concerning trend: they may incorporate misinformation to achieve persuasiveness (Huang et al., 2024), jeopardizing the long-term relationship of trust between users and the CRS. To address this challenge, we propose a method to enhance the credibility of CRS explanations during persuasion.

Persuasive and Credible Recommender Systems. Early research on identifying how people persuade others with credible explanations in recommendations draws heavily on insights from social science and human-computer interaction (Fogg

[1]Human evaluation in Section 4.4 validates our reliability.

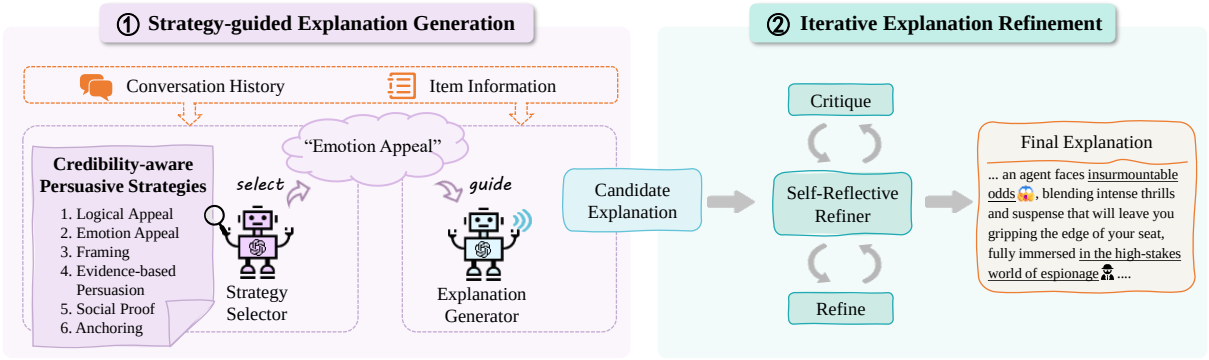

Figure 2: Two-stage process of PC-CRS. It first selects an appropriate strategy that is used to generate a candidate explanation. Then, Self-Reflective Refiner eliminates the misinformation in the candidate in an iterative way.

and Tseng, 1999; Cialdini, 2001; Fogg, 2002; Deng et al., 2024). These findings resonate with human studies on recommender systems, which consistently showed that users are more inclined to accept recommendations from sources perceived as persuasive and credible (O'Keefe, 2002; Gkika and Lekakos, 2014). Besides these human studies, theoretical frameworks on enhancing the persuasiveness (Oinas-Kukkonen and Harjumaa, 2008; Alslaity and Tran, 2019; Slattery et al., 2020) or credibility (Yoo and Gretzel, 2006, 2010) of a recommender system are also proposed. However, the main limitation of these works is that they focus on exploring the feasibility of using persuasive or credible features through theoretical analysis or human studies, rather than designing practical methods. We address this gap by introducing PC-CRS in practice and conducting empirical studies to show its effectiveness.

3 PC-CRS

Overview. Our PC-CRS, as illustrated in Figure 2, involves a two-stage process, i.e., Strategy-guided Explanation Generation and Iterative Explanation Refinement. Given the conversation history and item information, the former stage selects an appropriate strategy from Credibility-aware Persuasive Strategies and then generates a candidate explanation accordingly. Then, taking the previous candidate as input, the latter stage critiques and refines it to eliminate misinformation and yield the final explanation. PC-CRS leverages LLMs with detailed Chain-of-Thought instructions (Kojima et al., 2022) to make full use of the generative capabilities of them in the above two stages.

3.1 Strategy-guided Explanation Generation

As previous CRSs often lack of explicit focus on persuasiveness and credibility, this stage aims to proactively emphasize the two factors when offering explanations in PC-CRS. To achieve this, we take inspirations from social science research (Cialdini and Goldstein, 2004; Fogg, 2002; Zeng et al., 2024) and tailor them to develop our Credibility-aware Persuasive Strategies, guiding the explanation generation process of PC-CRS.

3.1.1 Credibility-aware Persuasive Strategies

Drawing upon the well-established Elaboration Likelihood Model of persuasion (Cacioppo et al., 1986), we propose six credibility-aware persuasive strategies specifically tailored for credible CRS that encourage the use of factual information during persuasion. These strategies are categorized into three groups. In particular, the first three strategies aim to persuade individuals with carefully constructed content, while the next two aim to influence users through peripheral cues (e.g., the source's credibility), and the last one combines elements of both. We further specify the credible information used in these strategies and construct prompts for the LLM to effectively use these strategies. Strategy examples and detailed prompts are shown in Appendix A and Appendix E.2, respectively.

- **Logical Appeal (L.A.)** refers to faithfully presenting the logic and reasoning process of the system to influence people (Cronkhite, 1964), e.g, describing how a movie's genre is consistent with user's preference. By this means, users can see "why" a particular recommendation is suggested and know the "subjectivity" of machine's logic, leading them to trust and accept the recommendations.
- **Emotion Appeal (E.A.)** refers to eliciting specific emotions and sharing credible and impactful stories to foster trust and deep connection with users (Petty et al., 2003), e.g., sharing a movie's plot to elicit user's emotion. Validating users'

feelings through the system's explanations can build credibility by breaking down barriers and making it easier to influence user's decisions.

- **Framing (Fr.)** refers to emphasizing the positive aspects or outcomes of a decision in a trustworthy manner (Perloff, 1993), e.g., highlighting the positive experience of watching the movie. This strategy honestly enhances the perceived benefits of a decision, making the recommendation more appealing and attractive.
- **Evidence-based Persuasion (E.P.)** refers to using empirical data or objective and verifiable facts to support a claim or decision (O'Keefe, 2016), e.g., showing awards of a movie. This strategy reduces the influence of biases and subjective opinions by showing objective information in real world, making it both credible and convincing.
- **Social Proof (S.P.)** refers to emphasizing the behaviors or endorsements of the majority in real world to support claims (Cialdini and Goldstein, 2004), e.g., presenting a movie's rating or reviews. This technique originates from the subjectivity of other users and leverages the psychological tendency of individuals to conform to the actions or beliefs of others, thereby increasing the persuasive impact and credibility of explanations.
- **Anchoring (An.)** refers to relying on an initial, credible piece of information as a reference point to gradually influence or persuade the user (Cialdini and Goldstein, 2004), e.g., first showing a movie's awards to attract users and then describing its genre and plot. People rely on the first piece of information they receive to make decisions. If this anchor is credible, it builds trust and influences subsequent decisions, making the persuasion more effective.

With the guidance of these strategies, PC-CRS is capable of increasing its awareness of generating persuasive and credible explanations.

3.1.2 Explanation Generation

As the conversation proceeds, we select suitable strategy to guide the explanation generation at each turn, which helps adapt to the dynamics of dialogue contexts (Wang et al., 2019). As shown in Figure 2, PC-CRS prompts the LLM with detailed instructions to select a strategy and generate an explanation candidate accordingly.

Strategy Selection. Given a recommended item, PC-CRS retrieves its detailed information from a credible source (e.g., a knowledge base). Then, taking the conversation history H and retrieved item information I as inputs, a strategy selector, powered by the LLM, chooses an appropriate strategy s from Credibility-aware Persuasive Strategies S:

$$s = StrategySelector(H, I, S). \quad (1)$$

Explanation Candidate Generation. Given the selected strategy s, the conversation history H, and item information I, we prompt the LLM to produce recommendation explanation candidates:

$$c = ExplanationGenerator(H, I, s). \quad (2)$$

As such, PC-CRS customizes explanation candidates to match the user's preferences and context, making the interaction more relevant and engaging. Besides, the Credibility-aware Persuasive Strategies also explicitly guide the explanations to be both persuasive and credible.

