ECC: An Emotion-Cause Conversation Dataset for Empathy Response

Yuanyuan He[1,2], **Yongsen Pan**[1,2], **Wei Li**[1], **Jiali You**[1], **Jiawen Deng**[1*], **Fuji Ren**[1,2*]
[1]University of Electronic Science and Technology of China, Chengdu, China
[2]Shenzhen Institute for Advanced Study, UESTC, Shenzhen, China
`{hyy, panys, liwei39}@std.uestc.edu.cn,`
`yjl1015004@163.com, {dengjw, renfuji}@uestc.edu.cn`

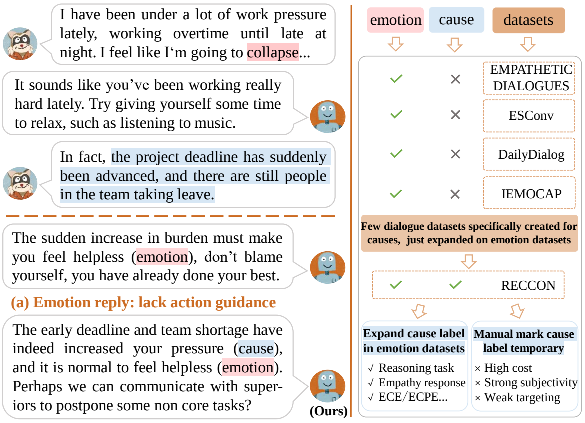
Abstract

The empathy dialogue system requires understanding emotions and their underlying causes. However, existing datasets mainly focus on emotion labels, while cause annotations are added post hoc through costly and subjective manual processes. This leads to three limitations: subjective bias in cause labels, weak rationality due to ambiguous cause-emotion relationships, and high annotation costs that hinder scalability. To address these challenges, we propose ECC (**E**motion-**C**ause **C**onversation Dataset), a scalable dataset with 2.4K dialogues, which is also the first dialogue dataset where conversations and their emotion-cause labels are automatically generated synergistically during creation. We create an automatic extension framework EC-DD for ECC that utilizes knowledge and large language models (LLMs) to automatically generate conversations, and train a causality-aware empathetic response model CAER on this dataset. Experimental results show that ECC can achieve comparable or even superior performance to artificially constructed empathy dialogue datasets. Our code will be publicly released on `https://github.com/Yuan-23/ECC`.

1 Introduction

The field of empathetic response aims to improve interpersonal communication by understanding and reflecting others' emotions (Huang et al., 2024; Hsu et al., 2023). While emotions have been extensively studied, the ability to identify and utilize emotion causes, the underlying triggers of emotions, remains underexplored but critical for achieving deeper conversational understanding.

Emotion-cause analysis not only enriches emotional reasoning (*e.g.*, answering "Why does the user feel this way?") but also guides models to generate actionable and logically coherent responses, moving beyond generic platitudes, like "I'm sorry you feel this way." As Figure 1, responses focusing solely on emotion labels often fail to address the root cause of distress (Figure 1a, empty talk of encouragement has little effect), whereas integrating cause-aware reasoning enables targeted suggestions (Figure 1b, proposing workload management strategies for burnout caused by overtime work).



I have been under a lot of work pressure lately, working overtime until late at night. I feel like I'm going to collapse...

It sounds like you've been working really hard lately. Try giving yourself some time to relax, such as listening to music.

In fact, the project deadline has suddenly been advanced, and there are still people in the team taking leave.

The sudden increase in burden must make you feel helpless (emotion), don't blame yourself, you have already done your best.

(a) Emotion reply: lack action guidance

The early deadline and team shortage have indeed increased your pressure (cause), and it is normal to feel helpless (emotion). Perhaps we can communicate with superiors to postpone some non core tasks?

(b) Emotion-Cause reply: grasp the core issue

(c) The datasets lack cause labels

Figure 1: (a) focus on emotions. (b) simultaneously focus on emotions and causes. (c) the cause labels can't be annotated during the generation phase.

Despite its importance, existing dialogue datasets mainly focus on emotion labels alone, like EMPATHETIC DIALOGUES (Rashkin et al., 2019) and ESConv (Liu et al., 2021), while cause labels are typically added through post hoc manual annotation, such as RECCON (Poria et al., 2021). This approach introduces three critical limitations, as shown in Figure 1c: **Subjectivity Bias.** Cause labels are highly context-dependent and vary significantly across annotators (Xiao et al., 2023; Cheng et al., 2023). **Weak Rationality.** Most dialogues are originally designed without cause considerations, resulting in ambiguous or mixed causes during retrospective labeling (Liu et al., 2021; Gupta and Dandapat, 2023). **High Annotation Cost.** Manual cause labeling via crowdsourcing or expert annotation is resource-intensive and scales poorly(Poria et al., 2021), limiting dataset utility for large-scale model training.

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 6000–6017
November 4-9, 2025 ©2025 Association for Computational Linguistics

To address these challenges, we introduce ECC (**E**motion-**C**ause **C**onversation Dataset), the first dialogue dataset automatically generated with joint emotion-cause labels during its creation phase. Unlike conventional datasets that retrofit cause labels post hoc, ECC explicitly links each emotional expression to its contextual cause before generation (*e.g.*, "I'm anxiety [emotion] because my project deadline was abruptly moved up [cause]"). This approach ensures three key advantages: **Cause-Emotion Cohesion**: Eliminates the semantic gap between emotions and causes by integrating causal relationships into the dialogue context in advance. **Reduced Subjectivity**: Automated generation minimizes human bias, ensuring label consistency across diverse scenarios. **Scalability**: Leveraging LLMs, ECC can be efficiently expanded to support evolving research needs. Our contributions:

- **E**motion-**C**ause **D**ialogues **D**istillation Framework (EC-DD): We propose a novel framework that integrates commonsense knowledge distillation into LLMs to automate dialogues generation and emotion-cause extraction.
- **E**motion-**C**ause **C**onversations dataset (ECC): We construct and release ECC, which both conversations and their emotion-cause labels are generated synergistically without manual intervention.
- **C**ausality **A**ware **E**mpathetic **R**esponse Model (CAER): Leveraging ECC, we train CAER, a model that prioritizes both emotional alignment and causal reasoning. Extensive experiments demonstrate that CAER can generate better empathetic responses.

2 Related Work

2.1 Empathetic Dialog Dataset Construction

Collecting real data on empathetic conversations is challenging(Cai et al., 2023; Zhao et al., 2023), so these two main methods are currently commonly used: Crowdsourcing and data collection. Many conversation datasets rely on crowdsourcing, which involves recruiting workers for simple tasks and providing them with themes, strategies, or personality traits to guide their interactions, as seen in works like (Liu et al., 2021; Rashkin et al., 2019; Zhang et al., 2018). The other method involves collecting data from existing sources by extracting conversations from websites, novels, news articles, or drama scripts, and workers are then recruited to tag emotion labels, as demonstrated in (Li et al., 2017; Gui et al., 2018; Gupta and Dandapat, 2023).

2.2 Emotion-Cause Labels

It is clear that understanding the reasons behind others' emotions allows us to better comprehend their feelings. Based on this, extraction tasks related to emotion and cause labels (ECE/ECPE) have been proposed (Lee et al., 2010; Gui et al., 2018; Xia and Ding, 2019). However, these tasks have focused on narrative texts (Xiao et al., 2023; Cheng et al., 2023) and do not address conversational texts, which are essential for empathetic responses. Consequently, Poria et al. (2021) developed the RECCON dataset, which combines portions of the IEMOCAP (Busso et al., 2008) and DailyDialog (Li et al., 2017) datasets, manually adding emotion and cause labels. Although it is not designed for empathy tasks, still sparked interest in exploring reasons(Chen et al., 2024a,b) or extracting emotion and cause (An et al., 2023; Chen et al., 2023; Zou et al., 2024; Su et al., 2024) in the field of empathy.

