Towards empathetic conversational recommender systems

Zhang, X.; Xie, R.; Lyu, Y.; Xin, X.; Ren, P.; Liang, M.; ... ; Joachims, T.

Towards Empathetic Conversational Recommender Systems

Xiaoyu Zhang*
Shandong University
tinyoctopus1999@gmail.com

Ruobing Xie
Tencent
xrbsnowing@163.com

Yougang Lyu
Shandong University
University of Amsterdam
youganglyu@gmail.com

Xin Xin
Shandong University
xinxin@sdu.edu.cn

Pengjie Ren
Shandong University
jay.ren@outlook.com

Mingfei Liang
Tencent
aesopliang@tencent.com

Bo Zhang
Tencent
nevinzhang@tencent.com

Zhanhui Kang
Tencent
kegokang@tencent.com

Maarten de Rijke
University of Amsterdam
m.derijke@uva.nl

Zhaochun Ren†
Leiden University
z.ren@liacs.leidenuniv.nl

ABSTRACT

Conversational recommender systems (CRSs) are able to elicit user preferences through multi-turn dialogues. They typically incorporate external knowledge and pre-trained language models to capture the dialogue context. Most CRS approaches, trained on benchmark datasets, assume that the standard items and responses in these benchmarks are optimal. However, they overlook that users may express negative emotions with the standard items and may not feel emotionally engaged by the standard responses. This issue leads to a tendency to replicate the logic of recommenders in the dataset instead of aligning with user needs. To remedy this misalignment, we introduce *empathy* within a CRS. With empathy we refer to a system's ability to capture and express emotions. We propose an **e**mpathetic **c**onversational **r**ecommender (ECR) framework.

ECR contains two main modules: emotion-aware item recommendation and emotion-aligned response generation. Specifically, we employ user emotions to refine user preference modeling for accurate recommendations. To generate human-like emotional responses, ECR applies retrieval-augmented prompts to fine-tune a pre-trained language model aligning with emotions and mitigating hallucination. To address the challenge of insufficient supervision labels, we enlarge our empathetic data using emotion labels annotated by large language models and emotional reviews collected from external resources. We propose novel evaluation metrics to capture user satisfaction in real-world CRS scenarios. Our experiments on the ReDial dataset validate the efficacy of our framework in enhancing recommendation accuracy and improving user satisfaction.

*Work done during an internship at 2023 Tencent Rhino-Bird Research Elite Program.
†Corresponding author.

RecSys '24, October 14–18, 2024, Bari, Italy

ACM ISBN 979-8-4007-0505-2/24/10
https://doi.org/10.1145/3640457.3688133

CCS CONCEPTS

• **Information systems** → **Recommender system**; **Users and interactive retrieval**.

KEYWORDS

Conversational recommender system, Empathetic response generation, User preference modeling, Prompt engineering

ACM Reference Format:
Xiaoyu Zhang, Ruobing Xie, Yougang Lyu, Xin Xin, Pengjie Ren, Mingfei Liang, Bo Zhang, Zhanhui Kang, Maarten de Rijke, and Zhaochun Ren. 2024. Towards Empathetic Conversational Recommender Systems. In *18th ACM Conference on Recommender Systems (RecSys '24), October 14–18, 2024, Bari, Italy*. ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3640457.3688133

1 INTRODUCTION

Advances in conversational systems have led to the integration of natural language conversations with recommender systems, culminating in the development of conversational recommender systems (CRSs) [23]. A crucial aspect of CRSs is to elicit user preferences through multi-turn dialogues, with two main subtasks: *item recommendation* and *response generation* [9]. A prominent challenge is the lack of sufficient contextual information for accurately modeling user preferences. Some research [2, 41] integrates knowledge graphs (KGs) and models user preferences based on entities from KGs mentioned in the dialogues. Recent work [8, 34, 35] centers on using pre-trained language models (PLMs) to enhance the system's understanding of dialogue context. Despite these advances, existing CRS models still do not fully align with user needs. These models are trained on conversational recommendation training datasets. But the presumption that the standard items and responses in the dataset are optimal leads to a tendency of CRS to replicate the logic of recommenders in the dataset instead of addressing user needs.

Using empathy to address misalignment. The above misalignment hinders the development of CRSs. Lerner et al. [20] have proposed that *emotions* are crucial in human decision-making processes.

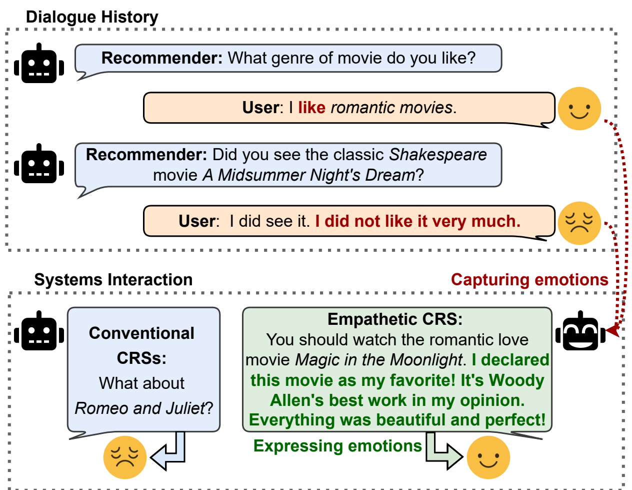

Figure 1: An example of a conversation on movie recommendation between a user and the system. Text conveying user emotions is highlighted in red font. System responses expressing emotions are marked in green font.

Their work suggests that capturing emotions expressed in user utterances within dialogues is prominent for achieving accurate user preference modeling for item recommendation. People tend to favor agents simulating human beings by exhibiting *emotions* [4]. Adopting emotion-rich expressions in response generation can enliven the user experience and contribute to user satisfaction. Serving users in a natural, human-like way by capturing and expressing emotions is a necessary development for CRSs in terms of aligning with user needs, thereby offering benefits for users and providers of recommender systems. We introduce *empathy* within a CRS, defining it as the system's capacity to capture and express emotions [16]. Through empathy, we aim to accurately distinguish and fulfill user needs, both during item recommendation and response generation.

Integrating empathy into recommendation and generation. We analyze the need to integrate empathy into item recommendation and response generation subtasks, respectively. For item recommendation, existing approaches often assume that all entities mentioned in dialogues reflect user preferences and that all items suggested by recommenders meet user expectations. This hypothesis disregards subtle cues of user emotions expressed in natural language for modeling user preferences. As illustrated in Figure 1, a conventional CRS might infer that the user likes "Shakespeare" and "A Midsummer Night's Dream" mentioned by the recommender, while overlooking that the user expresses negative emotions towards them during the dialogue. Thus such systems recommend the wrong item "Romeo and Juliet." For response generation, existing methods are trained on the standard responses from datasets, which tend to be short and lack narratives, often resulting in inconsistencies or a lack of emotional engagement. As shown in Figure 1 (bottom), the conventional CRSs' response only contains the item name, which may diminish user satisfaction when interacting with the system. In contrast, based on capturing and expressing emotions, an empathetic CRS recommends a reasonable item with a persuasive response.

Challenges. To construct empathetic CRSs, we face two major challenges: (i) how to accurately model user preferences using emotions; and (ii) how to generate emotional responses contributing to user satisfaction. To address these challenges, we propose an **e**mpathetic **c**onversational **r**ecommender (ECR) framework comprising two key modules: *emotion-aware item recommendation* and *emotion-aligned response generation*. For the emotion-aware item recommendation module, we integrate user emotions with entities in the utterance to augment user preference modeling. We also propose a training strategy to minimize the impact of incorrect labels in the dataset. For the emotion-aligned response generation module, we fine-tune a pre-trained language model (PLM) to express emotions. To avoid hallucination, we retrieve relevant knowledge from KGs as a part of generation prompts. Existing CRS datasets lack user emotion labels and emotional responses. To enlarge the available empathetic training data, we use large language models (LLMs) to discern nuanced emotions in the dialogue history; then, we collect emotional reviews as an informative external resource for fine-tuning the PLM to generate emotional responses.