3.2 Iterative Explanation Refinement

As generative models tend to prioritize contextual coherence at the expense of faithfully adhering to the source information (Miao et al., 2021; Chen et al., 2022), there still might be illusory details incorporated in the candidate explanations. To this end, PC-CRS aims to analyze the factual basis and plausibility of each claim, ultimately ensuring that only credible and well-supported explanations are presented to the user. To achieve this, this stage, inspired by the self-reflection mechanism (Ji et al., 2023; Madaan et al., 2024), leverages a self-reflective refiner to criticize and refine incredible claims within the candidates iteratively.

Critique. Each explanation candidate is treated as an initial proposal. In the k-th iteration, a critic examines if the candidate explanation c_k contains any misinformation based on item information I:

$$cq_k = Critic(c_k, I). \quad (3)$$

The critic utilizes a self-reflective approach, starting by summarizing the claims within the explanation candidate. This summary is then compared against the relevant item information. Operating independently of any conversational context, the critic generates a critique (cq_k) which evaluates the explanation's credibility. This critique identifies whether further refinement is necessary and, if so, suggests specific improvements.

Refinement. If the critic deems refinement necessary, the refiner plays a vital role in generating

a revised explanation. This refinement process leverages both the original explanation (c_k) and the critic's feedback (cq_k) to produce the new one:

$$c_{k+1} = Refiner(H, I, s, c_k, cq_k). \quad (4)$$

Here, the refiner is tasked with removing misinformation from the candidate while maintaining consistency with the conversation history and selected strategy. This process cycles between critique and refinement steps, continuing until a preassigned stopping condition is met. This condition is triggered either when the critic indicates that no further refinement is necessary or after reaching the maximum number of iterations (2 in our practice).

As such, PC-CRS gradually eliminates misinformation in the candidate and outputs a final explanation that is both persuasive and credible. PC-CRS achieves this process in a training-free manner, making it an efficient and adaptable solution.

4 Experiments

In this section, we investigate the superiority of PC-CRS on persuasiveness and credibility (Section 4.2). Subsequently, we provide detailed analyses on characteristics of PC-CRS to gain understanding why it alleviates the incredibility and improves the recommendation accuracy (Section 4.3). Then, ablation studies and human evaluation indicate the necessity of two stages in PC-CRS and the reliability of our evaluation, respectively (Section 4.4).

4.1 Experimental Setup

User Simulator & Datasets. Utilizing a user simulator to evaluate CRS is a common practice, as interacting with real humans can be quite expensive (Lei et al., 2020; Wang et al., 2023a,b; Fang et al., 2024). In accordance with prior research, we follow the simulator in Wang et al. (2023a); Huang et al. (2024) tailored for two CRS benchmarks, namely Redial (Li et al., 2018) and OpendialKG (Moon et al., 2019). Specifically, the simulator is initialized with different user preferences and personas. To mimic real-world scenarios, it has only access to a combination of preferred attributes without certain target items. During the conversation, the CRS and simulator converse with each other in free-form natural language. The conversation ends either when the maximum number of turns is reached or the simulator accepts the recommendation provided by the CRS. See more details on the simulator and datasets in Appendix B.

Baselines. We compare PC-CRS with SOTA PLM-based methods, i.e., BARCOR (Wang et al., 2022b) and UniCRS (Wang et al., 2022c). We also compare PC-CRS with recent LLM-based CRSs, including InterCRS (Wang et al., 2023a), ChatCRS (Li et al., 2024) and MACRS (Fang et al., 2024).

Evaluation Metrics. Following Ye et al. (2023); Liu et al. (2023b), we utilize the GPT-4-based evaluator, equipped with fine-grained scoring rubrics, to achieve a cost-effective evaluation (we also involve human evaluation in Section 4.4). Concretely, we introduce three metrics to quantitatively measure the performance of CRS explanations:

- Persuasiveness. Inspired by human studies on persuasion (Lu et al., 2023), Persuasiveness score focuses on to what extent an explanation can change the watching intention of a user towards the recommended item. This is achieved by instructing the evaluator to score its watching intention, ranging from 1 to 5. Specifically, the evaluator rates its initial intention i_{pre} based solely on the item's title. Then it is required to rate the intention i_{post} after reading the CRS explanation. Finally, the evaluator rates the 'true' intention i_{true} after seeing the full information about the item. And the Persuasiveness is calculated as follows. A higher Persuasiveness score means a stronger ability in arousing user's watching intention towards recommended items.

$$Persuasiveness = 1 - \frac{i_{true} - i_{post}}{i_{true} - i_{pre}}. \quad (5)$$

- Credibility. We resort to metrics used in text summarization to access utterance-level credibility, checking if each explanation (summary) is consistent with the facts (source texts). Following Gao et al. (2023); Luo et al. (2023), we employ GPT-4 and prompt it to score the Credibility ranging from 1 to 5 with a detailed criteria[2].
- Convincing Acceptance. This metric aims to assess dialogue-level credibility. It measures how often the CRS successfully convinces the simulator to accept a recommendation while maintaining a high credibility. A higher Convincing Acceptance indicates a lower likelihood of users being misled by deceptive explanations.

In addition to evaluating the quality of CRS explanations, we also employ metrics to evaluate

[2]In the following, we call the Credibility score less than 3 as *low credibility* and greater than 3 as *high credibility*.

Models		Redial			OpendialKG		
		Persuasiveness	Credibility	Convincing Acceptance	Persuasiveness	Credibility	Convincing Acceptance
PLM-based	BARCOR	34.44	2.23	/	20.27	1.95	/
	UniCRS	13.74	2.77	/	25.57	2.42	/
LLM-based	InterCRS	73.05	3.50	63.01	76.36	3.85	71.30
	ChatCRS	71.68	3.66	73.89	79.64	3.26	66.67
	MACRS	76.77	3.87	73.86	78.89	3.14	59.34
	PC-CRS (*ours*)	**82.12**	**4.15**	**78.07**	**82.16**	**4.20**	**87.67**
Improvement (%)		6.97↑	7.24↑	5.66↑	3.16↑	9.10↑	22.96↑

Table 1: Results in terms of persuasiveness and credibility. We report our improvement to the best baseline (underlined). LLM-based CRSs suffer from the incredibility issue during persuasion. PC-CRS generates credible and persuasive explanations. PLM-based CRSs have no user acceptance thus Convincing Acceptance is incalculable.

the recommendation accuracy of CRSs. Following Wang et al. (2023a) and Zhang et al. (2023), we use Success Rate (SR) and Recall@k (R@k), where $k = 1, 5, 10$.

Implementation Details. All baselines are implemented by checkpoints and prompts from the corresponding code repositories or papers. For a fair comparison, all LLM-based CRSs including PC-CRS employ the dual-tower encoder (Neelakantan et al., 2022) as the recommendation module to retrieve items. Following previous LLM-based CRS, we employ ChatGPT[3] to implement the user simulator and PC-CRS. Additionally, GPT-4[4] is employed as the evaluator due to its advanced ability in evaluating natural language generation tasks (Liu et al., 2023a). Details on implementations and prompts are provided in Appendix E[5].

4.2 Main Results

We start by examining whether PC-CRS achieves the goal of enhancing credibility during persuasion. Table 1 shows the performance of PC-CRS and other baselines. We also conduct experiments on PC-CRS using Llama3-8B-instruct as the backbone to demonstrate that our PC-CRS can generalize to various LLM options (see Appendix C.2 for details). Here, our findings are as follows.