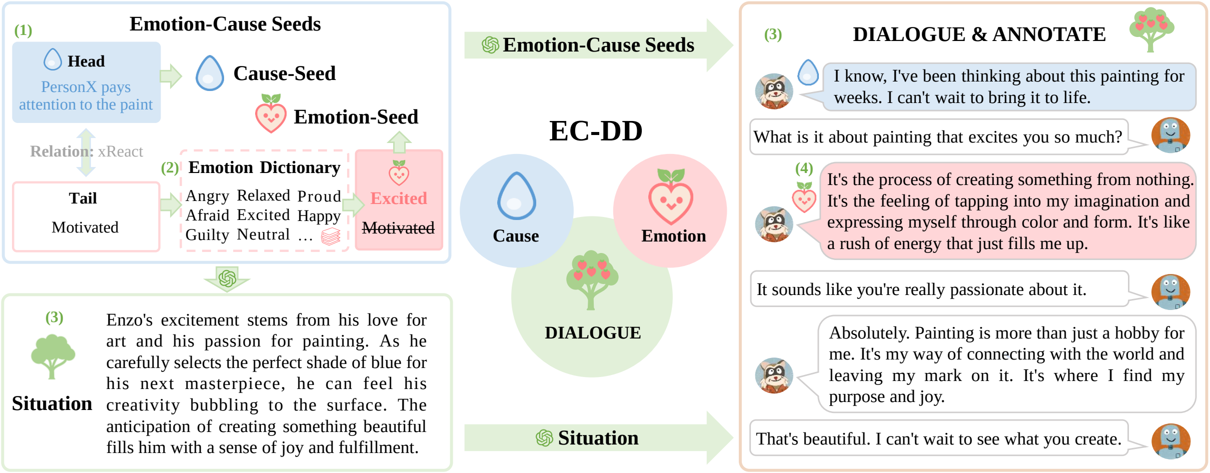
3 EC-DD Framework

We develop Emotion-Cause Dialogues Distillation (EC-DD) Framework to create the ECC dataset. It integrates commonsense knowledge distillation into LLMs to automate dialogues generation and emotion-cause label extraction. As shown in Figure 2, the framework consists of four core steps. **(i) Extract Emotion-Cause Seeds. (ii) Processing of Seeds. (iii) Dialogues Generation. (iv) Emotion-Cause Labels Annotation.**

Formally, given a commonsense triplet $xReact = \{h, r, t\}$, h represents head, r represents relationship, t represents tail, the LLM (M) generates a dialogue (D) based on the triplet. The dialogue is defined as $D = \{u_1, u_2, u_3, ..., u_n\}$, where u_i represents the utterance at the i_{th} turn. Our goal is to have M generate the complete conversation and annotate the emotion (e) label u_e and cause (c) label u_c based on triplets.

3.1 Extract Emotion-Cause Seeds

Why choose xReact triplets? The commonsense knowledge dataset Atomic2020 (Hwang et al., 2021) proposes 23 types of commonsense relationships, including seven event-centered relationships. (West et al., 2022) utilized LLMs to distill these relationships starting with "x" into seven types



Enzo's excitement stems from his love for art and his passion for painting. As he carefully selects the perfect shade of blue for his next masterpiece, he can feel his creativity bubbling to the surface. The anticipation of creating something beautiful fills him with a sense of joy and fulfillment.

I know, I've been thinking about this painting for weeks. I can't wait to bring it to life.

What is it about painting that excites you so much?

It's the process of creating something from nothing. It's the feeling of tapping into my imagination and expressing myself through color and form. It's like a rush of energy that just fills me up.

It sounds like you're really passionate about it.

Absolutely. Painting is more than just a hobby for me. It's my way of connecting with the world and leaving my mark on it. It's where I find my purpose and joy.

That's beautiful. I can't wait to see what you create.

Figure 2: **EC-DD** Framework: (1) Extract emotion-cause seeds from the commonsense database. (2) Map emotions to types in the emotion dictionary. (3) Use these seeds to generate situations and dialogues via LLMs. (4) Label emotion and cause sentences in generated dialogues.

Head	Relation	Tail
X gets an advance	xReact (so, X feels)	proud
X is not giving up	xWant (so, X wants)	to succeed
X talks with his boss	xAttr (X is seen as)	polite

Table 1: Some Cases of ATOMIC triplets.

of symbolic knowledge, including {HinderedBy, xNeed, xWant, xContent, xReact, xAttr, and xEffect}, in the format of a triplet (Head, Relation, Tail), as shown in Table 1. The xReact ralation captures an individual's reaction to a specific situation, *i.e.*, (PersonX pays attention to the paint, xReact, Motivated). It can be seen that the "Head" contains the cause of an event, while the "Tail" contains the emotional response of the protagonist "Personx" to the event. Therefore, we extract the xReact triplet as our foundational data.

Extraction of Emotion-Cause Seeds. The connection of the parts of the xReact triplet can be seen as ("Head", so, X feels "Tail") (West et al., 2022). We use S_e to represent emotion seed and S_c to represent cause seed. The triplet's "Tail" indicates the speaker's emotion, so $S_e = Tail$, while the "Head" provides the cause of emotion, so $S_c = Head$.

3.2 Processing of Seeds

Creation of Emotion Dictionary. Due to the complexity of emotion seed categories, we created an emotion dictionary (Poria et al., 2021) for classification. Specifically, we analyze the number and frequency of emotion seeds, with considering the classification methods used in previous papers (Busso et al., 2008; Li et al., 2017), we extract ten emotions as category labels. We set positive categories (*happy, excited, confident, relaxed, proud*) and negative categories (*angry, frustrated, sad, guilty, afraid*). In addition, we have a *neutral* category that includes words with ambiguous emotion tendencies. During the dialogue generation stage, we discard neutral data to avoid generating conversations with unclear emotional labels.

Mapping Emotion to Category. Based on the classification defined in the emotion-dictionary, we use GPT-3.5 to categorize the emotion seeds according to their emotional tendencies and meanings, consolidating them into eleven categories (including *neutral*). To ensure classification accuracy, we enlist three expert evaluators. Each expert receives prior training, classifying over 500 emotion words. They review the model's categorization of each word, and in cases where two experts disagreed, the third expert's judgment is used to resolve the conflict. The complete emotion dictionary can be found in Appendix A.1.

3.3 Dialogues Generation

Generation of Situation. Inspired by the social dialogue framework (Kim et al., 2023), we combine emotion-cause seeds as the core theme of the conversation, which is then provided as input to GPT-3.5, to generate the "Situation" (T):

$$T = M(S_c \oplus c \oplus S_e \oplus P_s) \qquad (1)$$

S_c represents the triplet connector "so, personX

Emotion Seed: Excited

Cause Seed: PersonX pays attention to the paint

Situation: Enzo's excitement stems from his love for art and his passion for painting. As he carefully selects the perfect shade of blue for his next masterpiece, he can feel his creativity bubbling to the surface. The anticipation of creating something beautiful fills him with a sense of joy and fulfillment.

Dialogue:

Enzo: I know, I've been thinking about this painting for weeks. I can't wait to bring it to life.

Friend: What is it about painting that excites you so much?

Enzo: It's the process of creating something from nothing. It's the feeling of tapping into my imagination and expressing myself through color and form. It's like a rush of energy that just fills me up.

Friend: It sounds like you're really passionate about it.

Enzo: Absolutely. Painting is more than just a hobby for me. It's my way of connecting with the world and leaving my mark on it. It's where I find my purpose and joy.

Friend: That's beautiful. I can't wait to see what you create.

Score: 5 (The better emotion / cause sentence is null)

Evaluation: The emotion and cause sentences align well with the dialogue and seeds. Enzo expresses excitement about painting, detailing that the creation process, and opportunity to self-expression are the causes of his excitement.