Since existing evaluation metrics ignore the impact of emotions, we introduce novel metrics for CRSs, aiming at better reflecting user satisfaction in real-world CRS scenarios. For item recommendation, we adopt the Area Under the Curve (AUC) metric to assess the model's accuracy in modeling user preferences. AUC requires that items receiving positive feedback from users should have a higher possibility of being recommended than those with negative feedback. For response generation, we move beyond traditional metrics like BLEU or ROUGE, opting instead to use five subjective metrics: emotional intensity, emotional persuasiveness, logic persuasiveness, informativeness, and lifelikeness. Experiments on the ReDial benchmark dataset confirm the effectiveness of our proposed framework.

Contributions. The contributions of this paper are as follows: (i) To bridge the gap between system outputs and user needs, we define empathy within a CRS and propose a novel framework ECR. (ii) We augment user preference modeling by integrating their emotions, with a new training strategy to minimize the impact of incorrect labels. (iii) We fine-tune a PLM to express emotions and apply retrieval-augmented prompts to mitigate hallucination. (iv) We use LLMs to annotate user emotions and collect emotional reviews from external resources as empathetic CRS training data, which facilitates future research in this area. (v) We propose new evaluation metrics tailored to user satisfaction in real-world CRS scenarios, and our experimental results demonstrate that ECR significantly outperforms baselines on the ReDial dataset.

2 RELATED WORKS

The literature on CRSs [23, 34, 41] can be classified into *attribute-based* CRSs and *generation-based* CRS [37]. Attribute-based CRSs [3, 42] predominantly employ fixed response templates and predefined actions for user interaction. The primary objective of most methodologies within this category is to minimize the number of turns required to complete the recommendation task [18, 19]. Deng et al. [7] and Lei et al. [19] use KGs to improve the recommendation performance. However, they still overlook the importance of generating high-quality natural language, which can be detrimental to the overall user experience.

Unlike attribute-based CRSs, generation-based CRSs [23, 43] focus on making recommendations using free-form text, which creates considerable flexibility to influence how a dialogue continues. Li et al. [23] use an auto-encoder for recommendation and a hierarchical RNN for response generation. However, a challenge these systems face is the lack of sufficient contextual information for accurately discerning user preferences [10]. Research indicates that CRSs can be enhanced by incorporating additional sources of knowledge. Chen et al. [2] integrates KGs to enhance the user representation and propose an end-to-end framework. Zhou et al. [41] incorporate both word-oriented and entity-oriented KGs. Through reasoning based on the entities from KGs mentioned in the dialogues [25, 28, 40, 44], this integration further enhances the logical accuracy of recommendation and response interpretability. Subsequent research also introduces reviews [27, 44] and in-text knowledge [24, 29, 38] to assist user preference modeling.

Recent work on generation-based CRSs have centered on integrating LLMs into CRSs [8, 34, 35]. UniCRS [34] addresses the recommendation and generation subtasks in a unified approach with prompt tuning. He et al. [10] conduct an in-depth analysis of LLMs for zero-shot CRS. And Wang et al. [33] develop an interactive evaluation method using LLM-based user simulators. However, Dai et al. [5] demonstrate that traditional collaborative filtering recommendation models, with adequate data, significantly outperform LLMs. Moreover, while LLMs are proficient in conversational aspects, they face limitations in conversational recommendation tasks, particularly in capturing user emotional engagement.

Our study aligns with the generation-based CRSs. A major problem of recent generation-based CRSs is their misalignment with user preferences. We integrate empathy into CRS, prioritizing user needs as our goal. Similarly to our approach, methods for empathetic response generation [21, 22, 36] detect and respond to user emotions. These methods are tailored for chat and not easily adapted to CRSs. Some traditional recommender systems have enhanced collaborative filtering by incorporating sentiment analysis [6, 13, 14]. However, these works only focus on the analysis of item reviews rather than real-time multi-turn natural language dialogues.

3 PRELIMINARIES

3.1 Problem Formulation

Notation. Given $t-1$ dialogue turns, the dialogue history D_{t-1} consists of a sequence of utterances from both recommenders and users, i.e., $D_{t-1} = \{u_k^r, u_k^u\}_{k=1}^{t-1}$, where each utterance $u_k^* = \{w_j\}_{j=1}^{|u_k^*|}$ is composed of a sequence of words. For simplicity, we concatenate all utterances from D_{t-1} into a single word sequence $D = \{w_q\}_{q=1}^{n_w}$, where n_w represents the total number of words in D_{t-1}. To incorporate knowledge about entities mentioned in the dialogue, we set an external knowledge graph (e.g., DBpedia [1]) as $\mathcal{G} = (\mathcal{E}, \mathcal{L})$, consisting of triples $\mathcal{T} = \langle e_h, l, e_t \rangle$, where $e_h \in \mathcal{E}$ and $e_t \in \mathcal{E}$ are the head and tail entities, $l \in \mathcal{L}$ reflects the relation between e_h and e_t. $\mathcal{E}$ and $\mathcal{L}$ denote the sets of entities and relations. We define I as the entire set of items, all of which are included in the entities of $\mathcal{G}$, i.e., $I \in \mathcal{E}$. Entities in each utterance u_k^* are identified as $E_k^* = \{e_j\}_{j=1}^{|E_k^*|}$. Each item i_j within E_k^* is linked with user feedback f_{i_j}, indicating whether the user likes it. Similarly, we combined all entities mentioned in D_{t-1} into an entity list $E_l = \{e_q\}_{q=1}^{n_e}$, where n_e is the count of entities in the dialogue history. Here, we refer to the entities mentioned in the dialogue history as *local entities*. Correspondingly, we refer to entities co-occurring with the local entities in the training dataset as *global entities*, which will be detailed in Section 4.2.1.

Task outputs. At the t-th turn, a CRS (i) selects a set of target items $I_t = \{i_k\}_{k=1}^{|I_t|}$ from the entire item set I, and (ii) generates a response utterance u_t^r for the user.

3.2 Backbone Framework

Since UniCRS [34] unifies the recommendation and generation into a prompt learning paradigm by PLM, i.e., DialoGPT [39], which is the state-of-the-art method in using PLMs, we adopt it as our backbone framework. It encompasses three primary modules:

(1) **Semantic fusion module**: Initially, UniCRS fuses the semantic spaces of dialogues and KGs for knowledge alignment and enrichment. It obtains a word embedding matrix and a local entity embedding matrix. Then it associates two kinds of embedding matrices via a bilinear transformation, yielding the fused word representations $\tilde{\boldsymbol{W}} = [\tilde{\boldsymbol{w}}_1; \ldots; \tilde{\boldsymbol{w}}_{\boldsymbol{n_w}}]$, and the fused local entity representations $\tilde{\boldsymbol{E}}_{\boldsymbol{l}} = [\tilde{\boldsymbol{e}}_1; \ldots; \tilde{\boldsymbol{e}}_{\boldsymbol{n_e}}]$.

(2) **Response generation module**: UniCRS prompts a PLM to generate the response u_t^r, which is designated as "*recommendation response*." The prompt for this module consists of the fused word representations $\tilde{\boldsymbol{W}}$, generation task-specific soft tokens $\boldsymbol{S_{gen}}$, and the dialogue history D:

$$C_{gen}^r = [\tilde{\boldsymbol{W}}; \boldsymbol{S_{gen}}; D]. \tag{1}$$

Note that UniCRS replaces all items appearing in the recommendation response with a special token [MASK], which is later filled following the item recommendation subtask.