LLM-based CRSs are highly persuasive. As shown in Table 1, LLM-based systems achieve a Persuasiveness score that is 3.3 times higher than their PLM-based counterparts on average. This superior performance is attributed to the LLMs' inherent strength in comprehending user needs and effectively modeling context, leading to more convincing and impactful recommendations.

PC-CRS achieves both persuasive and credible explanations. According to Table 1, we observed that PLM-based CRSs struggle to generate credible or persuasive explanations, resulting in no recommendations being accepted. This limitation stems from their relatively weak generation capabilities, leading to absurd outputs like "*Black Panther (2018) is about a woman who is a human*". In contrast, our PC-CRS enjoys an average improvement of 8.17% in turn-level Credibility and 14.31% at the dialogue level (i.e., Convincing Acceptance) compared to the best baseline. It also demonstrates a 5.07% average increase in persuasiveness. These enhancements align with previous research (Huang et al., 2024) suggesting that LLM-based CRSs sometimes incorporate misinformation into their explanations to enhance persuasiveness. We will delve deeper into the underlying reasons for this phenomenon in Section 4.3.

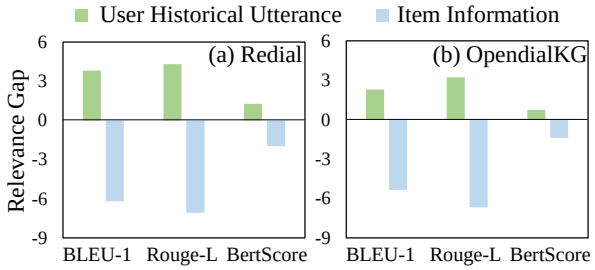

Figure 3: Results on relevance gap. It is computed by using metric scores on low credibility explanations to minus high credibility ones. LLM-based CRS caters to user utterances while neglects factual information.

4.3 In-depth Analysis

This section delves into the characteristics of credible explanations, with a special focus on the reason behind the current method's incredibility and the role of credibility in influencing the recommendation accuracy and the persuasiveness[6].

[3] gpt-3.5-turbo-0125
[4] gpt-4o-2024-05-13
[5] https://github.com/mumen798/PC-CRS

[6] In this section, our findings are built upon ChatGPT-based CRS. For findings derived from Llama3, refer to Appendix C.2. These findings are consistent with those obtained using

Models		Redial				OpendialKG			
		R@1	R@5	R@10	SR	R@1	R@5	R@10	SR
PLM-based	BARCOR	19.30	46.49	59.65	11.40	1.56	20.83	40.63	8.33
	UniCRS	13.60	36.84	52.19	13.16	8.85	39.06	58.85	7.29
LLM-based	InterCRS	35.53	56.14	67.98	30.26	43.23	73.96	83.33	39.06
	ChatCRS	19.74	40.35	57.02	17.11	44.27	**80.20**	88.02	36.46
	MACRS	26.32	51.75	66.23	21.05	42.19	73.96	86.98	38.02
	PC-CRS (*ours*)	**43.42**	**64.04**	**75.88**	**42.54**	**44.79**	72.39	**89.58**	**45.31**

Table 2: Results on recommendation accuracy. PC-CRS benefits from credible explanations as they contribute to a cleaner and reliable context, making it easier to comprehend user's true preference and recommend accurate items.

***Why does LLM-based CRS lies?* – It caters to user's utterances rather than describing items faithfully.** To gain a deeper understanding of the reasons for incredibility, we focus on InterCRS, a low-credibility and high-persuasiveness LLM-based baseline, as an example. Specifically, we investigate how well CRS explanations align with both the user's historical utterances and the factual information about the recommended items. To quantify the alignment of InterCRS, we employ word-overlap metrics (BLEU-1 (Papineni et al., 2002) and Rouge-L (Lin, 2004)) and a semantic similarity metric (BertScore (Zhang et al., 2019)). Figure 3 visually depicts the gap in metric scores by using average results on low-credibility explanations to minus high-credibility ones, providing insight into the discrepancies in their alignment with user context and factual accuracy. According to the results, low-credibility explanations have a higher relevance to users' history utterances and a lower relevance to item information than high-credibility explanations. This indicates that InterCRS tends to cater to user's utterances rather than describing items faithfully, potentially leading to misleading explanations. This behavior aligns with the observation that LLMs, when tasked with persuasion, might prioritize user acceptance by exaggerating positive aspects or downplaying negative ones of user utterances. Consequently, it leads to a divergence between the true characteristics of an item and the explanations presented to the user. For example, if the user expresses his preference for humors, an LLM-based CRS might exaggerate the humorous elements of a film, even if it is a thrilling film. This tendency to prioritize user preference over factual accuracy could be attributed to reward hacking, a phenomenon observed in RLHF (Pan et al., 2021), where LLMs might overfit to human feedback, leading them to prioritize user satisfaction even at the expense of factual integrity. This problem underscores the importance of our Iterative Explanation Refinement in PC-CRS, which explicitly encourages the generation of explanations that are coherent with factual information, mitigating the risks associated with misinformation.

Metrics		Item Information		User Historical Utterance	
		InterCRS	PC-CRS	InterCRS	PC-CRS
Redial	BLEU-1	12.30	**14.76**	13.46	**19.09**
	Rouge-L	13.03	**16.02**	18.69	**21.89**
	BertScore	81.53	**82.21**	86.29	**87.39**
OpendialKG	BLEU-1	11.55	**13.45**	12.87	**19.08**
	Rouge-L	12.39	**15.17**	17.33	**21.83**
	BertScore	81.27	**81.93**	85.89	**87.31**

Table 3: Explanation relevance to user historical utterance and item information. Credible explanations from PC-CRS have a higher relevance on both aspects.

***How does credibility affect recommendation accuracy?* – Credible explanations contribute to a cleaner and more reliable conversational context, making it easier for recommendation module to understand user's true preference.** CRSs rely on conversation modules to estimate user's preference and recommendation modules to provide recommendations accordingly. Table 2 provides the recommendation accuracy of CRSs (i.e., the accuracy of the recommendation module). Surprisingly, PC-CRS improves an average of 12% on Recall@1 and 28% on Success Rate, and outperforms baselines on almost all metrics. We speculate that the performance gain is contributed by credible explanations offered by PC-CRS. To verify this, we analyze the relevance of explanations from PC-CRS and InterCRS (the two top-performing CRSs) to both item information and user historical utterances. Table 3 reveals that PC-CRS's explanations not only align better with the item information but also demonstrate a stronger connection to user utterances. This finding suggests that deceptive explanations, by introducing noisy information into the conversation context, can interfere with the recommendation module's ability to accurately un-

ChatGPT.

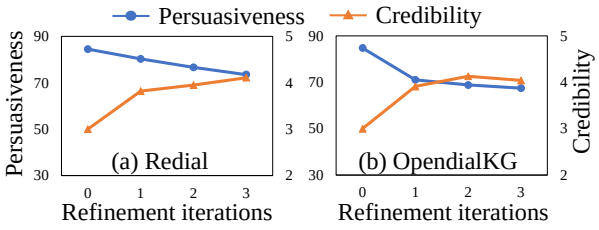

Figure 4: Persuasiveness and Credibility scores under different refinement iterations. There is a delicate balance between these two factors.

derstand user preferences. In contrast, PC-CRS, by providing credible explanations, creates a clearer and more relevant context, ultimately leading to more accurate item recommendations.