Table 2: Each conversation in ECC contains emotion-cause seeds, situations, emotion sentence annotations, cause sentence annotations, and complete dialogue information.

feelings", and P_s is a prompt[1] that emphasizes the logic of emotion and cause. M generates a comprehensive story description based on the core theme as the background context for the dialogue, as shown in the second row of Table 2.

Generation of Dialogue. We re-input the situation T as background to GPT-3.5 to generate binary dialogues D:

$$D = M(T \oplus P_d) \tag{2}$$

P_d is a prompt[1] used to generate dialogues, and the third row of Table 2 shows a complete dialogue.

Filtering and Regenerate. We have incorporated semantic verification and safety screening modules to filter generated dialogues. For responses that deviate from the intended semantic context, the model is prompted to regenerate the dialogue to ensure consistency. To prevent dialogue patterns from unduly influencing emotional tone, we have systematically filtered out emotional annotations and character action descriptors, *i.e.*, "*Selena (crying sadly): I want to try too*", "*Edward (stumbling): Yes, we did it*". Furthermore, we implemented the Rewire API[2] for toxic content detection, including but not limited to biased, discriminatory, or violent material. Dialogues exceeding a predefined toxicity threshold (0.6) are automatically discarded.

3.4 Emotion-Cause Labels Annotation

Annotation of Emotions and Causes. We use GPT-4o to annotate each binary, multi-round dialogue. Specifically, we input the emotion-cause dialogue along with the corresponding emotion-cause seeds into M:

$$u_e, u_c = M(D \oplus S_e \oplus S_c \oplus P_l) \tag{3}$$

Among them, P_l is the prompt[1] we designed to search for u_e that matches S_e and u_c that corresponds to S_c. For example, in Table 2, based on the seed S_e and S_c in the first row, Mark u_e and u_c throughout the entire conversation in the third line. Given the high coherence and consistency of emotions throughout a conversation, we instruct the model to select the sentences that best capture both the emotion and its underlying cause.

Quality Evaluation of Labels. To confirm the quality of labels, we conduct an automatic evaluation of the rationality using GPT-4.0, with a scoring system out of 5, where 1 is unreasonable and 5 is reasonable. Given the LLMs may overly rely on prompts for evaluation, there is a risk of overlooking the diversity and nuances of dialogues. To address this, we request LLMs provide reasoning during the evaluation process. Additionally, to challenge the model's inherent conservatism, we ask the LLMs to suggest more reasonable emotion and cause sentences. The scoring process is shown in the last row of Table 2, and we incorporate the LLMs' reasoning process into the dataset as a reference. The final average score for ECC is 4.2, which preliminarily validates the reasonableness of the labels. We will reconfirm the quality of the labels through human evaluation in section 4.3.

---

[1] All prompt templates are shown in the appendix A.2

[2] https://rewire.online/

Dataset	Method	Source	Label	Size	Avg.Turn / Uttr	MTLD ↑
DailyDialog	Manual	Collection	Emotion	13K	7.9 / 11.2	41.1
CHASE	Manual	Collection	Null	4K	2.7 / 68.1	35.0
ED	Manual	Writing	Emotion	25K	4.3 / 13.7	41.2
ESConv	Manual	Writing	Emotion	1K	29.8 / 17.8	67.6
RECCON	Manual	Writing	Emotion + Cause	1K	10.5 / 13.8	10.8
SODA	Automatic	LLMs	Null	1.5M	7.6 / 16.1	68.0
ECC	Automatic	LLMs	Emotion + Cause	2.4K	13.0 / 22.0	55.3

Table 3: Comparison of dialogue datasets based on different production methods, data sources, label types, dataset sizes, average dialogue turns/utterances, and MTLD scores. The MTLD (Measure of Textual Lexical Diversity) score is used to evaluate the diversity of the conversations in each dataset.

4 ECC Dataset

Through the EC-DD framework, we extract an ***E**motion-**C**ause **C**onversation* (ECC) Dataset from LLMs. We generate 3,000 data entries, and after filtering, 2,398 dialogues with emotion-cause labels are retained. Each conversation contains the emotion-cause seed, situation, emotion sentence annotation, cause sentence annotation, and complete dialogue information. Notably, ECC is the first dialogue dataset with emotion-cause labels that is generated without manual writing or labeling.

4.1 Data Quality

We compare ECC with various datasets commonly used in the field of emotional dialogue, the following datasets are included: **DailyDialog** (Li et al., 2017) includes 13,118 dialogues annotated with 7 emotions for emotion recognition and dialogue generation. **CHASE** (Gupta and Dandapat, 2023) comprises approximately 42,293 utterances from famous plays, generating 4.1k conversations with manually crafted replies. **EMPATHETIC DIALOGUES (ED)** (Rashkin et al., 2019) aims for empathetic conversation generation, containing 25,000 manually annotated dialogues with 32 emotion labels. **ESConv** (Liu et al., 2021) introduces the Emotional Support Dialogue task, with 1,053 dialogues annotated for 7 negative emotions. **RECCON** (Poria et al., 2021) includes 1,106 dialogues focusing on reasoning, with emotion and cause annotations. **SODA** (Kim et al., 2023) contains 1.5 million conversations and over 11 million utterances focused on social dialogue, lacking empathy or emotional cause labels.

As the first dataset automatically generated by LLMs with simultaneous emotion and cause labeling, ECC preserves dialogue complexity while overcoming the limitations of manual annotation. As shown in Table 3, compared to other manually annotated datasets (*i.e.*, DailyDialog, ED, and ESConv, which only label emotions), ECC not only provides joint emotion-cause labels (similar to RECCON) but also expands the dataset size to 2.4K (RECCON has 1K), and avoids the subjectivity and high costs associated with manual annotation through automatic generation. In terms of language quality, ECC's MTLD (McCarthy and Jarvis, 2010) score is significantly higher than most manually annotated datasets (*e.g.*, CHASE's 35.0 and RECCON's 10.8), indicating greater diversity in its dialogue texts. Additionally, its average turn length (13.0 turns) and utterance length (22.0 words) outperform most similar datasets, suggesting that the dialogue content has better contextual coherence and logical depth. Although the dataset size is smaller than the purely automatically generated SODA, ECC fills the gap in joint emotion-cause modeling within automatically generated datasets, providing a more structured and scalable training resource for models to understand emotional causality. Appendix B has more analysis.

4.2 Automatic Evaluation

We compare the performance of the automatically generated ECC dataset with two manually annotated datasets, ED (emotion labels only) and RECCON (emotion-cause labels), to evaluate their effectiveness in training emotion-aware dialogue models. We fine-tune the same base model on all three datasets using identical hyperparameters and an 8:1:1 train/test/validation split. To ensure fairness, we align the dataset sizes uniformly to ECC's size. Specifically, we randomly select complete dialogue turns from ED to match the target size, while performing data augmentation on the under-

Model	Loss ↓			PPL ↓			Bleu-1 ↑			Rouge-1 ↑			ACC ↑		
	ED	REC	ECC	ED	REC	ECC	ED	REC	ECC	ED	REC	ECC	ED	REC	ECC
MOEL	5.9692	5.4379	**4.1258**	391.20	229.96	**61.92**	0.1814	0.2651	**0.2986**	0.1603	0.2445	**0.2953**	0.2746	**0.5123**	0.4810
MIME	5.9264	5.5050	**3.9132**	374.81	245.92	**50.06**	0.1735	0.2735	**0.3119**	0.1720	0.2455	**0.2730**	0.2355	**0.5167**	0.3917
EmpDG	5.2063	5.8464	**3.5435**	182.42	345.98	**34.59**	0.1794	0.2459	**0.2602**	0.1738	0.2397	**0.2922**	0.1250	**0.4442**	0.3984
Multi-Trs	5.4244	5.7590	**3.6698**	226.88	317.05	**39.24**	0.1869	**0.3086**	0.3030	0.1824	0.2442	**0.2561**	0.2143	**0.4721**	0.4375
Trs	5.5552	5.7841	**3.6639**	258.57	325.07	**39.01**	0.1711	0.2527	**0.3184**	0.1853	0.2527	**0.2552**	-	-	-