(3) **Item recommendation module**: Given u_t^r from the response generation subtask, the recommendation prompt consists of the fused local entity representations $\tilde{\boldsymbol{E}}_{\boldsymbol{l}}$, recommendation task-specific soft tokens $\boldsymbol{S_{rec}}$, the dialogue history D, and u_t^r:

$$C_{rec} = [\tilde{\boldsymbol{E}}_l; \boldsymbol{S_{rec}}; D; u_t^r]. \tag{2}$$

The response generation and item recommendation modules both use cross-entropy loss for prediction.

Although UniCRS shows promise in using PLMs, its optimization still relies on standard answers provided by datasets and ignores user emotions, which limits its ability to track user needs. It inspires our subsequent endeavors in instantiating ECR based on UniCRS. Note that our proposed framework can extend beyond UniCRS and be seamlessly adapted to other CRSs with modifications.

4 METHOD

In this section, we introduce our empathetic data enlargement process (Section 4.1) and two key modules of ECR: emotion-aware item recommendation (Section 4.2) and emotion-aligned response generation (Section 4.3). Figure 2 shows an overview of ECR.

4.1 Empathetic Data Enlargement

4.1.1 User Emotion Extraction. Existing datasets lack explicit supervisory signals for identifying user emotions. To address the problem,

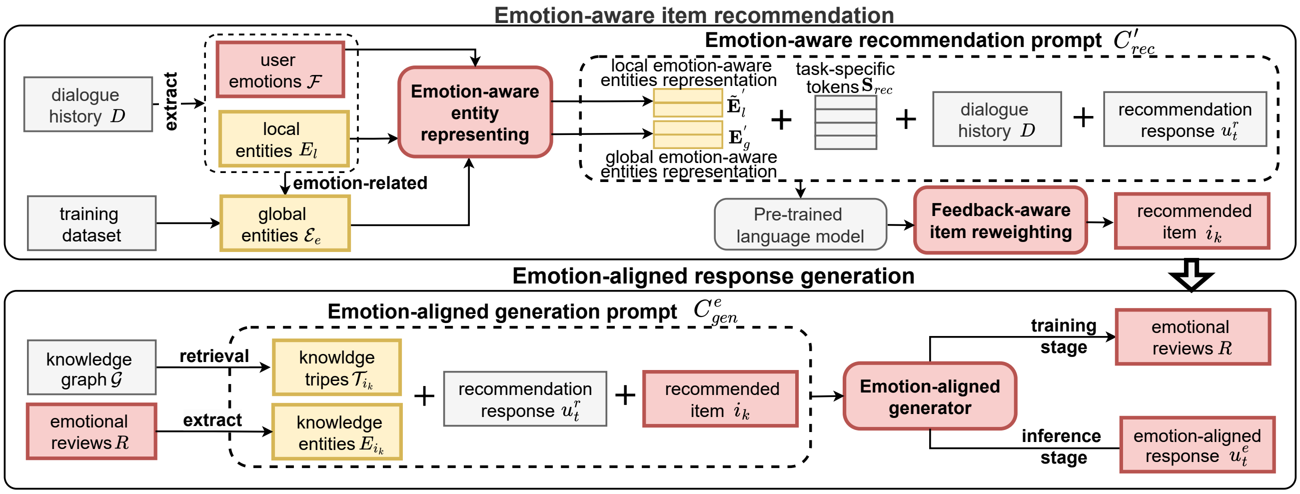

Figure 2: Overview of ECR. ECR has two key modules: (i) emotion-aware item recommendation for better user preference understanding, and (ii) emotion-aligned response generation for engaging conversations.

we employ GPT-3.5-turbo [45] to initially annotate user emotions in 5,082 utterances from the ReDial dataset. We limit the number of annotated emotions per utterance to a maximum of two labels. In annotating with GPT-3.5-turbo [45] for utterance-level user emotions, we adopted nine emotion labels: "like," "curious," "happy," "grateful," "negative," "neutral," "nostalgia," "agreement," and "surprise." The "negative" label, denoting adverse emotions, accounted for 8.0%. See Appendix A.1 for a more detailed annotation process. Based on the annotations, we fine-tune a GPT-2 model, which achieves 87.75% in terms of Recall@2 in categorizing emotions. We applied this model to annotate emotions for each user utterance u^u_k in the ReDial dataset and set a threshold β to retain relevant emotion label f. For each utterance u^u_k, we obtain its utterance-level user emotions $\mathcal{F}_{u^u_k} = \{f_j\}_{j=1}^{|\mathcal{F}_{u^u_k}|}$, along with the probabilities associated with each emotional label, denoted as $\mathcal{P}_{u^u_k} = \{p_j\}_{j=1}^{|\mathcal{P}_{u^u_k}|}$, where $|\mathcal{F}_{u^u_k}| = |\mathcal{P}_{u^u_k}|$.

4.1.2 Emotional Response Construction. There is an abundance of reviews about consumer experiences on the web. These reviews are frequently imbued with the writers' personal experiences and emotions, with comprehensive information about various facets of the items being reviewed. Hence, emotional reviews serve as optimal resources for emotional responses. Considering the emphasis of our study is on recommendation task and positive emotions contribute more to effectively persuasive and engaging interactions [15], we only adopted top-rated (10/10) reviews rich in positive emotions, rather than expressing different types of emotions according to user emotions. To construct a emotional review database R, we collect movie reviews from IMDb.[1] Each emotional review is retrieved according to the item i_r in the training dataset. The constructed emotional review database $R = \{r_j\}_{j=1}^{|R|}$ comprises a sequence of emotional review sentences. Each emotional review sentence $r = \{w_j\}_{j=1}^{|r|}$ consists of a sequence of words.

[1] https://www.imdb.com/

From each emotional review sentence r, we extract a list of entities $E_r = \{e_j\}_{j=1}^{|E_r|}, e_j \in \mathcal{E}$. We then retrieve a set of knowledge triples $\mathcal{T}_r = \{\langle i_r, l, e_j\rangle, e_j \in E_r\}_{j=1}^{|\mathcal{T}_r|}$ from the knowledge graph $\mathcal{G}$, where the item i_r is the head entity, the entity $e_j \in E_r$ is the tail entity.

4.2 Emotion-aware Item Recommendation

4.2.1 Emotion-aware Entity Representation. Entities are essential for reflecting user preferences. Thus we aim to model the effect of user emotions on the entities. Since local entities E_l only reflect user interest exhibited in the ongoing dialogue D_{t-1}, which is insufficient for comprehensively exploring user preferences, we collect global entities from the training dataset filtered by user emotions, which encompass the collaborative knowledge shared by all users. In general, we model the effect of user emotions on entities both in the dialogue history and in the training data, respectively.

Local emotion-aware entity representing. To model the effect of user emotions on local entities, we characterize utterance-level user emotions $\mathcal{F}_{u^u_k} = \{f_j\}_{j=1}^{|\mathcal{F}_{u^u_k}|}$ as reflecting emotions towards entities mentioned by the user in the current utterance E^u_k and by the recommender in the preceding one E^r_k. Consequently, each local entity $e_j \in E_l$ is linked to an utterance-level user emotions, represented as $\mathcal{F}_{e_j} = \{f_i\}_{i=1}^{|\mathcal{F}_{e_j}|}$ along with corresponding probabilities $\mathcal{P}_{e_j} = \{p_i\}_{i=1}^{|\mathcal{P}_{e_j}|}$. Hence, we calculate the user emotion representation of each local entity e_j as:

$$\boldsymbol{\mathcal{F}}_{\boldsymbol{e_j}} = \sum_{i=1}^{|\mathcal{F}_{e_j}|} p_i * v(f_i), \tag{3}$$

where $v(f_i)$ denotes the learnable representation of the i-th emotion label in the utterance-level user emotions. Then, we fuse the user emotion representation $\boldsymbol{\mathcal{F}}_{\boldsymbol{e_j}}$ with the local entity representation of e_j to get a local emotion-aware entity representation $\tilde{\boldsymbol{e}}'_j$, as follows:

$$\tilde{\boldsymbol{e}}'_j = \left[\tilde{\boldsymbol{e}}_j; \boldsymbol{\mathcal{F}}_{\boldsymbol{e_j}}\right] \mathbf{W}_t^\top + \boldsymbol{b}, \tag{4}$$

where $[\cdot;\cdot]$ denotes vector concatenation; $\mathbf{W_t}$ and $\boldsymbol{b}$ are learnable parameters aimed at projecting the dimension of the concatenated representation back to the dimension of $\tilde{\boldsymbol{e}}_{\boldsymbol{j}}$. We stack all local emotion-aware entity representations into a matrix, denoted as $\tilde{\boldsymbol{E}}_{\boldsymbol{l}}' = \left[\tilde{\boldsymbol{e}}_1'; \ldots; \tilde{\boldsymbol{e}}_{\boldsymbol{n_e}}'\right]$.