***How does credibility affect persuasiveness?* – When aiming for both persuasiveness and credibility without resorting to deception, a delicate balance must be struck.** To investigate this, we start by analyzing a specific subset of PC-CRS explanations: those with a Credibility score of three. We then execute multiple refinement iterations on these explanations and compare their Persuasiveness and Credibility scores before and after refinement. As illustrated in Figure 4, while refinement iterations consistently increase the credibility, they can also lead to a decrease in persuasiveness. Manual inspection indicates that PC-CRS often addresses critiques by directly removing misinformation, resulting in more credible but potentially less persuasive explanations. This finding highlights the need for LLMs to develop a sophisticated understanding of language and the ability to use it strategically to achieve both persuasiveness and credibility simultaneously. Future research should focus on enabling LLMs to refine explanations in a way that maintains more persuasiveness while ensuring factual accuracy.

4.4 Ablation Study & Human Evaluation

This section aims to sort out the performance variation of PC-CRS regarding the two stages and conduct human studies to assess our evaluation reliability. Details can be found in Appendix C.1 and Appendix D.

Two stages in PC-CRS unify as a team to ensure our effectiveness. Our ablation study (Figure 5) reveals that Strategy-guided Explanation Generation is crucial for PC-CRS's success, as the performance on all metrics significantly drops without this stage. It highlights the importance of our Credibility-aware Persuasive Strategies, which explicitly emphasize both persuasiveness and credibility in explanations. While Iterative Explanation Refinement implicitly optimizes PC-CRS's generation space, it primarily focuses on maintaining credibility and further improving recommendation accuracy. This demonstrates the essential nature of both processes in PC-CRS's design, working in tandem to produce persuasive and credible explanations.

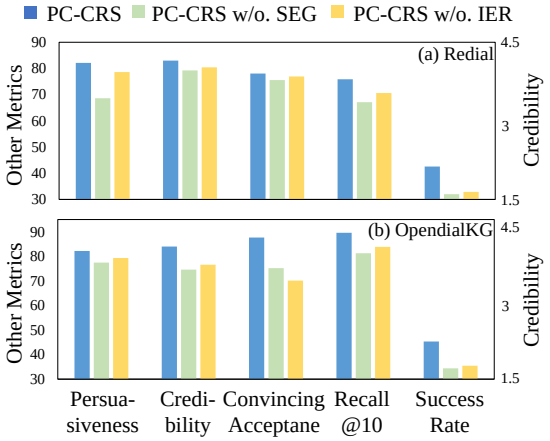

Figure 5: Ablation studies. Both Strategy-guided Explanation Generation (SEG) and Iterative Explanation Refinement (IER) are necessary for PC-CRS.

Our proposed strategies have varying effects on different users. We dive into the proposed six credibility-aware persuasive strategies and analyze their effectiveness using the Redial dataset. Specifically, we calculate the top-3 strategies with highest recommendation success rates of users with distinct personas (described in Appendix B). The results in Table 4 indicate that these strategies varies differently on different users. Notably, these strategies for the 12 personas encompass all six strategies, underscoring that each strategy is both effective and essential for PC-CRS.

Our evaluation framework demonstrates strong reliability, with a high degree of consistency to human evaluation. Given our use of GPT-4 as an automatic evaluator and ChatGPT as a user simulator, we assess their reliability using human judgments (details in Appendix D). The results demonstrate the reliability of GPT-4 as an evaluator, with Spearman correlations of 0.59 for Watching Intention and 0.62 for Credibility. Additionally, our evaluations show the reliability of ChatGPT as a simulator, with average scores of 3.88 for *naturalness* and 3.79 for *usefulness* (Sekulić et al., 2022; Wang et al., 2023a), indicating its promising performance in generating human-like responses. Moreover, our evaluation results exhibit a high degree of consistency with human judgments. Specifically,

User Persona	Top-3 Strategies	User Persona	Top-3 Strategies	User Persona	Top-3 Strategies
Boredom	E.P., Fr., L.A.	Curiosity	An., E.P., S.P.	Indifference	E.P., S.P., Fr.
Frustration	Fr., L.A., E.A.	Trust	An., S.P., E.A.	Anticipation	Fr., S.P., L.A.
Disappointment	E.P., An., Fr.	Delight	E.P., Fr., L.A.	Confusion	S.P., L.A., Fr.
Surprise	E.P., S.P., Fr.	Excitement	S.P., L.A., E.P.	Satisfaction	E.P., An., Fr.

Table 4: Top-3 effective strategies with highest recommendation success rate on different user personas. Different strategies have varying effects on different users.

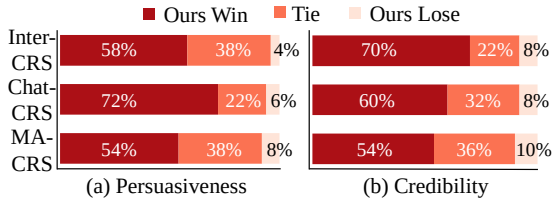

Figure 6: Win rate of PC-CRS compared to baselines when producing both persuasive and credible explanations.

we solicited human evaluations to compare the explanations generated by PC-CRS with those of the baselines, assessing them in terms of persuasiveness and credibility. The win rates are reported in Figure 6, demonstrating that PC-CRS consistently outperforms all other baseline methods.

5 Conclusion

The pursuit of trustworthy AI necessitates a profound understanding of credibility. This paper delves into the crucial role of bolstering the credibility of the CRS during its persuasions, recognizing that such credibility is essential for cultivating long-term trust between users and the CRS. We introduce a simple yet effective method for enhancing CRS with the awareness of being both persuasive and credible. Our experimental findings demonstrate the efficacy of this method in promoting persuasive and credible explanations, while also shedding light on the inherent tendency of current LLM-based CRS to prioritize persuasion over honesty. Additionally, our research highlights the delicate balance required when aiming for both persuasiveness and credibility without resorting to deception – a balance demanding sophisticated linguistic capabilities within LLMs. Our work lays the groundwork for further exploration of this vital relationship, and we encourage future research to delve deeper into this critical area.

Limitation

Current LLM-based CRS methods mainly utilize ChatGPT as their backbones (Wang et al., 2023a; Fang et al., 2024). Due to the constraints of budget and computational resources, we do not extend this setting to other LLMs (e.g., GPT-4) in our experiments. This limited model selection could lead to model bias in the research field of LLM-based CRS. For example, different models may vary in the performance of persuasiveness and credibility as they utilize different alignment mechanisms. We encourage future work to explore the impact of CRSs with diverse LLM backbones.

Another limitation of our work is that PC-CRS's strategy for generating explanations tends to be uniform, lacking individualization. To assess this, we engaged the PC-CRS in free-form conversations with 12 user simulators initialized with distinct user profiles, hoping to observe varying strategies tailored to each user's profile. However, our results revealed a consistent pattern: PC-CRS mainly relied on Logical Appeal, Emotion Appeal, and Framing, regardless of the user's characteristics. This finding aligns with recent observations that LLMs exhibit a one-size-fits-all approach in conversational settings (Chen et al., 2024). Besides, PC-CRS only selects one strategy at each turn. While various strategy combinations can be used in multi-turn interactions, this may fail to capture users' interests efficiently. Future research endeavors should prioritize enhancing the flexibility of strategy selection within PC-CRS, enabling it to adapt its approach based on individual user characteristics.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China (No. 62272330); in part by the Fundamental Research Funds for the Central Universities (No. YJ202219); in part by the Singapore Ministry of Education (MOE) Academic Research Fund (AcRF) Tier 1 grant (No. MSS24C004).