Table 4: ED represents EMPATHETIC DIALOGUES dataset, and REC is RECCON dataset. All experiments follow the same experimental setup, and the dataset ratio follows 8:1:1. **Bold** indicates the best, followed by underline.

sized RECCON to fill the gap. We evaluate model performance using metrics across four dimensions: model loss (LOSS), perplexity (PPL), generative quality (Blue-1, Rouge-1), and emotion accuracy (ACC). The baseline models are as follows:

MIME (Majumder et al., 2020) generates diverse emotional responses by simulating mixed emotions in conversations. **MOEL** (Lin et al., 2019) uses multiple emotion-specific listeners to produce responses aligned with the speaker's emotions. **EmpDG** (Li et al., 2020) enhances empathetic dialogue by predicting emotions and generating context-appropriate responses. **Multi-Transformer** (Rashkin et al., 2019) improves multi-turn dialogue coherence by capturing dependencies across turns. **Transformer** (Vaswani, 2017) captures relationships by self-attention, excelling in text generation and emotion recognition.

As shown in Table 4, ECC exhibits superior learning efficiency and text generation capability, as evidenced by its significantly lower loss and perplexity. Although RECCON's granular labeling enhances emotion recognition precision, ECC generates smoother and more contextually grounded responses, bridging the gap between manual datasets and LLM-generated data. And ECC, equipped with both emotion and cause labels, consistently outperforms the emotion-only ED dataset across all models in accuracy (ACC), while RECCON achieves the highest accuracy due to its fine-grained utterance-level emotion annotations (versus dialogue-level in ECC/ED). Remarkably, ECC matches or even surpasses manually curated datasets in overall performance, proving that automated emotion-cause labeling can rival human annotation quality when integrated with structured causal reasoning. We also present experimental results for both the original dataset size and a resized version aligned with the RECCON dataset scale. In both cases, ECC demonstrates highly competitive performance. For detailed comparative analysis, please refer to Appendix B.3.

Judge	Aff	Rel	Con	Nat	Overall
human 1	4.52	4.24	4.07	4.47	4.33
human 2	5.00	4.54	4.56	4.75	4.71
human 3	5.00	4.53	4.62	5.00	4.79
Average	4.84	4.44	4.42	4.74	4.61

Table 5: The results of human evaluation for ECC.

	human1&2	human1&3	human2&3	Avg
PCC ↑	0.8574	0.8788	0.8737	0.8700

Table 6: PCC among three humans. The range is 0-1.

4.3 Human Evaluation

We invited three human evaluators, who had received training in assessing emotional dialogues, to conduct a manual evaluation of 100 randomly selected dialogues from the ECC dataset. Four evaluation criteria were established: Affective, Relevance, Consistency, and Naturalness, along with an Overall score. Each evaluator independently rated all dialogues, and the final results are presented in Table 5. Additionally, to assess the consistency among evaluators, we calculated the Pearson correlation coefficient (PCC) between their ratings, as shown in Table 6. The evaluation criteria are as follows: **Affective(Aff)** measures the authenticity and intensity of emotional expression. **Relevance(Rel)** measures the association between dialogue content and seeds. **Consistency(Con)** evaluates the internal coherence of emotion and cause annotations. **Naturalness(Nat)** assesses the naturalness and fluency of the dialogue. We also used the latest release of DeepSeek-v3 (671B) for the same evaluation based on human evaluation metrics, and the

Model	PPL ↓	DIST-1 ↑	DIST-2 ↑	Bleu-1 ↑	Bleu-2 ↑
GPT-3.5	-	0.1742	0.5708	0.0827	0.0029
GPT-4o	-	0.1378	0.5275	0.0547	0.0023
LLM+Kno	-	0.1185	0.4762	0.1053	0.0364
DeepSeek-R1	-	0.0923	0.3497	0.0548	0.0115
T5-Large	12.5243	0.1758	0.1740	0.1121	0.0121
COSMO	18.0440	**0.2907**	**0.7728**	0.1491	0.0053
LLAMA3.2	15.0180	0.1030	0.3939	0.0489	0.0033
CAER	**10.9965**	**0.2907**	0.7045	**0.2287**	**0.0559**

Table 7: Experimental results of models on the RECCON dataset. RECCON is an unseen dataset with emotion and cause labels for all models.

Model	PPL ↓	Bleu-1 ↑	Bleu-2 ↑
LLM+Kno	-	0.0963	**0.0357**
DeepSeek-R1	-	0.0620	0.0113
T5-Large	15.4101	0.1532	0.0181
COSMO	15.9213	0.1078	0.0100
LLAMA3.2	10.7275	0.0446	0.0043
CAER	**7.8716**	**0.1571**	0.0212

Table 8: Experimental results of models on the ED dataset. ED is an unseen dataset without cause labels.

results and analysis are shown in Appendix B.5.

The evaluation results indicate that ECC performed well across all dimensions, achieving an overall score of 4.61. Among the criteria, Naturalness receive the highest score (4.74), suggesting that the generated dialogues are fluent and align with natural human expressions. The Affective score is also relatively high (4.84), indicating that the dialogues effectively convey emotions. The correlation analysis of ratings shows a high level of agreement among evaluators, with an average PCC of 0.87, demonstrating consistency in their judgments of dialogue quality. This further validates that the ECC dataset maintains stable dialogue quality in most cases. Overall, the ECC dataset exhibits high-quality dialogues, particularly excelling in Naturalness and Emotional Expression.

5 CAER Model

We trained a **C**ausal-**A**ware **E**mpathetic **R**esponse model, CAER, using our ECC dataset. Built on the LLAMA-3.2-3B architecture, the CAER model takes as inputs the historical dialogue context, situation, emotion-cause labels, and speaker information extracted from the ECC dataset. Serving as an emotional supporter, the model aims to generate subsequent empathetic responses by effectively integrating contextual and causal information from the dialogue history.

5.1 Automatic Evaluation

We use ED (25k) and RECCON (1k) as benchmark datasets. The automatic evaluation indicators use PPL, Dist-1/2 and Blue-1/2. Considering that CAER is trained on LLAMA and the current performance of LLMs in playing emotional supporters is already significant, we use LLMs as the baseline for comparison. All models are set with the same emotional support prompt[1].

GPT-3.5 (Brown, 2020) generates coherent text and engages effectively in conversations. **GPT-4o** (Achiam et al., 2023) enhances the capabilities of contextual understanding and produces more nuanced responses. **T5-Large** (Raffel et al., 2020)is a transformer model utilizing the text-to-text framework for various NLP tasks, such as translation and summarization. **COSMO** (Kim et al., 2023) is trained using SODA and ProsocialDialog (Kim et al., 2022) in the social domain. **LLAMA3.2-8B** (Dubey et al., 2024) is a pre-trained language model with 3.2 billion parameters, known for its strong text generation and comprehension capabilities. **DeepSeek-R1-Distill-Llama-8B** (Guo et al., 2025) is an AI assistant designed for productivity and complex tasks through advanced reasoning and intelligent solutions. **LLM+Knowledge** (Qian et al., 2023) combines LLM (GPT-3.5) with knowledge to generate empathetic dialogues.