Global emotion-aware entity representing. We first use utterance-level user emotions to filter global entities and then aggregate their representations. Concretely, we assume that if a user exhibits similar emotions towards both e_j and e_i in a conversation, then e_i is globally *emotion-related* to e_j. Similarly, we define a local entity e_j to be emotion-related to a set of global entities $\mathcal{E}_{e_j} = \{e_i\}_{i=1}^{|\mathcal{E}_{e_j}|}$, where e_j and e_i overlap in the most probable n_f emotion labels in $\mathcal{F}_{e_j}$ and $\mathcal{F}_{e_i}$ during their co-occurrence in a conversation. Additionally, we calculate the co-occurrence probability of each local entity e_j with its emotion-related global entity e_i from the training dataset, denoted as $P(e_i|e_j)$. Then, we aggregate the representation of all global entities emotion-related to the local entity e_j as a global entity representation $\boldsymbol{\mathcal{E}}_{\boldsymbol{e_j}}$:

$$\boldsymbol{\mathcal{E}}_{\boldsymbol{e_j}} = \sum_{i=1}^{|\mathcal{E}_{e_j}|} \boldsymbol{e_i} * P(e_i|e_j), \tag{5}$$

where $\boldsymbol{e_i}$ denotes the representation of e_i obtained from the RGCN [30]. Following the Eq. 4, we calculate the global emotion-aware entity representation $\boldsymbol{\mathcal{E}}_{\boldsymbol{e_j}}'$ by integrating $\boldsymbol{\mathcal{E}}_{\boldsymbol{e_j}}$ and $\boldsymbol{\mathcal{F}}_{\boldsymbol{e_j}}$. Finally, we stack global emotion-aware entity representation for each local entity $e_j \in E_l$ into a matrix, denoted as $\boldsymbol{E}_{\boldsymbol{g}}' = \left[\boldsymbol{\mathcal{E}}_{\boldsymbol{e_1}}'; \ldots; \boldsymbol{\mathcal{E}}_{\boldsymbol{e_{n_e}}}'\right]$.

Emotion-aware recommendation prompt. To comprehensively model user preferences with their emotions, we use the local emotion-aware entity representation matrix $\tilde{\boldsymbol{E}}_{\boldsymbol{l}}'$ and global emotion-aware entity representation matrix $\boldsymbol{E}_{\boldsymbol{g}}'$ to update the prompt in Eq. 2. So we formulate an emotion-aware recommendation prompt as:

$$C_{rec}' = \left[\tilde{\boldsymbol{E}}_{\boldsymbol{l}}'; \boldsymbol{E}_{\boldsymbol{g}}'; \boldsymbol{S_{rec}}; D; u_t^r\right]. \tag{6}$$

4.2.2 Feedback-aware Item Reweighting. In the preceding section, we employ utterance-level user emotions to track user entity-based preferences. In this section, we develop a reweighting strategy that is aware of user feedback f_{i_k} on each recommended item i_k, aligning with the supervision labels provided in the dataset but ignored by most CRS methods. Specifically, we introduce a mapping function $m(f_{i_k})$ that converts each user feedback f_{i_k} as a weight scalar. The mapping function converts negative or unclear feedback into a lower weight. Based on the weight scalars, we rewrite the cross-entropy loss for item recommendation subtask as:

$$L_{rec} = -\sum_{k=i}^{N} m(f_{i_k}) \log Pr(i_k \mid C_{rec}'), \tag{7}$$

where N represents the total number of training instances, and $Pr(i_k \mid C_{rec}')$ refers to the predicted probability of the recommended item i_k given the emotion-aware recommendation prompt C_{rec}'.

4.3 Emotion-aligned Response Generation

To support an engaging user experience, we generate an *emotion-aligned response* u_t^e to enrich the recommendation response u_t^r generated by the UniCRS. In this section, we construct an emotion-aligned generation prompt and train an emotion-aligned generator.

4.3.1 Emotion-aligned Generation Prompt. While PLMs can memorize information from their training corpus, Ji et al. [12] have shown that PLMs often exhibit hallucinations, which may diminish users' satisfaction with their usage. Teaching PLMs to accurately retain knowledge is resource-intensive and challenging. Thus, we construct an emotion-aligned generation prompt based on retrieved knowledge to enhance the informativeness in the responses while mitigating hallucination.

Specifically, during the training stage, given the extracted knowledge entities E_r and the retrieved knowledge triples $\mathcal{T}_r$, we transform the entities and triples into word sequences, represented as $S_{\mathcal{T}_r}$ and S_{E_r}. The prompt for generating emotional review r consists of the word sequence of the knowledge entities S_{E_r}, knowledge triples $S_{\mathcal{T}_r}$, and the item name S_{i_r}. Then, we incorporate the recommendation response u_t^r into the prompt, guiding the model to generate contextually relevant responses. The emotion-aligned generation prompt is formally denoted as:

$$C_{gen}^e = \left[S_{E_r}; S_{\mathcal{T}}; S_{i_r}; u_t^r\right]. \tag{8}$$

During the inference stage for generating emotion-aligned responses u_t^e, we followed the same prompt design as in Eq. 8. We retrieve knowledge triples $\mathcal{T}_{i_k} = \{\langle i_k, l, e_j\rangle\}_{j=1}^{|\mathcal{T}_{i_k}|}$ from the KG $\mathcal{G}$ using the predicted item i_k as the head entity. And we collect a list of knowledge entities $E_{i_k} = \{e_j\}_{j=1}^{|E_{i_k}|}$ that is mentioned at least twice in the reviews corresponding to i_k. Then, we filter pn_t triples from $\mathcal{T}_{i_k}$ and pn_e entities from $E_{i_k}^r$ as a part of the emotion-aligned generation prompt. We simplify the filtering process by random selection, leaving more complex approaches to be explored in the future.

4.3.2 Emotion-aligned Generator. To align the model with the persuasive dialogue style and emotions, we fine-tune a PLM as an emotion-aligned generator to generate the emotion-aligned responses u_t^e using the constructed emotional review databases R. Specifically, based on the emotional reviews r, we employ cross-entropy for training the emotion-aligned generator, as follows:

$$L_{gen}^e(r) = -\sum_{j=1}^{|r|} \log \Pr(w_j \mid C_{gen}^e; w_{<j}), \tag{9}$$

where $\Pr(w_j \mid C_{gen}^e; w_{<j})$ denotes the predicted probability of the word w_j given the prompt C_{gen}^e and the words proceeding the j-th position. Ultimately, we combine the emotion-aligned response u_t^e with the recommendation response u_t^r to formulate the final response delivered to the users.

Following Wang et al. [34], we choose DialoGPT [39] as the PLM for emotion-aligned response generation (ECR[DialoGPT]). Given DialoGPT's limited parameters, which inherently restrict its linguistic capabilities, we introduce an alternative version by using Llama 2-7B-Chat [31] to which we refer as ECR[Llama 2-Chat]. This choice is motivated by that Llama 2-Chat is an open-source, powerful LLM instruction-tuned on chat tasks. This variation allows us to evaluate our framework's performance based on LLMs.