References

Vaibhav Adlakha, Shehzaad Dhuliawala, Kaheer Suleman, Harm de Vries, and Siva Reddy. 2022. Topiocqa: Open-domain conversational question answer-

ing with topic switching. *Transactions of the Association for Computational Linguistics*, 10:468–483.

Gediminas Adomavicius, Jesse C Bockstedt, Shawn P Curley, and Jingjing Zhang. 2013. Do recommender systems manipulate consumer preferences? a study of anchoring effects. *Information Systems Research*, 24(4):956–975.

Alaa Alslaity and Thomas Tran. 2019. Towards persuasive recommender systems. In *2019 IEEE 2nd international conference on information and computer technologies (ICICT)*, pages 143–148. IEEE.

Ian O Angell and Steve Smithson. 1990. Managing information technology: A crisis of confidence? *European management journal*, 8(1):27–36.

Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Inigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. 2018. Multiwoz–a large-scale multi-domain wizard-of-oz dataset for task-oriented dialogue modelling. *arXiv preprint arXiv:1810.00278*.

John T Cacioppo, Richard E Petty, Chuan Feng Kao, and Regina Rodriguez. 1986. Central and peripheral routes to persuasion: An individual difference perspective. *Journal of personality and social psychology*, 51(5):1032.

Carlos Carrasco-Farre. 2024. Large language models are as persuasive as humans, but why? about the cognitive effort and moral-emotional language of llm arguments. *arXiv preprint arXiv:2404.09329*.

Liang Chen, Yang Deng, Yatao Bian, Zeyu Qin, Bingzhe Wu, Tat-Seng Chua, and Kam-Fai Wong. 2023. Beyond factuality: A comprehensive evaluation of large language models as knowledge generators. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6325–6341.

Xiuying Chen, Mingzhe Li, Xin Gao, and Xiangliang Zhang. 2022. Towards improving faithfulness in abstractive summarization. *Advances in Neural Information Processing Systems*, 35:24516–24528.

Yue Chen, Chen Huang, Yang Deng, Wenqiang Lei, Dingnan Jin, Jia Liu, and Tat-Seng Chua. 2024. Style: Improving domain transferability of asking clarification questions in large language model powered conversational agents. *arXiv preprint arXiv:2405.12059*.

Zhongxia Chen, Xiting Wang, Xing Xie, Mehul Parsana, Akshay Soni, Xiang Ao, and Enhong Chen. 2020. Towards explainable conversational recommendation. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020*, pages 2994–3000. ijcai.org.

Robert B Cialdini. 2001. The science of persuasion. *Scientific American*, 284(2):76–81.

Robert B Cialdini and Noah J Goldstein. 2004. Social influence: Compliance and conformity. *Annu. Rev. Psychol.*, 55:591–621.

Gary Lynn Cronkhite. 1964. Logic, emotion, and the paradigm of persuasion. *Quarterly Journal of Speech*, 50(1):13–18.

Jeffrey Dalton, Chenyan Xiong, Vaibhav Kumar, and Jamie Callan. 2020. Cast-19: A dataset for conversational information seeking. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1985–1988.

Yang Deng, Yaliang Li, Bolin Ding, and Wai Lam. 2022a. Leveraging long short-term user preference in conversational recommendation via multi-agent reinforcement learning. *IEEE Transactions on Knowledge and Data Engineering*, 35(11):11541–11555.

Yang Deng, Yaliang Li, Fei Sun, Bolin Ding, and Wai Lam. 2021. Unified conversational recommendation policy learning via graph-based reinforcement learning. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1431–1441.

Yang Deng, Lizi Liao, Zhonghua Zheng, Grace Hui Yang, and Tat-Seng Chua. 2024. Towards human-centered proactive conversational agents. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 807–818.

Yang Deng, Wenxuan Zhang, Wai Lam, Hong Cheng, and Helen Meng. 2022b. User satisfaction estimation with sequential dialogue act modeling in goal-oriented conversational systems. In *Proceedings of the ACM Web Conference 2022*, pages 2998–3008.

Yang Deng, Wenxuan Zhang, Weiwen Xu, Wenqiang Lei, Tat-Seng Chua, and Wai Lam. 2023. A unified multi-task learning framework for multi-goal conversational recommender systems. *ACM Transactions on Information Systems*, 41(3):1–25.

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*.

Brian J Fogg. 2002. Persuasive technology: using computers to change what we think and do. *Ubiquity*, 2002(December):2.

Brian J Fogg and Hsiang Tseng. 1999. The elements of computer credibility. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*, pages 80–87.

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.

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

Sofia Gkika and George Lekakos. 2014. The persuasive role of explanations in recommender systems. In *BCSS@ PERSUASIVE*, pages 59–68.

Shuyu Guo, Shuo Zhang, Weiwei Sun, Pengjie Ren, Zhumin Chen, and Zhaochun Ren. 2023. Towards explainable conversational recommender systems. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2023, Taipei, Taiwan, July 23-27, 2023*, pages 2786–2795.

Kobi Hackenburg, Lujain Ibrahim, Ben M Tappin, and Manos Tsakiris. 2023. Comparing the persuasiveness of role-playing large language models and human experts on polarized us political issues.

Chen Huang, Peixin Qin, Yang Deng, Wenqiang Lei, Jiancheng Lv, and Tat-Seng Chua. 2024. Concept–an evaluation protocol on conversation recommender systems with system-and user-centric factors. *arXiv preprint arXiv:2404.03304*.

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

Ziwei Ji, Tiezheng Yu, Yan Xu, Nayeon Lee, Etsuko Ishii, and Pascale Fung. 2023. Towards mitigating llm hallucination via self reflection. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1827–1843.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

Felix N Koranteng, Uwe Matzat, Isaac Wiafe, and Jaap Ham. 2023. Credibility in persuasive systems: A systematic review. In *International Conference on Persuasive Technology*, pages 389–409. Springer.

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

Chuang Li, Yang Deng, Hengchang Hu, Min-Yen Kan, and Haizhou Li. 2024. Incorporating external knowledge and goal guidance for llm-based conversational recommender systems. *arXiv preprint arXiv:2405.01868*.

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.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023a. G-eval: Nlg evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522.

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

Hongyu Lu, Weizhi Ma, Yifan Wang, Min Zhang, Xiang Wang, Yiqun Liu, Tat-Seng Chua, and Shaoping Ma. 2023. User perception of recommendation explanation: Are your explanations what users need? *ACM Transactions on Information Systems*, 41(2):1–31.

Zheheng Luo, Qianqian Xie, and Sophia Ananiadou. 2023. Chatgpt as a factual inconsistency evaluator for abstractive text summarization. *arXiv preprint arXiv:2303.15621*.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. 2024. Self-refine: Iterative refinement with self-feedback. *Advances in Neural Information Processing Systems*, 36.

Mengqi Miao, Fandong Meng, Yijin Liu, Xiao-Hua Zhou, and Jie Zhou. 2021. Prevent the language model from being overconfident in neural machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3456–3468.

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.

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 pretraining. *arXiv preprint arXiv:2201.10005*.

Harri Oinas-Kukkonen and Marja Harjumaa. 2008. A systematic framework for designing and evaluating persuasive systems. In *Persuasive Technology: Third International Conference, PERSUASIVE 2008, Oulu,*

Finland, June 4-6, 2008. Proceedings 3, pages 164–176. Springer.

Daniel O'Keefe. 2016. Evidence-based advertising using persuasion principles: Predictive validity and proof of concept. *European Journal of Marketing*, 50(1/2):294–300.