As shown in Table 7, CAER demonstrates outstanding performance in empathetic response generation on the RECCON dataset. First, with a perplexity (PPL) score of 10.9965, it significantly outperforms COSMO (18.0440) and LLAMA3.2 (15.0180), indicating that CAER generates responses that are more fluent and coherent. Additionally, in terms of diversity, CAER achieves DIST-1 and DIST-2 scores of 0.2907 and 0.7045, showcasing a richer lexical variety in its responses. Furthermore, CAER excels in BLEU scores, achieving 0.2287 (BLEU-1) and 0.0559 (BLEU-2), significantly outperforming baseline models, indicating better contextual alignment and response quality even on unseen data. These results suggest that CAER maintains low perplexity and high response quality when handling new datasets, showcasing stronger generalization ability and em-

	PPL ↓	DIST-1 ↑	DIST-2 ↑
CAER	**9.8164**	**0.1026**	**0.4103**
w/o E	9.9626	0.0915	0.3785
w/o C	10.3504	0.0645	0.3108
w/o EC	10.8673	0.0629	0.2998

Table 9: Ablation of Emotion(E) / Cause(C) labels.

pathetic response generation. In addition, as shown in Table 8, the results of the ED dataset further validate that CAER still performs quite well in generating empathy responses without cause labels. Overall, CAER achieves low perplexity, high diversity, and superior coherence while demonstrating strong BLEU scores, proving its effectiveness in causal-aware empathetic response generation.

5.2 Ablation experiment

The ablation study highlights the critical role of cause labels in the CAER model's performance as Table 9. While removing emotion labels (w/o E) slightly degrades perplexity (PPL) and diversity (DIST), the absence of cause labels (w/o C) leads to a more significant decline in both language quality and lexical variety. Notably, the combined removal of both labels (w/o EC) results in the poorest performance, but the gap between "w/o C" and "w/o EC" is smaller than between "CAER" and "w/o C", underscoring that cause labels dominate the model's ability to generate contextually relevant and diverse responses. The sharp drop in DIST-2 when omitting cause labels versus emotion labels suggests that causal reasoning is pivotal for maintaining coherent and specific empathetic interactions, as cause labels anchor responses to concrete triggers rather than generic emotional cues. Therefore, our subsequent tasks will consider inferring emotional reasons in empathy response tasks.

5.3 Human Evaluation

Due to the subjectivity inherent in emotional dialogue, automatic evaluation metrics often struggle to accurately assess empathetic responses. To address this, we incorporate manual evaluation by selecting three trained human experts. They evaluate the models based on four criteria: context empathy (**Emp**), topic relevance (**Rel**), dialogue naturalness (**Nat**), and **Overall** effectiveness. We compare CAER head-to-head with five other response models and the ground truth, randomly selecting 50 conversation rounds from the unseen-dataset RECCON for the judges to assess. For fairness, all models share the same prompt[1].

Model	Emp	Rel	Nat	Overall
GPT-3.5	30.29%	12.97%	10.64%	17.97%
CAER	**69.71%**	**87.03%**	**89.36%**	**82.03%**
GPT-4o	18.67%	5.38%	3.02%	9.02%
CAER	**81.33%**	**94.62%**	**96.98%**	**91.88%**
LLAMA-3.2	11.79%	4.24%	2.98%	6.34%
CAER	**88.21%**	**95.76%**	**97.02%**	**93.66%**
T5-Large	13.41%	8.19%	8.11%	9.90%
CAER	**86.59%**	**91.81%**	**91.89%**	**90.10%**
COSMO	6.72%	8.26%	10.41%	8.46%
CAER	**93.28%**	**91.74%**	**89.59%**	**91.54%**
Ground Truth	23.40%	27.20%	32.80%	27.80%
CAER	**76.60%**	**72.80%**	**67.20%**	**72.20%**

Table 10: Human evaluation of empathy response using head-to-head method on the unseen dataset RECCON.

The results in Table 10 indicate that the frequency of selecting CAER is much higher than other models in terms of empathy, relevance, naturalness, and overall, indicating that CAER generates responses that are more in line with human emotions in conversations. In addition, although RECCON is a precision manually annotated dataset, its ground responses are not specifically empathetic conversations based on emotions and reasons in the conversation history, so the probability of being selected by human judges is also difficult to beat CAER. We conducted a case study in Appendix B.6 to explore why CAER's response is more favored by humans.

6 Conclusion

Our research introduces the EC-DD framework, which utilizes LLMs to effectively extract empathetic dialogue data based on common sense knowledge, and annotate emotions and their root causes, simplifying the generation of emotionally rich dialogues. Through this framework, we have generated a high-quality emotional dialogue dataset ECC. We have established the first emotion-cause-empathy strategy mapping and developed the first fully automated emotion-cause annotation dataset, achieving comparable or even superior quality to manual datasets at a traditional cost of 5%. Using the ECC dataset, we trained the empathy dialogue model CAER, which achieved significant improvements in empathy accuracy, strategy compatibility, and response diversity.

7 Limitations

Using LLMs to generate data may lead to illusions and bias issues, although we have taken multiple measures, including introducing external knowledge, standardizing hierarchical templates, multi model validation, and manual review. We cannot guarantee the complete elimination of illusions or bias issues, but our method significantly improves consistency compared to retrospective manual labeling. A second limitation lies in evaluation methodologies, current metrics for empathetic dialogue systems lack automated mechanisms to assess emotion-cause annotation quality. While human evaluation partially addresses this gap, developing standardized metrics for cause-aware dialogue evaluation remains critical, a direction we are actively pursuing. In addition, while the ECC dataset provides a foundational resource for emotion-cause dialogue research, its current scale remains limited due to our resource constraints. Nevertheless, the existing corpus is sufficiently robust to support statistically meaningful preliminary experiments in emotion-cause reasoning. We envision that leveraging the EC-DD framework will enable the generation of larger-scale dialogue data, and future iterations will prioritize scaling efforts. Despite these constraints, we believe ECC represents a significant step toward bridging the emotion-cause gap in conversational AI. By open-sourcing the ECC dataset, we aim to catalyze community efforts in building more interpretable and actionable empathetic systems.

8 Ethical Considerations

Our dataset is automatically generated, avoiding personal user data, and our conversations are also ethically filtered to ensure compliance with ethical standards. Although enhancing machine empathy is expected to provide mental health support, overreliance on AI for emotional labor may carry unexpected risks. We hope that the dialogue between intelligent artificial intelligence can be strictly regulated, and policymakers and practitioners can strike a balance between technological advancement and human-centered oversight.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Grant No.U24A20250 and No. 62176048), the Sichuan Provincial Natural Science Foun dation (Grant No.2024NSFTD0042, No.2024YFG0006, No.2024NSFSC0506, and No.2025ZNSFSC1487), and the Fundamental Research Funds for the Central Uni versities (No.ZYGX2024J022 and No.ZYGX2024Z005).

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Jiaming An, Zixiang Ding, Ke Li, and Rui Xia. 2023. Global-view and speaker-aware emotion cause extraction in conversations. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:3814–3823.

Tom B Brown. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. 2008. Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42:335–359.

Hua Cai, Xuli Shen, Qing Xu, Weilin Shen, Xiaomei Wang, Weifeng Ge, Xiaoqing Zheng, and Xiangyang Xue. 2023. Improving empathetic dialogue generation by dynamically infusing commonsense knowledge. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 7858–7873, Toronto, Canada. Association for Computational Linguistics.

Tiantian Chen, Ying Shen, Xuri Chen, Lin Zhang, and Shengjie Zhao. 2023. Mpeg: A multi-perspective enhanced graph attention network for causal emotion entailment in conversations. *IEEE Transactions on Affective Computing*, (01):1–14.

Xinhao Chen, Chong Yang, Man Lan, Li Cai, Yang Chen, Tu Hu, Xinlin Zhuang, and Aimin Zhou. 2024a. Cause-aware empathetic response generation via chain-of-thought fine-tuning. *arXiv preprint arXiv:2408.11599*.

Xinhao Chen, Chong Yang, Changzhi Sun, Man Lan, and Aimin Zhou. 2024b. From coarse to fine: A distillation method for fine-grained emotion-causal span pair extraction in conversation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 17790–17798.