5 EXPERIMENTS

We address the following research questions: (**RQ1**) Does ECR learn user preferences by capturing their emotions to improve the accuracy of recommendation? (**RQ2**) Is ECR capable of expressing emotions in response generation, thereby improving the user satisfaction? (**RQ3**) How does each component of ECR contribute to its overall performance?

5.1 Dataset

The ReDial dataset [23] is a large-scale CRS dataset, carefully curated by human workers [9]. Consequently, it effectively reflects real-world CR scenarios and fully validates the effectiveness of our method. Considering the significant cost of emotion annotations and evaluations in the generation subtask, we use the ReDial dataset for experiments and plan to extend ECR to other datasets in future work. The ReDial dataset is composed of two-party dialogues between a user and a recommender in the movie domain. It contains 10,006 conversations consisting of 182,150 utterances related to 51,699 movies. The user feedback towards items recommended in the dataset includes three categories: "like," "dislike," and "not say." Previous works [28, 34, 44] simply treat all the recommended items as positive labels. However, according to Li et al. [23], the "dislike" and "not say" labels are distributed separately at 4.9% and 14%, indicating the previous works introduce a large number of incorrect item labels. In contrast, we distinguish between those items with different user feedback. For emotional responses construction, we filter 34,953 reviews related to 4,092 movies for DialoGPT, and 2,459 reviews related to 1,553 movies for Llama 2-7B-Chat. The filtering process is detailed in Appendix A.2. Following [2], we extract entities mentioned in each utterance and review from DBpedia.

5.2 Baselines

For the item recommendation subtask, we compare our method with several CRS approaches: **KBRD** [2], **KGSF**[41], **RevCore**[28], **UCCR**[44] and **UniCRS** [34] to evaluate the effectiveness of ECR. Specifically, KBRD first uses KGs to enhance the semantics fusion in recommendation and generation systems. It uses transformer [32] for response generation with enhanced modeling of word weights. KGSF integrates both word-oriented and entity-oriented KGs to refine user representations and employs the transformer for response generation. RevCore introduces a review-enhanced framework, using item reviews for improving recommendations and response generation, with a focus on sentiment-aware review selection. UCCR focuses on comprehensive user modeling by considering multi-aspect information from current and historical dialogues, as well as data from look-alike users. UniCRS unifies the recommendation and generation into a prompt learning paradigm by PLM.

For the response generation subtask, our comparison involves the state-of-the-art CRS model **UniCRS** [34], the powerful open-source dialogue LLM **Llama 2-7B-Chat** [31] and two advanced OpenAI models: **GPT-3.5-turbo-instruct** and **GPT-3.5-turbo** [45]. To make the output deterministic, we set temperature = 0 when calling the API. These LLMs are prompted to chat with users, aiming to recommend the item predicted by the recommendation module of ECR. They are all provided with the dialogue history for consistency in evaluation.

5.3 Emotion-enhanced Evaluation Metrics

Our evaluation encompasses subjective and objective metrics to assess recommendation and generation performance respectively, which considers the user satisfaction in real-world CRS scenarios. We discuss more details in Appendix D.

Objective evaluation metrics. For recommendation evaluation, we employed Recall@n (R@n, where $n = 1, 10, 50$) to verify if the top-n recommended items include the target item suggested by the dataset's recommenders. To validate the model's effectiveness in estimating user preferences while negating the logged errors in the dataset, we calculate Recall_True@n (RT@n, where $n = 1, 10, 50$). This metric refines Recall@n but only considers the items that get the user feedback of "like" as the standard answers. Additionally, we incorporate the Area Under the Curve (AUC) metric, which emphasizes the ranking order between recommended items linked to the users' positive and negative feedback.

Subjective evaluation metrics. The generation quality is evaluated across five dimensions: emotional intensity (Emo Int), emotional persuasiveness (Emo Pers), logic persuasiveness (Log Pers), informativeness (Info), and lifelikeness (Life). (a) *Emotional intensity* measures the strength of emotions conveyed to users. (b) *Emotional persuasiveness* gauges the capacity to connect with the user emotionally to persuade users. (c) *Logic persuasiveness* evaluates the use of logical reasoning and coherent arguments to persuade users. (d) *Informativeness* determines the utility of useful information provided by the system. (e) *Lifelikeness* assesses how vivid and engaging the responses are, reflecting their resemblance to natural human communication. The scoring range for these metrics is 0 to 9.

Following Wang et al. [33], we employ an LLM-based scorer capable of automatically assigning scores based on specific prompts to alleviate the evaluation reliance on human annotations and randomly sampling 1,000 examples for evaluation. In this context, GPT-4-turbo from the OpenAI serves as the scoring tool. Given the inherent instability in LLMs, we invite three human annotators to assess the reliability of our LLM-based scorer's evaluation results. The annotators are enlisted to rate 200 examples. Additionally, to ensure the robustness of our evaluation, GPT-4 is also employed as an auxiliary scorer, with results detailed in Appendix D.

5.4 Experimental Settings

We implement ECR with PyTorch. The embedding size of the emotion label is 48. The threshold β and n_f are set to 0.1 and 3. The amount of knowledge triples pn_t and entity pn_e in the emotion-aligned generation prompt is set to 2 and 4. For the feedback-aware item reweighting strategy, we assign the weight scalar of user feedback "like," "dislike," and "not say" to 2.0, 1.0, and 0.5, respectively. The analysis for the hyperparameters, i.e., the weight scalar of user feedback and the amount of knowledge used in the emotion-aligned generation prompt, can be found in Appendix B. In the emotion-aligned response generation process, we use AdamW [26] to optimize the tunable parameters of DialoGPT and fine-tune Llama2-chat with LoRA [11]. We set the learning rate for DialoGPT and Llama2-chat to $1e-4$ and $5e-5$, respectively. The batch size is set to 128 for the emotion-aware item recommendation and 16 for the emotion-aligned response generation. The prompts used for the LLM-based scorers and baselines are detailed in Appendix E.

Table 1: Objective evaluation of item recommendation. The boldface indicates the best result. Significant improvements over best baseline results are marked with * (t-test, $p < 0.05$).

Model	AUC	RT@1	RT@10	RT@50	R@1	R@10	R@50
KBRD	0.503	0.040	0.182	0.381	0.037	0.175	0.360
KGSF	0.513	0.043	0.195	0.383	0.040	0.182	0.361
RevCore	0.514	0.054	0.230	0.410	0.046	0.209	0.390
UCCR	0.499	0.038	0.208	0.423	0.039	0.198	0.407
UniCRS	0.506	0.052	0.229	0.439	0.047	0.212	0.414
ECR	**0.541***	**0.055**	**0.238***	**0.452***	**0.049**	**0.220***	**0.428***

Table 2: Subjective evaluation of LLM-based scorer (GPT-4-turbo) and human annotators for response generation. The boldface indicates the best result. Significant improvements over best baseline results are marked with * (t-test, $p < 0.05$).

	Model	Emo Int	Emo Pers	Log Pers	Info	Life
LLM-based scorer	UniCRS	0.400	0.942	0.793	0.673	2.241
	GPT-3.5-turbo-instruct	1.706	3.043	3.474	2.975	4.182
	GPT-3.5-turbo	2.215	3.754	4.782	4.147	5.338
	Llama 2-7B-Chat	3.934	6.030	5.886	5.904	7.129
	ECR[DialoGPT]	4.011	4.878	4.736	5.094	5.906
	ECR[Llama 2-Chat]	**6.826***	**7.724***	**6.702***	**7.653***	**8.063***
Human annotator	UniCRS	0.947	0.775	1.158	0.380	1.805
	GPT-3.5-turbo-instruct	2.048	2.555	3.265	1.822	3.648
	GPT-3.5-turbo	2.890	3.678	5.323	3.233	5.125
	Llama 2-7B-Chat	4.432	6.152	6.393	5.713	7.463
	ECR[DialoGPT]	5.097	4.817	5.398	4.628	6.385
	ECR[Llama 2-Chat]	**7.130***	**7.575***	**7.403***	**7.172***	**8.468***

5.5 Evaluation on Item Recommendation (RQ1)

We address RQ1 by evaluating the performance of item recommendation; see Table 1. KGSF and RevCore, introducing external knowledge in CRSs, have demonstrated superior performance compared to KBRD, underscoring the significance of external knowledge in recommendations. UCCR also performs well on RT@50 and R@50 by extracting user-centric data from cross-session interactions. UniCRS, which integrates PLM into CRSs, exhibits the best performance among all baselines on RT@n and R@n. Regarding the AUC, a metric previously overlooked but essential for evaluating a model's full alignment with users' needs, we find that all baselines exhibit poor performance, with AUC values approaching 0.5. This finding highlights the considerable challenge faced by CRSs in distinguishing between items receiving positive and negative feedback.