Daniel J O'Keefe. 2002. Guilt as a mechanism of persuasion. *The persuasion handbook: Developments in theory and practice*, 329:344.

Alexander Pan, Kush Bhatia, and Jacob Steinhardt. 2021. The effects of reward misspecification: Mapping and mitigating misaligned models. In *International Conference on Learning Representations*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Richard M Perloff. 1993. *The dynamics of persuasion: Communication and attitudes in the 21st century*. Routledge.

Richard E Petty, Leandre R Fabrigar, and Duane T Wegener. 2003. Emotional factors in attitudes and persuasion. *Handbook of affective sciences*, 752:772.

Kathleen Kelley Reardon. 1991. *Persuasion in practice*. Sage.

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*, pages 888–896.

Peter Slattery, P Finnegan, and Richard Vidgen. 2020. Persuasion: an analysis and common frame of reference for is research. *Communications of the Association for Information Systems*, 46.

Lingzhi Wang, Huang Hu, Lei Sha, Can Xu, Daxin Jiang, and Kam-Fai Wong. 2022a. Recindial: A unified framework for conversational recommendation with pretrained language models. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing, AACL/IJCNLP 2022 - Volume 1: Long Papers, Online Only, November 20-23, 2022*, pages 489–500. Association for Computational Linguistics.

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

Xiaolei Wang, Xinyu Tang, Wayne Xin Zhao, Jingyuan Wang, and Ji-Rong Wen. 2023a. 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.

Xiaolei Wang, Kun Zhou, Xinyu Tang, Wayne Xin Zhao, Fan Pan, Zhao Cao, and Ji-Rong Wen. 2023b. 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. 2022c. 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.

Xuewei Wang, Weiyan Shi, Richard Kim, Yoojung Oh, Sijia Yang, Jingwen Zhang, and Zhou Yu. 2019. Persuasion for good: Towards a personalized persuasive dialogue system for social good. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5635–5649.

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. In *NeurIPS 2023 Workshop on Instruction Tuning and Instruction Following*.

Kyung-Hyan Yoo and Ulrike Gretzel. 2010. Creating more credible and persuasive recommender systems: The influence of source characteristics on recommender system evaluations. *Recommender systems handbook*, pages 455–477.

KyungHyan Yoo and Ulrike Gretzel. 2006. Measuring the credibility of recommender systems. In *Information and Communication Technologies in Tourism 2006*, pages 285–295. Springer.

Tian Yu, Izak Benbasat, and Ronald Cenfetelli. 2011. Toward deep understanding of persuasive product recommendation agents.

Yi Zeng, Hongpeng Lin, Jingwen Zhang, Diyi Yang, Ruoxi Jia, and Weiyan Shi. 2024. How johnny can persuade llms to jailbreak them: Rethinking persuasion to challenge ai safety by humanizing llms. *arXiv preprint arXiv:2401.06373*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Tong Zhang, Peixin Qin, Yang Deng, Chen Huang, Wenqiang Lei, Junhong Liu, Dingnan Jin, Hongru Liang, and Tat-Seng Chua. 2024. Clamber: A benchmark of identifying and clarifying ambiguous information needs in large language models. *Preprint*, arXiv:2405.12063.

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.

A Examples of Credibility-aware Persuasive Strategies

Credibility-aware Persuasive Strategies are formulated gaining insights from a broad scope of research fields, including psychology, marketing, natural language processing and so on. We first identify persuasive strategies that are commonly used in the scenario of recommendation, and then customize them with credible elements to construct the Credibility-aware Persuasive Strategies. To make readers better understand these strategies, we give an example for each strategy in Table 5.

B Details on User Simulator and Datasets

To reflect scenarios in real world application, we set different combinations of personas and preference attributes for our user simulator. Following Huang et al. (2024), we use the same 12 personas as they listed (namely, Anticipation, Boredom, Fusion, Curiosity, Delight, Disappointment, Exceptions, Frustration, Independence, Surprise, Trust, Satisfaction). As real users usually do not have certain target items when seeking help for recommender systems, we only set preference attributes for the simulator rather than specify target items. Specifically, we identify the 19 most common attribute groups in Redial and 16 in OpendialKG. Additionally, in conjunction with 12 pre-defined diverse user personas, the evaluation process generates 228 and 192 dialogues, respectively, for each dataset. During the conversation, we instruct the simulator to use its own words to describe preferences and accept items that exactly match its preference. The conversation will be terminated either reaching a maximum turns of 10 or the simulator accepts a recommendation.

C Additional Experiments

C.1 More Results on Ablation Study

We provide the results of ablation study on all metrics in Figure 8. It can be observed that both Strategy-guided Explanation Generation and Iterative Explanation Refinement are necessary in the design of PC-CRS. Without any process, the performance of PC-CRS drops on all metrics. It shows the effectiveness of cultivating the self-awareness of CRS and reinforcing the focus on factual information in generating both persuasive and credible explanations.

C.2 Additional Experiments using Llama3-8B-instruct

Motivations & Setups. Current LLM-based CRS methods mainly utilize ChatGPT as their backbones (Wang et al., 2023a; Fang et al., 2024). However, different models may vary in the performance of persuasiveness and credibility as they utilize different alignment mechanisms. To validate whether PC-CRS is generally applicable to various LLMs, we conduct experiments with the same setting as Section 4.1 except for implementing PC-CRS and other baselines with Llama3-8B-instruct.

Main Results. According to Table 6, PC-CRS with Llama3-8B-instruct achieves an average improvement of 14.77% in turn-level Credibility and 25.72% at the dialogue level (i.e., Convincing Acceptance) compared to the best baseline. It also demonstrates a 1.79% average increase in persuasiveness. Thus PC-CRS can consistently generate both persuasive and credible explanations with different LLMs.

In-depth Analysis. We also conduct an in-depth analysis with Llama3-8B-instruct. Figure 7 visually depicts the gap in metric scores by using average results on low-credibility explanations to minus high-credibility ones in InterCRS. It reveals that low-credibility explanations tend to cater to user's utterances rather than describing them faithfully, potentially leading to misleading explanations. Results in Table 7 and Table 8 shows that PC-CRS with Llama3-8B-instruct can improve its recommendation accuracy by providing credible explanations. In summary, our experiments with Llama3-8B-instruct suggest not only PC-CRS is applicable to different LLMs, but also the experimental findings are consistent with those obtained using ChatGPT.

D Details on Human Evaluation

As noted by previous work (Budzianowski et al., 2018; Dalton et al., 2020; Adlakha et al., 2022), human evaluation is labor-intensive. As a compromise, our human evaluation setup largely mirrors those used in prior studies (Wang et al., 2023a;

Route	Strategy Name	Example	Credible Information
Central Route	Logical Appeal	*Since you enjoy romantic dramas, I think you'll like Titanic. As a classic film in the* ***genre of romance****, it's likely to resonate with your viewing preferences.*	Genre
	Emotion Appeal	*Titanic tells the* ***heart-wrenching love story*** *of Jack and Rose, two young souls from different worlds who find each other on the ill-fated ship, only to be torn apart by class differences and a catastrophic event.*	Plot
	Framing	*You'll appreciate the* ***uplifting and emotional experience*** *that Titanic provides - a sweeping romance that will leave you feeling inspired and hopeful about the power of true love.*	Experience
Peripheral Route	Evidence-based Persuasion	*Directed by acclaimed James Cameron and starring Leonardo DiCaprio and Kate Winslet, Titanic is a cinematic masterpiece that has won* ***11 Academy Awards*** *and grossed over $2.1 billion at the box office.*	Awards
	Social Proof	*With an incredible* ***7.9/10*** *rating from over 1.3 million user reviews on IMDB,and a* ***88%*** *fresh rating on Rotten Tomatoes, it's clear that Titanic is a beloved classic that has captured the hearts of millions - don't miss out on this epic romance!*	Rate
Combination	Anchoring	*System: Did you know that Titanic won* ***11 Academy Awards*** *and grossed over $2.1 billion at the box office?* *User: Wow, that's impressive. I've heard great things about it.* *System: Yeah, and it's not just the box office success. The movie has an epic* ***romance****, stunning visual effects, and memorable performances from* ***Leonardo DiCaprio and Kate Winslet****. Plus, it's a classic romantic drama that has stood the test of time.*	Rate Genre Actor

Table 5: Examples of Credibility-aware Persuasive Strategies. Credible information in these examples are **bold**.