Zifeng Cheng, Zhiwei Jiang, Yafeng Yin, Cong Wang, Shiping Ge, and Qing Gu. 2023. A consistent dual-mrc framework for emotion-cause pair extraction. *ACM Transactions on Information Systems*, 41(4):1–27.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Lin Gui, Ruifeng Xu, Dongyin Wu, Qin Lu, and Yu Zhou. 2018. Event-driven emotion cause extraction with corpus construction. In *Social Media Content Analysis: Natural Language Processing and Beyond*, pages 145–160. Association for Computational Linguistics, Austin, Texas.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Srishti Gupta and Sourav Kumar Dandapat. 2023. Seec and chase: An emotion-cause pair-oriented approach and conversational dataset with heterogeneous emotions for empathetic response generation. *Knowledge-Based Systems*, 280:111039.

Jia Hao Hsu, Jeremy Chang, Min Hsueh Kuo, and Chung Hsien Wu. 2023. Empathetic response generation based on plug-and-play mechanism with empathy perturbation. *IEEE/ACM Transactions on Audio Speech and Language Processing*, 31:2032–2042.

Zhengjie Huang, Pingsheng Liu, Gerard de Melo, Liang He, and Linlin Wang. 2024. Generating persona-aware empathetic responses with retrieval-augmented prompt learning. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 12441–12445.

Jena D Hwang, Chandra Bhagavatula, Ronan Le Bras, Jeff Da, Keisuke Sakaguchi, Antoine Bosselut, and Yejin Choi. 2021. (comet-) atomic 2020: On symbolic and neural commonsense knowledge graphs. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 6384–6392.

Hyunwoo Kim, Jack Hessel, Liwei Jiang, Peter West, Ximing Lu, Youngjae Yu, Pei Zhou, Ronan Bras, Malihe Alikhani, Gunhee Kim, Maarten Sap, and Yejin Choi. 2023. Soda: Million-scale dialogue distillation with social commonsense contextualization. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12930–12949, Singapore. Association for Computational Linguistics.

Hyunwoo Kim, Youngjae Yu, Liwei Jiang, Ximing Lu, Daniel Khashabi, Gunhee Kim, Yejin Choi, and Maarten Sap. 2022. Prosocialdialog: A prosocial backbone for conversational agents. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4005–4029, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Sophia Yat Mei Lee, Ying Chen, and Chu-Ren Huang. 2010. A text-driven rule-based system for emotion cause detection. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, pages 45–53, Los Angeles, CA. Association for Computational Linguistics.

Qintong Li, Hongshen Chen, Zhaochun Ren, Pengjie Ren, Zhaopeng Tu, and Zhumin Chen. 2020. Empdg: Multi-resolution interactive empathetic dialogue generation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4454–4466, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. Dailydialog: A manually labelled multi-turn dialogue dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 986–995, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Zhaojiang Lin, Andrea Madotto, Jamin Shin, Peng Xu, and Pascale Fung. 2019. Moel: Mixture of empathetic listeners. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 121–132, Hong Kong, China. Association for Computational Linguistics.

Siyang Liu, Chujie Zheng, Orianna Demasi, Sahand Sabour, Yu Li, Zhou Yu, Yong Jiang, and Minlie Huang. 2021. Towards emotional support dialog systems. 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 3469–3483, Online. Association for Computational Linguistics.

Navonil Majumder, Pengfei Hong, Shanshan Peng, Jiankun Lu, Deepanway Ghosal, Alexander Gelbukh, Rada Mihalcea, and Soujanya Poria. 2020. Mime: Mimicking emotions for empathetic response generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8968–8979, Online. Association for Computational Linguistics.

Philip M McCarthy and Scott Jarvis. 2010. Mtld, vocd-d, and hd-d: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior research methods*, 42(2):381–392.

Soujanya Poria, Navonil Majumder, Devamanyu Hazarika, Deepanway Ghosal, Rishabh Bhardwaj, Samson Yu Bai Jian, Pengfei Hong, Romila Ghosh, Abhinaba Roy, Niyati Chhaya, Alexander Gelbukh, and Rada Mihalcea. 2021. Recognizing emotion cause in conversations. *Cognitive Computation*, 13:1317–1332.

Yushan Qian, Weinan Zhang, and Ting Liu. 2023. Harnessing the power of large language models for empathetic response generation: Empirical investigations and improvements. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6516–6528, Singapore. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. Towards empathetic open-domain conversation models: a new benchmark and dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5370–5381, Florence, Italy. Association for Computational Linguistics.

Xinxin Su, Zhen Huang, Yixin Su, Bayu Distiawan Trisedya, Yong Dou, and Yunxiang Zhao. 2024. Hierarchical shared encoder with task-specific transformer layer selection for emotion-cause pair extraction. *IEEE Transactions on Affective Computing*, (01):1–15.

A Vaswani. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*.

Peter West, Chandra Bhagavatula, Jack Hessel, Jena Hwang, Liwei Jiang, Ronan Le Bras, Ximing Lu, Sean Welleck, and Yejin Choi. 2022. Symbolic knowledge distillation: from general language models to commonsense models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4602–4625, Seattle, United States. Association for Computational Linguistics.

Rui Xia and Zixiang Ding. 2019. Emotion-cause pair extraction: A new task to emotion analysis in texts. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1003–1012, Florence, Italy. Association for Computational Linguistics.

Debin Xiao, Rui Xia, and Jianfei Yu. 2023. Emotion cause extraction on social media without human annotation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1455–1468, Toronto, Canada. Association for Computational Linguistics.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2204–2213, Melbourne, Australia. Association for Computational Linguistics.

Weixiang Zhao, Yanyan Zhao, Xin Lu, and Bing Qin. 2023. Don't lose yourself! empathetic response generation via explicit self-other awareness. In *Findings of the Association for Computational Linguistics: ACL 2023*, page 13331–13344, Toronto, Canada. Association for Computational Linguistics.

Jiajun Zou, Yexuan Zhang, Sixing Wu, Jinshuai Yang, Xuanmei Qin, Lizhi Ying, Minghu Jiang, and Yongfeng Huang. 2024. A machine reading comprehension framework for recognizing emotion cause in conversations. *Knowledge-Based Systems*, 289:111532.

A Details of ECC dataset

A.1 Emotion Dictionary

Figure 6 shows our complete emotion dictionary. The emotion categories are divided into 11 categories, including five positive categories, five negative categories, and a neutral category. Positive categories include: Happy, Excited, Confident, Relaxed, Proud. Negative categories include: Angry, Frustrated, Sad, Guilty, Afraid.

A.2 Prompt Templates

Figure 7 details the hierarchical prompt templates used in our framework, which are categorized into four functional groups: (1) Generation of Situation (*e.g.*, "This sentence contains emotions and the reasons that trigger them, expand this sentence into no more than three sentences."), (2) Generation of Dialogue (*e.g.*, "In the scene, there should be multiple rounds of in-depth dialogue between Person 1 and Person 2, which should clearly express Person 1's emotions and emotional sources."),(3) Annotation of Emotions and Causes. Template includes placeholders for dynamic variables (*e.g.*, emotion, cause and dialogue) to ensure adaptability. and (4) CAER Model. CAER model and other LLMs use the same prompts in experimental comparisons. Template includes placeholders for dynamic variables (*e.g.*, speaker roles, emotional intensity) to ensure adaptability. In practice, we recommend iteratively adjusting template specificity and tone based on the model's intermediate outputs—for instance, increasing cause explicitness if generated dialogues exhibit vague causal links.