We observe that ECR significantly outperforms all the baselines. Specifically, it shows an improvement of 3.9% and 3.0% over UniCRS in RT@10 and RT@50, respectively. Additionally, ECR demonstrates a significant lead in the AUC metric, as indicated by a 6.9% improvement over UniCRS. These findings confirm the importance of capturing user emotions in enhancing CRSs ability to accurately estimate user preferences and effectively mitigate the impact of incorrect item labels in the dataset.

5.6 Evaluation on Response Generation (RQ2)

To analyze whether ECR is capable of expressing emotions for better user satisfaction, we conduct a comparison of response generation, evaluated by both LLM-based scorer (GPT-4-Turbo) and human annotators, as shown in Table 2. We observe that the evaluation results from the LLM-based scorer and human annotators are essentially consistent. In comparison to all baseline models, we discovered that LLMs in the zero-shot setting significantly outperform UniCRS, which is fine-tuned on the entire ReDial dataset. This indicates the subpar quality of the dataset's standard responses.

We observe that ECR[Llama 2-Chat] surpasses all the baselines. Moreover, ECR[DialoGPT] achieves a comparable performance with GPT-3.5-turbo, despite having notably fewer parameters. In particular for the evaluation results of the LLM-based scorer, ECR shows a large improvement in emotional intensity, with ECR[Llama 2-Chat] and ECR[DialoGPT] increasing by 73.5% and 2.0% compared to Llama 2-7B-Chat. Correspondingly, ECR[Llama 2-Chat] and ECR[DialoGPT] achieve increases of 28.1% and 29.9% over Llama 2-7B-Chat and GPT-3.5-turbo in emotional persuasiveness. It confirms ECR's capability, enhanced by supervised fine-tuning on emotional reviews, to express emotions and improve the user experience on an emotional level. Furthermore, ECR incorporates relevant knowledge of recommended items as a part of the generation prompt, which has a beneficial effect on response generation. This is evidenced by the informativeness metric where ECR[Llama 2-Chat] and ECR[DialoGPT] outperform Llama 2-7B-Chat and GPT-3.5-turbo by 29.6% and 22.8%. Meanwhile, ECR[Llama 2-Chat] also increases by 13.9% over Llama 2-7B-Chat on logic persuasiveness. Lastly, in terms of lifelikeness, ECR[Llama 2-Chat] and ECR[DialoGPT] surpass Llama 2-7B-Chat and GPT-3.5-turbo by 13.1% and 10.6%. Overall, these findings illustrate that ECR is more human-like by expressing emotions, thereby enlivening user experience. This enhancement goes beyond aesthetic improvements, it significantly strengthens user identification and increases satisfaction with the system, which may lead to greater user attachment and increased frequency of use.

Reliability of LLM-based scorer. Considering the wide scoring range and the variability in absolute value across evaluators for scoring on each metric, we use the model rankings within these metrics to calculate Cohen's kappa [17]. The average Cohen's kappa within annotators is 0.82 and between LLM-based scorer and annotators is 0.62, indicating substantial agreement. This result suggests the reliability of the LLM-based scorer. Upon analyzing the human evaluation results, discrepancies are primarily observed in logic persuasiveness, whereas evaluations in the emotional dimension are highly consistent, proving the stability of ECR's capabilities in expressing emotions and increasing user emotional satisfaction. In contrast to the LLM-based scorer's evaluations, human annotators identify that ECR[DialoGPT] exhibits superiority over GPT-3.5-turbo on logic persuasiveness by 1.4%, which is attributable to its more effective highlights on the advantages of recommended items. This finding suggests that even GPT-4-turbo still has a slight gap with humans when performing subjective tasks, which reconfirms the significance of our proposed ECR to satisfy practical user needs.

User satisfaction. To confirm ECR's capability in improving user satisfaction by expressing emotions, we direct the human annotators to rate user satisfaction. Our findings indicate that the proposed emotion-enhanced evaluation metrics effectively reflect user satisfaction, especially lifelikeness, which shows a high correlation with user satisfaction. This evidence confirms that by adopting emotion-rich and human-like expressions, ECR significantly improves user experience and satisfaction. See Appendix D.2 for more details.

Table 3: Results of ablation studies for item recommendation. The boldface indicates the best result. Significant improvements are marked with * (t-test, $p < 0.05$).

Model	AUC	RT@10	RT@50	R@10	R@50
UniCRS	0.506	0.229	0.439	0.212	0.414
ECR[L]	0.535	0.229	0.444	0.213	0.420
ECR[LS]	**0.541**	0.232	0.442	0.216	0.420
ECR[LG]	0.535	0.232	**0.453**	0.216	**0.428**
ECR	**0.541**	**0.238***	0.452	**0.220***	**0.428**

6 MODEL ANALYSIS (RQ3)

6.1 Ablation Studies

ECR has a set of components to improve the performance. To verify their effectiveness, we conduct an ablation study and report the results in Table 3. We considered three variants: (i) ECR[L] retains only the local emotion-aware entity representation; (ii) ECR[LS] includes the local emotion-aware entity representation and the feedback-aware item reweighting strategy; and (iii) ECR[LG] contains local and global emotion-aware entity representations.

Our ablation study indicates that each component contributes positively to the recommendation process, as evidenced by performance drops in all three variants. Specifically, the comparison between ECR[L] and UniCRS demonstrates that integrating user emotions into preference modeling significantly improves the accuracy in recommendation, resulting in marked improvements in AUC, RT@50, and R@50. Additionally, ECR[LS] shows further enhancement in AUC. It shows that by being aware of the feedback of users in the training process, ECR reduces the impact of the incorrect item labels in the dataset. Lastly, ECR[LG] outperforms ECR[L] in RT@*n* and R@*n* by combining collaborative knowledge. The major reason is that the use of user emotions as a filter enables the distillation of useful knowledge from the noisy global entity information.

6.2 Case Studies

In this section, we present an example from the ReDial dataset to intuitively demonstrate how ECR works in generating human-like emotional responses. For more cases please refer to Appendix C. Given the dialogue history (Hist) between a recommender and a user, the responses (Response) from ECR, GPT-3.5-turbo and Llama 2-7B-Chat as well as the corresponding evaluation results of LLM-based scorer are presented in Table 4. We also provide the standard responses from the dataset for comparison. We observe that the standard response rarely expresses emotions and is uninformative, which is hardly attractive and convincing to the users. LLMs, i.e., GPT-3.5-turbo and Llama 2-7B-Chat, prioritize factual and logical interpretability. In the provided example, they try to amplify the user interest in the recommended movie by echoing the mentioned keyword "comedy." Additionally, they often deliver objective movie descriptions, such as the plot, genre, and cast. While this approach could make them informative and logic persuasiveness, it neglects the emotional needs of users, inadequately engaging the user. In contrast, our model simulates the vivid conversational style between human beings, incorporating experiences, subjective viewpoints, and emotions. This strategy constructs a deeper emotional connection with users, enlivens their experience, and leads to increased user satisfaction [4]. Furthermore, ECR maintains high-quality factual information to support its emotional rendering and highlight the strengths of the movie. Therefore, even in logic persuasiveness and informativeness – areas where LLMs traditionally excel – ECR remains highly competitive.