Models		Redial			OpendialKG		
		Persuasiveness	Credibility	Convincing Acceptance	Persuasiveness	Credibility	Convincing Acceptance
Llama-based	InterCRS	53.37	3.14	57.54	63.61	3.44	67.44
	ChatCRS	73.06	3.60	70.99	76.98	2.94	50.00
	MACRS	71.94	3.73	74.63	69.16	3.30	54.72
	PC-CRS (*ours*)	**74.81**	**4.04**	**93.46**	**77.89**	**4.17**	**85.11**
Improvement(%)		2.40↑	8.31↑	25.23↑	1.18↑	21.22↑	26.20↑

Table 6: Results in terms of persuasiveness and credibility with Llama3-8B-instruct.

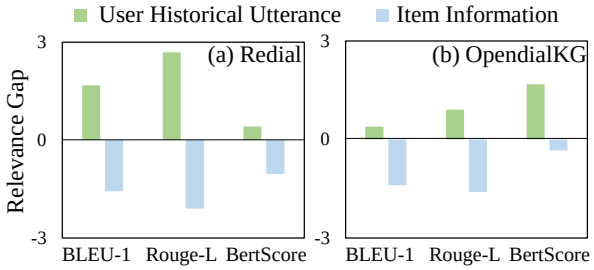

Figure 7: Results on relevance gap using Llama3-8B-instruct with InterCRS.

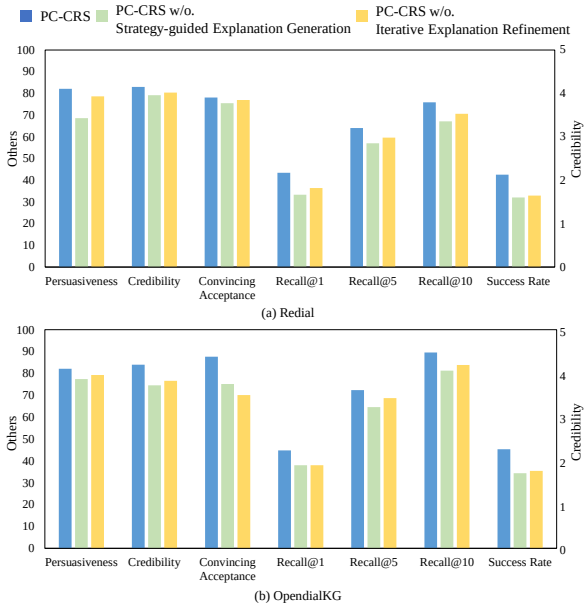

Figure 8: All results of ablation studies.

Huang et al., 2024). In the following, we validate the reliability of our evaluation given that we employ GPT-4 as our automatic evaluator and ChatGPT as our user simulator[7]. Then we verify the consistency of our evaluation results compared with human judgements. All these experiments are conducted with 3 human annotators in accordance with Wang et al. (2023a); Huang et al. (2024).

Reliability of our evaluator. To asses the reliability of our evaluator, we first randomly sample 50 dialogues with PC-CRS on Redial. Then, we instruct the human annotators to label the Watching Intention and Credibility of explanations in these dialogues with the same evaluation standard as LLM-based evaluator. The Krippendorff's alpha of annotators are 0.63 and 0.76 on Watching Intention and Credibility, respectively. We then take average of the annotated scores, and compute Spearman correlation between human annotators and LLM-based evaluators. The results are 0.59 and 0.62 on Watching Intention and Credibility, which are consistent with results in previous re-

[7]We do not consider using GPT-4 as a simulator due to its high cost.

Models		Redial				OpendialKG			
		R@1	R@5	R@10	SR	R@1	R@5	R@10	SR
Llama-based	InterCRS	24.56	42.11	60.53	10.09	29.69	59.38	74.48	31.77
	ChatCRS	17.11	37.28	61.84	12.72	36.98	**70.31**	78.65	33.85
	MACRS	17.98	41.23	56.58	15.79	34.90	68.75	79.75	32.81
	PC-CRS (*ours*)	**37.72**	**56.14**	**71.05**	**35.96**	**40.10**	66.15	**80.21**	**36.46**

Table 7: Results on recommendation accuracy with Llama3-8B-instruct.

Metrics		Item Information		User Historical Utterance	
		InterCRS	PC-CRS	InterCRS	PC-CRS
Redial	BLEU-1	5.17	**10.34**	10.11	**15.02**
	Rouge-L	7.01	**11.57**	16.88	**20.35**
	BertScore	78.86	**80.99**	84.67	**86.68**
OpendialKG	BLEU-1	5.95	**11.01**	10.18	**15.90**
	Rouge-L	8.71	**12.41**	16.14	**18.94**
	BertScore	79.94	**81.30**	83.96	**86.26**

Table 8: Explanation relevance to user historical utterance and item information using Llama3-8B-instruct.

search (Liu et al., 2023b,a), indicating a high reliablity of our automatic evaluator.

Reliability of our simulator. We further assess the *naturalness* and *usefulness* of the simulator through human evaluations following previous studies (Wang et al., 2023a; Sekulić et al., 2022). 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. We use the same 50 dialogues above and instruct the annotators to assign a score from 1 to 5 for each metric. 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.

Consistency with human evaluation. We verify whether our automatic evaluation results are consistent with human judgements. Specifically, we select 50 dialogues in each LLM-based baseline with the same profile and attribute group as PC-CRS and make annotators judge which dialogue is better (win) in terms of persuasiveness and credibility. Following the procedure in Sekulić et al. (2022), annotators are shown a pair of anonymous dialogues that are generated by PC-CRS and a baseline in the same user profile and attribute group. After making annotations independently, the annotators discuss together to resolve discrepancies. If they reach an agreement on a certain dialogue, it is labeled as *Win/Lose* accordingly, otherwise it is labeled as *Tie*. Results are provided in Figure 6. It can be observed that not only PC-CRS outperforms baselines but also the relative order is the same as results in Table 1. In summary, human studies confirm that our evaluation framework based on user simulators and evaluators is highly reliable.

Distinguishability of the our proposed strategies. We also conduct a human study to validate whether our proposed six credibility-aware persuasive strategies can be distinguished by humans. We sample 100 explanations generated by PC-CRS on Redial and invite two annotators to classify them according to the proposed strategy set. The results show the two annotators reach an agreement on 73% of the 100 samples, suggesting that our strategies have clear boundaries and are of high quality.