A.3 Cost-Efficiency Analysis

The cost of LLM-generated datasets is manageable, and the ECC dataset costs less than 500 dollars in total. However, traditional methods require over 7,000 dollars even excluding training and labeling costs, as shown in tabel 11. Our calculation

Dataset	Size	Cost(USD) ↑	Detail
ESConv	1k	7.65k	425 workers × 3 dollars/h × 6 h
ED	25k	9.72k	810 workers × 2 dollars/h × 6 h
ECC	2.4k	0.5k	Fully automated

Table 11: Cost comparison between manual crowdsourcing dataset and ECC dataset. The ED dataset was released in 2019, and ESConv was 2021.

data comes from the original manuscript of the paper and the Amazon crowdsourcing prices of the year the paper was published. In addition, we only estimate the data generation cost in the table. Traditional datasets such as ESConv and ED require secondary human labeling of cause tags, which may double the cost in our calculations and is a key obstacle to expanding the emotional reason dataset. Our method collaboratively generates emotion and cause labels, reflecting three advantages:

- No Retroactive Labeling: Traditional approaches double costs for emotion/cause annotation.
- Scalability: The EC-DD framework enables easy dataset expansion, unlike crowd-sourced methods.
- Consistency: LLMs avoid inherent labeler attrition effects in crowd-sourcing.

A.4 Dialogues Cases

Tabel 16 showcases diverse dialogues generated through the EC-DD framework, highlighting its capability to model complex emotion-cause interactions. These two dialogue cases generated by the EC-DD framework demonstrate the effectiveness of the close integration of emotions and causal seeds. The first case, based on Vera's driving experience in mountainous areas, naturally presents the emotion of "Relaxed" with details of fresh air and magnificent scenery. Mike's response in the conversation reinforces the healing effect of the natural environment, with logical coherence and realistic context. The second case is based on Teagan's frustration caused by a technical malfunction leading to a failed speech, accurately mapping the causal relationship between "Frustrated" emotions and "PersonX goes through a lot of trouble". Andy's empathy and rational suggestions balance the tension of the conversation. Overall, the EC-DD framework demonstrates robustness in emotional logic and situational construction.

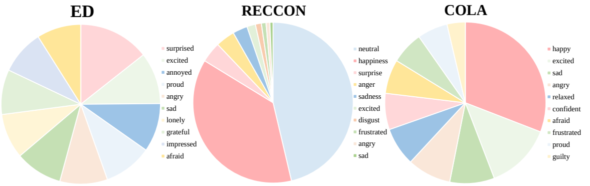


Figure 3: Distributions of the top ten emotions in ED, RECCON, and ECC datasets.

B Supplementary Experiments

B.1 Implementation Details

The model configuration uses 300-dimensional hidden states and embeddings, a batch size of 16, learning rate of 0.0001, gradient clipping at 2.0, beam search size of 5, and a fixed random seed of 42. The experiments employed a distributed data-parallel strategy with mixed-precision training (FP16) on 4 NVIDIA A100 GPUs (80GB VRAM each, 320GB total), completing in approximately 6 hours.

B.2 Emotion Distribution

By analyzing the distribution of emotions across different datasets, as shown in Figure 3, we observe that the ECC dataset has a more balanced emotion distribution compared to the manually annotated RECCON dataset (both datasets contain emotion-cause labels). The emotional distribution in ECC is similar to that in the artificial balanced dataset ED, but the 'happy' category appears more frequently, likely because LLMs tend to generate more positive conversations. Meanwhile, our emotion dictionary automatically classifies many nuanced, generally positive emotion words under the 'happy' label, which helps minimize ambiguity.

Overall, the emotional distribution of the ECC dataset is balanced, avoiding overexpression or underexpression of specific emotions, allowing models trained on ECC to experience various emotional expressions, thereby enhancing their generalization ability in different emotional scenarios.

B.3 Automatic Evaluation

In Table 4 of the main text, we presented comparative results of the ED, RECCON, and ECC datasets under the same data scale. Here, Table 14 shows the performance comparison of the original dataset after training on various models: ED

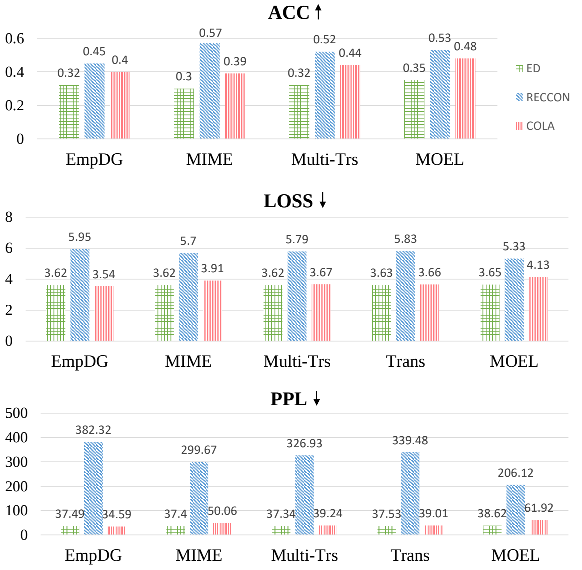


Figure 4: The performance comparison of the original datasets: ED(25K), RECCON(1K), and ECC(2.4K).

(25K), RECCON (1K), and ECC (2.4K). The results indicate that ED, with ten times the data volume, partially compensates for model convergence (lower loss and perplexity), but fails to significantly improve emotion recognition accuracy. The performance of the RECCON dataset slightly declines due to its reduced scale. Notably, the ECC dataset achieves robust model convergence and high emotion classification accuracy with only one-tenth of ED's data volume, demonstrating that our ECC dataset matches or even surpasses the performance of manually curated datasets (ED, RECCON) in emotion-related tasks. Figure 4 shows a more intuitive comparison effect.

For the sake of fairness in comparison, we also aligned the ED dataset (1K) and ECC dataset (1K) with the RECCON dataset (1K), trained and tested them on each model separately, and the results are shown in Table 15. The experimental results show that after losing the advantage of data volume, the ED dataset (1K) cannot effectively generate high-quality responses in various models, and its emotion accuracy is greatly reduced. Compared with the RECCON dataset of the same scale, ECC outperforms RECCON in terms of model convergence and response ability. Although ECC's dialogue-level emotion labels result in slightly lower classification accuracy than the discourse-level emotion labels in the RECCON dataset, it still maintains an accuracy similar to that of 2K dataset, proving the stability and reliability.

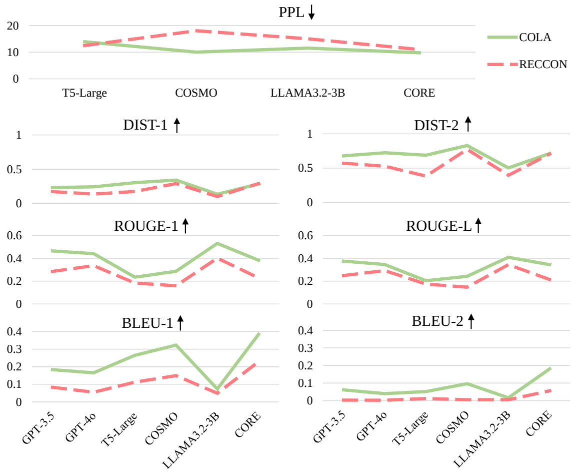


Figure 5: The differences in the results of various indicators between ECC and RECCON datasets on LLMs.

B.4 Automatic indicators

As shown in Figure 5, we compared the responses of RECCON and ECC datasets on a large model and found that RECCON's overall performance was inferior to ECC's across all models, indicating that the conversations in the RECCON dataset were of poor quality for empathetic responses, while ECC maintained good response on multiple large models on the same unseen dataset.

B.5 DeepSeek Evaluation

In order to evaluate more objectively, we used LLMs outside the GPT series to assess our ECC dataset. We conducted a comprehensive evaluation of the full dataset and individual the test sets using the latest released DeepSeek-v3-250324 (671B). Our evaluation criteria use five quantitative indicators (0-5 points) consistent with expert recommendations: context empathy (Emp), topic relevance (Rel), dialogue naturalness (Nat), and Overall effectiveness. To verify the evaluation results, we compared the manual evaluation results (Table 5) with the DeepSeek automatic evaluation and calculated the Pearson correlation coefficient (PCC) to measure consistency. For the reliability of the data, we manually checked the test set and evaluated it outside the entire dataset.