Table 4: A sampled case extracted from the ReDial dataset. Text in a green box represents expressing emotions.

Hist.	**Recommender**: Hello there. **User**: Musicals, *comedy*, action.
Response	**Standard response**: Do you like the Marvel movies, like Avengers: Infinity War?
	GPT-3.5-turbo: If you enjoy *comedy*, I highly recommend watching "The Hangover." It's a hilarious movie that follows a group of friends on a wild and unforgettable adventure in Las Vegas. I think you'll really enjoy it! ***Emo Int:*** *2;* ***Emo Pers:*** *4;* ***Log Pers:*** *4;* ***Info:*** *5;* ***Life:*** 5;
	Llama 2-7B-Chat: Great! Based on your interests, I highly recommend "The Hangover"! It's a hilarious *comedy* movie that will have you laughing out loud. The movie follows a group of friends who travel to Las Vegas for a bachelor party and wake up the next morning with no memory of what happened the night before. They then have to retrace their steps and figure out what happened. It's a really funny and entertaining movie that I think you'll enjoy. Let me know if you have any other questions! ***Emo Int:*** *2;* ***Emo Pers:*** *5;* ***Log Pers:*** *5;* ***Info:*** *7;* ***Life:*** *6*;
	ECR[DialoGPT]: Have you seen The Hangover? I have to say that I was very surprised at how good this movie was. I thought it was going to be a typical "chick flick" but it was much more than that. The story was great and the acting was great. I was laughing so hard I almost cried. I would recommend this movie to anyone who likes to laugh. ***Emo Int:*** *5;* ***Emo Pers:*** *6;* ***Log Pers:*** *6;* ***Info:*** *6;* ***Life:*** 7;
	ECR[Llama 2-Chat]: Have you seen The Hangover? I don't think I have ever laughed so hard at a movie. The acting by Ed Helms was excellent. I have seen every episode of The Office and I have always enjoyed him in it. The other actors did a great job too. I have been looking forward to this movie for months and I wasn't disappointed. The movie was very entertaining and I was laughing throughout the whole movie. The jokes were not crude and the story was good. I would definitely recommend this movie to anyone who enjoys a good *comedy*. ***Emo Int:*** *4;* ***Emo Pers:*** *7;* ***Log Pers:*** *7;* ***Info:*** *8;* ***Life:*** *8* ;

6.3 Generalization of Response Generation

In ECR, we use reviews to supervised fine-tune the emotion-aligned generator, endowing it with the ability to express emotions. This process has resulted in some recommended items being "seen" within the reviews used for training. To determine whether ECR acquires a general ability to generate high-quality emotion-aligned responses, especially for items not encountered in the reviews for the training process, we categorized the 1,000 examples used for the LLM-based scorer evaluation in Section 5.6 into "seen" and "unseen." The results are presented in Table 5. We observe a minimal difference in the generation performance between the "seen" and "unseen" categories. This indicates that ECR, when provided with knowledge relevant to the recommended item as a part of the

Table 5: Subjective evaluation of LLM-based scorer (GPT-4-turbo) for generalization of response generation.

	Model	Emo Int	Emo Pers	Log Pers	Info	Life
Seen	ECR[DialoGPT]	4.035	4.945	4.912	5.282	6.003
	ECR[Llama 2-Chat]	6.759	7.704	6.532	7.616	8.046
Unseen	ECR[DialoGPT]	3.995	4.834	4.620	4.970	5.842
	ECR[Llama 2-Chat]	6.844	7.730	6.749	7.663	8.067

emotion-aligned generation prompt, can generalize to generate persuasive and vivid responses for any item, whether or not it is within the training dataset. We observe ECR[Llama 2-Chat] shows better generalization ability than ECR[DialoGPT]. This is likely due to the superior understanding and representation capabilities of Llama 2-7B-Chat because of its large parameter size. Therefore, Llama 2-7B-Chat inherently provides a certain degree of generalization.

7 CONCLUSION AND FUTURE WORK

To bridge the gap between system outputs and user needs, we proposed ECR framework to introduce empathy into CRSs. It is composed of two key modules: (i) emotion-aware item recommendation, which employs user emotions to augment their preference modeling, and (ii) emotion-aligned response generation, which fine-tunes a PLM to express emotions with retrieval-augmented prompts. For data enlargement, we use LLMs to annotate user emotions and collect emotional reviews from external resources. We proposed new evaluation metrics tailored to user satisfaction in real-world CRS scenarios. Extensive experiments verify the effectiveness of ECR in improving recommendation accuracy and user satisfaction. During the experiments, the evaluation results revealed an insight that informativeness significantly increases when recommending multiple items simultaneously. Moving forward, we plan to explore recommending multiple items concurrently while maintaining the logical coherence of response generation.

REPRODUCIBILITY

This work uses publicly available data. To facilitate reproducibility of the results reported in this paper, the code used is available at https://github.com/zxd-octopus/ECR.

ACKNOWLEDGMENTS

This work was supported by the Natural Science Foundation of China (62272274, 62072279,61902219, 61972234, 62102234, 62202271, 61672324), the Young Elite Scientists Sponsorship Program by CAST (2023QNRC001), the 2023 Tencent Rhino-Bird Research Elite Program, the National Key R&D Program of China with grants No. 2020YFB1406704 and No. 2022YFC3303004, the Natural Science Foundation of Shandong Province (ZR2021QF129, ZR2022QF004), the Key Scientific and Technological Innovation Program of Shandong Province (2019JZZY010129), the Fundamental Research Funds of Shandong University, the Tencent WeChat Rhino-Bird Focused Research Program (WXG-FR-2023-07), Shandong University multi-disciplinary research and innovation team of young scholars (No. 2020QNQT017), the Dutch Research Council (NWO), under project numbers 024.004.022, NWA.1389.20.183, and KICH3.LTP.20.006, and the European Union's Horizon Europe program under grant agreement No 101070212. All content represents the opinion of the authors, which is not necessarily shared or endorsed by their respective employers and/or sponsors.

REFERENCES

[1] Christian Bizer, Jens Lehmann, Georgi Kobilarov, Sören Auer, Christian Becker, Richard Cyganiak, and Sebastian Hellmann. 2009. DBpedia - A Crystallization Point for the Web of Data. *J. Web Semant.* 7, 3 (2009), 154–165.

[2] 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 EMNLP-IJCNLP*. 1803–1813.

[3] Konstantina Christakopoulou, Filip Radlinski, and Katja Hofmann. 2016. Towards Conversational Recommender Systems. In *Proceedings of SIGKDD*. 815–824.

[4] Lorenzo Cominelli, Francesco Feri, Roberto Garofalo, Caterina Giannetti, Miguel A. Meléndez-Jiménez, Alberto Greco, Mimma Nardelli, Enzo Pasquale Scilingo, and Oliver Kirchkamp. 2021. Promises and Trust in Human–robot Interaction. *Scientific Reports* 11 (2021).

[5] Sunhao Dai, Ninglu Shao, Haiyuan Zhao, Weijie Yu, Zihua Si, Chen Xu, Zhongxiang Sun, Xiao Zhang, and Jun Xu. 2023. Uncovering ChatGPT's Capabilities in Recommender Systems. In *Proceedings of RecSys*. 1126–1132.

[6] Cach N. Dang, María N. Moreno García, and Fernando de la Prieta. 2021. An Approach to Integrating Sentiment Analysis into Recommender Systems. *Sensors* 21, 16 (2021), 5666.

[7] 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 SIGIR*. 1431–1441.

[8] 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 Trans. Inf. Syst.* 41, 3 (2023), 77:1–77:25.

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

[10] Zhankui He, Zhouhang Xie, Rahul Jha, Harald Steck, Dawen Liang, Yesu Feng, Bodhisattwa Prasad Majumder, Nathan Kallus, and Julian J. McAuley. 2023. Large Language Models as Zero-Shot Conversational Recommenders. In *Proceedings of CIKM*. 720–730.