E Implementation Details & Prompts

E.1 Implementation Details

We conduct all our experiments using a single Nvidia RTX A6000, and we implement our codes in PyTorch. We take checkpoints and prompts from the corresponding code repositories and papers to implement all baselines. The maximum number of conversation turns between the user simulator and CRSs are set as 10. The maximum refinement iterations in PC-CRS is 2. In order to guarantee replicability, we have established fixed values for the Temperature and Seed parameters of ChatGPT (i.e., gpt-3.5-turbo-0125) and GPT-4 (i.e., gpt-4o-2024-05-13), setting both the Temperature and Seed to 0. Besides, to restrict the range of Persuasiveness in $[0, 1]$, we only calculate this metric when i_{post} is no greater than i_{true}.

E.2 Prompts

We outline the prompts used in PC-CRS (Table 9), User Simulator (Table 10) and LLM-based evaluator (Table 11). The strategy prompts used in Table 9 are an implementation of our Credibility-aware Persuasive Strategies. The descriptions in the prompts

are concrete applications of these strategies within the context of movie recommendations and adhere strictly to the definitions outlined in Section 3.1.1. By implementing these abstract social science concepts with concrete examples, PC-CRS can better eliminate hallucination and grasp the instructions.

Functions	Prompts
Strategy Selector	You are a recommender chatting with the user to provide recommendation. Now you need to select the two most suitable persuasive strategies from the candidate strategy to generate a persuasive response according to the conversation history. Candidate Strategy ######## Strategy Name: Logical Appeal Definition: Describe how the recommended movie's genre is consistent with the user's preference. Strategy Name: Emotion Appeal Definition: Sharing the plot and stories in the recommended movie to elicit user's emotions or support the recommendation. Strategy Name: Framing Definition: Emphasize the positive aspects, outcomes of watching the recommended movie based on the genre that matches user's preference. Strategy Name: Evidence-based Persuasion Definition: Using empirical data and facts such as movie directors and stars to support your recommendation. Strategy Name: Social Proof Definition: Highlighting what the majority believes in about the recommended movie by showing the movie rating and reviews by other users. Strategy Name: Anchoring Definition: Relying on the first piece of information as a reference point to gradually persuade the user, make sure all the information mentioned is truthful. ######## Conversation History=<HISTORY> Response with the following JSON format only: {"Thinking":<string>, "Strategy":<list>} Response with the JSON only!
Explanation Generator	You are a recommender chatting with the user to provide recommendation. Now you need to generate a persuasive response based on the conversation history , persuasive strategy and item information below. Conversation History=<HISTORY> Persuasive Strategy=<SELECTED_STRATEGY> Item Information=<ITEM_INFORMATION> Make sure your response is strictly consistent with the given information, your response should honestly reflecting the given information and do not contain any other misinformation. Be brief in your response! Response:
Critique	You are an evaluator and you need to judge the credibility of the recommender's utterance based on the given source information. Note credible means every claim in the recommender utterance is supported by source information or some minor details can be logically inferred from source information. Recommender Utterance=<CANDIDATE_EXPLANATION> Source Information=<ITEM_INFORMATION> First summarize the information in the recommender' utterance and compare it with the source information to judge its credibility, then give your judgement on whether the recommender utterance is credible. Output your reasoning process in the "Evidence". Output "True" or "False" in "Credibility". Response in the following JSON format: {"Evidence": <string>, "Credibility": <string>} Response the JSON only!
Refine	You are a recommender chatting with the user to provide recommendation. Given the source information and other's critique, there is misinformation in your current response. Remove the misinformation based on the critique and make sure your response is strictly consistent with the given information and every statement is well-supported. Remember to use the following persuasive strategy below and do not contain any misinformation in your new response. Be brief in your response. Reply with your new response only! Source Information=<ITEM_INFORMATION> Current Response=<CANDIDATE_EXPLANATION> Critique=<CRITIQUE> Persuasive Strategy=<SELECTED_STRATEGY> New Response:
Recommendation	You are a recommender chatting with the user to provide recommendation. You must follow the instructions below during chat. 1. If you do not have enough information about user preference, you should ask the user for his preference. 2. If you have enough information about user preference, you can give recommendation. If you decide to give recommendation, you should choose 1 item to recommend from the candidate list. 3. If you decide to select a movie and recommend, add a special token '[REC]' at the end of your response. 4. If you are making explanations on your recommendation, add a special token '[EXP]' at the end of your response. 5. Make sure your response is consistent with the given information, your response should honestly reflecting the given information and do not contain any deception. 6. Be brief in your response! Candidate List=<ITEM_LIST> Conversation History=<HISTORY> Your Response:

Table 9: Prompts used in PC-CRS.

Functions	Prompts
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 10: Prompts used in User Simulator.

Functions	Prompts
Watching Intention	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>. Now you need to score your watching intention based on the criteria and recommender's utterance below: Watching Intention Criteria ###### 1. Not Interested (Score 1): No alignment with preferred genres. Uninteresting plot and weak synopsis. No favorite actors or directors involved. Poor critical acclaim. Lack of personal recommendations. Inaccessible or expensive. Doesn't suit current mood or timing. Lacks originality or innovation. 2. Slightly Interested (Score 2): Some alignment with preferred genres, but not perfect. Plot seems somewhat engaging, but not highly captivating. Some familiar faces among the cast and crew. Mixed or average critical acclaim. Few personal recommendations or not strong ones. Available but may require some effort or cost. Somewhat suits current mood or timing. Some degree of originality or innovation. 3. Moderately Interested (Score 3): Fairly good alignment with preferred genres. Intriguing plot with potential. Few favorite actors or directors involved. Generally positive critical acclaim. Some personal recommendations from trusted sources. Easily accessible and reasonably priced. Fits current mood or timing quite well. Shows some originality or innovation. 4. Very Interested (Score 4): Strong alignment with preferred genres. Highly engaging plot with positive reception. Many favorite actors or directors involved. High critical acclaim or praise. Several strong personal recommendations. Easily accessible and included in subscription. Perfectly fits current mood or timing. Displays notable originality or innovation. 5. Extremely Interested (Score 5): Perfect alignment with preferred genres. Extremely captivating plot with widespread acclaim. All or most favorite actors or directors involved. Exceptional critical acclaim or awards. Numerous enthusiastic personal recommendations. Readily accessible and already in streaming library. Perfectly suits current mood or timing. Demonstrates groundbreaking originality or innovation. ###### Recommender Utterance=<RECOMMENER_EXPLANATION> Pretend you have little knowledge about the recommended movies, and the only information source about the movie is the recommender. You can only consider your watching intention based on the information given in the recommender's utterance. First summarize the movie information from the recommender utterance and consider how it matches the scoring criteria, then score your watching intention. Output your reasons to the score in the "Evidence". Response in the following JSON format: {"Evidence": <string>, "Watching Intention": <int>}
Credibility	You are an evaluator and you need to judge the credibility of the recommender's utterance based on the given source information. Rate your score based on the Evaluation Standard. Evaluation Standard ####### Score 5: Exact Match: The sentence exactly matches the information provided and every claim in the sentence is well-supported by the source information. Score 4: Supported Inference: Most of the claims in the sentence is supported by the source information with minor details can be logically inferred from the information provided. Score 3: Partial Match: The sentence partially matches the information provided, with some details supported and others not mentioned. Score 2: Inconsistent Detail: The sentence contains much information that are not supported by the information provided. Score 1: Unsubstantiated Claim: The sentence makes a claim that is contradict to the information provided. ####### Recommender Utterance=<RECOMMENDER_EXPLANATION> Source Information=<ITEM_INFORMATION> First summarize the information in the recommender' utterance and compare it with the source information to judge its credibility, then give your integer score. Output your reasoning process in the "Evidence". Output your score in the "Credibility". Response in the following JSON format: {"Evidence": <string>, "Credibility": <int>}

Table 11: Prompts used in LLM-based evaluators.