[11] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-Rank Adaptation of Large Language Models. In *Proceedings of ICLR*.

[12] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of Hallucination in Natural Language Generation. *ACM Comput. Surv.* 55, 12 (2023), 248:1–248:38.

[13] Ikram Karabila, Nossayba Darraz, Anas El-Ansari, Nabil Alami, and Mostafa El Mallahi. 2023. Enhancing Collaborative Filtering-Based Recommender System Using Sentiment Analysis. *Future Internet* 15, 7 (2023), 235.

[14] Sudhanshu Kumar, Kanjar De, and Partha Pratim Roy. 2020. Movie Recommendation System Using Sentiment Analysis From Microblogging Data. *IEEE Trans. Comput. Soc. Syst.* 7, 4 (2020), 915–923.

[15] Robert J. Kwortnik and William T. Ross. 2007. The Role of Positive Emotions in Experiential Decisions. *International Journal of Research in Marketing* 24, 4 (2007), 324–335.

[16] Claus Lamm and Giorgia Silani. 2014. Insights into Collective Emotions from the Social Neuroscience of Empathy. In *Collective Emotions*. Oxford Academic, 63–77.

[17] J. Richard Landis and Gary G. Koch. 1977. The Measurement of Observer Agreement for Categorical Data. *Biometrics* 33 1 (1977), 159–74.

[18] 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 WSDM*. 304–312.

[19] Wenqiang Lei, Gangyi Zhang, Xiangnan He, Yisong Miao, Xiang Wang, Liang Chen, and Tat-Seng Chua. 2020. Interactive Path Reasoning on Graph for Conversational Recommendation. In *Proceedings of KDD*. 2073–2083.

[20] Jennifer S. Lerner, Ye Li, Piercarlo Valdesolo, and Karim S. Kassam. 2015. Emotion and Decision Making. *Annual Review of Psychology* 66 (2015), 799–823.

[21] Qintong Li, Hongshen Chen, Zhaochun Ren, Pengjie Ren, Zhaopeng Tu, and Zhumin Chen. 2020. EmpDG: Multi-resolution Interactive Empathetic Dialogue Generation. In *Proceedings of COLING*. 4454–4466.

[22] Qintong Li, Piji Li, Zhaochun Ren, Pengjie Ren, and Zhumin Chen. 2022. Knowledge Bridging for Empathetic Dialogue Generation. In *Proceedings of AAAI*. 10993–11001.

[23] Raymond Li, Samira Ebrahimi Kahou, Hannes Schulz, Vincent Michalski, Laurent Charlin, and Chris Pal. 2018. Towards Deep Conversational Recommendations. In *Proceedings of NeurIPS*. 9748–9758.

[24] Shuokai Li, Ruobing Xie, Yongchun Zhu, Xiang Ao, Fuzhen Zhuang, and Qing He. 2022. User-Centric Conversational Recommendation with Multi-Aspect User Modeling. In *Proceedings of SIGIR*. 223–233.

[25] Wendi Li, Wei Wei, Xiaoye Qu, Xian-Ling Mao, Ye Yuan, Wenfeng Xie, and Dangyang Chen. 2023. TREA: Tree-Structure Reasoning Schema for Conversational Recommendation. In *Proceedings of ACL*. 2970–2982.

[26] Ilya Loshchilov and Frank Hutter. 2019. Decoupled Weight Decay Regularization. In *Proceedings of ICLR*.

[27] Yu Lu, Junwei Bao, Yan Song, Zichen Ma, Shuguang Cui, Youzheng Wu, and Xiaodong He. 2021. RevCore: Review-Augmented Conversational Recommendation. In *Proceedings of ACL/IJCNLP*. 1161–1173.

[28] Wenchang Ma, Ryuichi Takanobu, and Minlie Huang. 2021. CR-Walker: Tree-Structured Graph Reasoning and Dialog Acts for Conversational Recommendation. In *Proceedings of EMNLP*. 1839–1851.

[29] Zhaochun Ren, Zhi Tian, Dongdong Li, Pengjie Ren, Liu Yang, Xin Xin, Huasheng Liang, Maarten de Rijke, and Zhumin Chen. 2022. Variational Reasoning about User Preferences for Conversational Recommendation. In *Proceedings of SIGIR*. 165–175.

[30] Michael Sejr Schlichtkrull, Thomas N. Kipf, Peter Bloem, Rianne van den Berg, Ivan Titov, and Max Welling. 2018. Modeling Relational Data with Graph Convolutional Networks. In *Proceedings of ESWC*. 593–607.

[31] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton-Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurélien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models. *CoRR* abs/2307.09288 (2023).

[32] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is All You Need. In *Proceedings of NeurIPS*. 5998–6008.

[33] Xiaolei Wang, Xinyu Tang, Xin Zhao, Jingyuan Wang, and Ji-Rong Wen. 2023. Rethinking the Evaluation for Conversational Recommendation in the Era of Large Language Models. In *Proceedings of EMNLP*. 10052–10065.

[34] Xiaolei Wang, Kun Zhou, Ji-Rong Wen, and Wayne Xin Zhao. 2022. Towards Unified Conversational Recommender Systems via Knowledge-Enhanced Prompt Learning. In *Proceedings of KDD*. 1929–1937.

[35] Bowen Yang, Cong Han, Yu Li, Lei Zuo, and Zhou Yu. 2022. Improving Conversational Recommendation Systems' Quality with Context-Aware Item Meta-Information. In *Proceedings of NAACL*. 38–48.

[36] Zhou Yang, Zhaochun Ren, Wang Yufeng, Xiaofei Zhu, Zhihao Chen, Tiecheng Cai, Yunbing Wu, Yisong Su, Sibo Ju, and Xiangwen Liao. 2023. Exploiting Emotion-Semantic Correlations for Empathetic Response Generation. In *Proceedings of EMNLP*. 4826–4837.

[37] Gangyi Zhang. 2023. User-Centric Conversational Recommendation: Adapting the Need of User with Large Language Models. In *Proceedings of RecSys*. 1349–1354.

[38] 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 WSDM*. 231–239.

[39] Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2020. DIALOGPT : Large-Scale Generative Pre-training for Conversational Response Generation. In *Proceedings of ACL*. 270–278.

[40] 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 EMNLP*. 4324–4334.

[41] Kun Zhou, Wayne Xin Zhao, Shuqing Bian, Yuanhang Zhou, Ji-Rong Wen, and Jingsong Yu. 2020. Improving Conversational Recommender Systems via Knowledge Graph based Semantic Fusion. In *Proceedings of KDD*. 1006–1014.

[42] Kun Zhou, Wayne Xin Zhao, Hui Wang, Sirui Wang, Fuzheng Zhang, Zhongyuan Wang, and Ji-Rong Wen. 2020. Leveraging Historical Interaction Data for Improving Conversational Recommender System. In *Proceedings of CIKM*. 2349–2352.

[43] Kun Zhou, Yuanhang Zhou, Wayne Xin Zhao, Xiaoke Wang, and Ji-Rong Wen. 2020. Towards Topic-Guided Conversational Recommender System. In *Proceedings of COLING*. 4128–4139.

[44] Yuanhang Zhou, Kun Zhou, Wayne Xin Zhao, Cheng Wang, Peng Jiang, and He Hu. 2022. C2-CRS: Coarse-to-Fine Contrastive Learning for Conversational Recommender System. *CoRR* abs/2201.02732 (2022).

[45] Barret Zoph, Christina Kim, Jacob Hilton, Jacob Menick, Jiayi Weng, Juan Felipe, Ceron Uribe, Liam Fedus, Luke Metz, Michael Pokorny, Rapha Gontijo Lopes, and Shengjia Zhao. 2022. ChatGPT: Optimizing Language Models for Dialogue. OpenAI (2022).