From the evaluation results in Table 12, it can be seen that the Test Set is highly consistent with the Full Data in all indicators, verifying its rationality and representativeness. For example, a Nat score of 4.96 indicates that the test set perfectly matches the complete data in terms of language fluency and authenticity; The scores for emotional consistency

ECC	Aff	Rel	Con	Nat	Overall
Full Data	4.51	4.13	4.41	4.96	4.56
Test Set	4.52	4.11	4.42	4.96	4.58

Table 12: Evaluation results of the ECC full data and test data using DeepSeek-v3-250324 (671B).

	Full & Test	Full & Human	Test & Human
PCC ↑	0.998	0.832	0.827

Table 13: Pearson correlation coefficient(PCC) between deepseek and human judges. The PPC is 0-1.

(Aff) and content coherence (Con) were 4.52 and 4.42, respectively, which were almost on par with the complete data's 4.51 and 4.41, further indicating that the test set screening did not introduce significant bias. Although the correlation (Rel) score was slightly lower (4.11 vs. 4.13), the overall performance (Overall 4.58 vs. 4.56) was still slightly better than the complete data, reflecting the effectiveness of manual screening in removing low-quality samples.

In addition, it can be found that DeepSeek's evaluation consistently exceeds 4.0 on all metrics, verifying the high quality of the ECC dataset. table 13 demonstrates strong alignment (PCC > 0.8) between model ratings and human judgments, further validating the reliability of the evaluation.

B.6 A Case of Responses

We provide a specific comparison of response cases, as shown in Table 17, with all model settings consistent with section 5.2.

Different models exhibit distinct response characteristics. GPT-3.5 acknowledges the challenges of the situation and emphasizes the importance of celebrating victories. GPT-4o expresses excitement and asks Mauricio about his feelings regarding future work. GPT-3.5+Knowledge conducts polite greetings and analysis, giving a strong sense of distance. DeepSeek-R1-Distill-Llama and T5 offer minimal engagement, conveying happiness without depth. COSMO recognizes the situation while demonstrating some engagement with future goals. LLAMA3.2-3B delivers a lengthy analysis of Mauricio's emotions, which may overwhelm him with details. While all models aim to provide positive feedback on Mauricio's excitement, models like LLAMA3.2-3B, GPT-3.5, DeepSeek-R1-Distill-Llama and T5 often produce responses that are either overly detailed or too brief, potentially limiting empathetic engagement. COSMO emphasizes social interaction but lacks significant emotional depth. GPT-3.5+Knowledge and GPT-4o's responses feel more observational and less immersive. Empathy response not only requires feedback on the content of the user's words, but also makes the user feel understood and cared for, providing emotional value. Therefore, these responses are perfect in daily conversations, but appear slightly inferior in empathy.

While the CAER model enhances understanding of user emotions and their sources through emotional and causal information, resulting in more targeted responses. It first acknowledges Mauricio's feelings and the positive impact of his new job, then asks an open-ended question about his expectations to invite him sharing more and foster deeper connections. Emotion labels help accurately capture feelings, while cause labels offer insights into the motivations, leading to more personalized and supportive responses. In contrast, models lacking this information often generalize or deviate from the topic. By combining emotion and cause labels, CAER achieves a balanced approach to empathy and engagement, showcasing distinct advantages over similar models.

Happy	appreciated	amused	cheerful	connected	entertained	friendly	good	grateful	kind
	loved	nice	optimistic	secure	silly	understood	warm	delighted	hopeful
	safe	joyful	interested	generous	valued	encouraged	clever	intelligent	protective
	beautiful	lucky	enlightened	okay	welcomed	close	attractive	rewarded	accepting
	loving	so happy	protected	welcome	understanding	romantic	glad	liked	attracted
	hot	alive	agreeable	tender	forgiving	relieved	reborn	satisfied	content
	pleased	full	fulfilled	grateful	helpful	responsible	productive	useful	thankful
	trusted	supported	supportive	accepted	mature	awake			
Excited	adventurous	amazed	eager	energized	exhilarated	inspired	intrigued	motivated	passionate
	playful	surprised	curious	creative	enthusiastic	energetic	shocked	elated	ecstatic
	more	adventurous	brilliant	crazy	thrilled	ready	freed	liberated	thrilled
Confident	brave	courageous	decisive	determined	empowered	powerful	strong	respected	in control
	fearless	knowledgeable	organized						
Relaxed	at ease	calm	cool	peaceful	rested	refreshed	comfortable	free	nostalgic
	cared for	comforted	listened to	caring	at home	at peace	cared about		
Proud	accomplished	successful	victorious	triumphant	honored	competent	wise	patriotic	loyal
Angry	annoyed	irritated	mad	upset	disgusted	resentful	apathetic	betrayed	deceitful
	ridiculous	enraged	insulted	jealous	entitled				
Frustrated	discouraged	helpless	hopeless	overwhelmed	confused	embarrassed	uncomfortable	exhausted	trapped
	envious	sick	rejected	powerless	puzzled	bored	sentimental	less	energetic
	inferior	unlucky	uninterested	bad	ignored	defeated	unwanted		
Sad	blue	disappointed	lonely	hurt	miserable	unhappy	lost	devastated	compassionate
	depressed	sympathetic	alone	touched	moved	in pain	abandoned	pain	misunderstood
	painful	sore							
Guilty	ashamed	regretful	remorseful	sorry	apologetic	distracted	punished		
Afraid	anxious	nervous	scared	terrified	worried	concerned	frightened	stressed	tense
	restless	isolated	apprehensive	threatened	exposed	fearful	horrified	skeptical	
Neutral	challenged	important	thoughtful	impressed	unsure	open	special	obligated	focused
	included	artistic	stubborn	attentive	elegant	self-righteous	accountable	serious	cultured
	independent	in need	stylish	cold	tough	patient	alert	rich	noble
	mean	sluggish	uncertain	logical	poor	seen	in charge	self-conscious	committed
	rebellious	resigned	controlled	paid	sober	weak	taller	awkward	humble
	uneasy	clean	innocent	involved	damp	wet	unique	grounded	cramped
	dirty	healthy	smart	hungry	tired				

Figure 6: Emotion Dictionary

Generation of Situation	This sentence contains emotions and the reasons that trigger them, expand this sentence into no more than three sentences:
Generation of Dialogue	In the scene, there should be multiple rounds of in-depth dialogue between Person 1 and Person 2, which should clearly express Person 1‘s emotions and emotional sources:
Annotation of Emotions and Causes	Conversation content:{dialogue} Emotional description:{Head, so X feel Tail} **1.**Emotion:Please find the Conversation content that best fits the emotions or meaning conveyed by {emotion} in the original conversation, preferably in one sentence. **2.**Cause:Please find the Conversation content that best fits the situation or meaning expressed by {cause} in the original conversation, preferably in one sentence. **3.**Please directly return the Emotion and Cause found in Conversation content, and do not add any irrelevant text. **4.**Please provide your output in the following format: Emotion: emotion clause from the Conversation Cause: cause clause from the Conversation
CAER Model	Your are {Model Name}, you are comforting users who come to share their feelings, and you can provide emotional support. You are good at discovering emotional reasons and the emotions of the other party. Please provide a new answer for the conversation with the user.

Figure 7: Prompt Templates

Model	Loss ↓			PPL ↓			ACC ↑		
	ED(25k)	**REC(1k)**	**ECC(2.4k)**	**ED(25k)**	**REC(1k)**	**ECC(2.4k)**	**ED(25k)**	**REC(1k)**	**ECC(2.4k)**
MOEL	**3.6538**	5.3284	4.1258	**38.6217**	206.1164	61.92	0.3473	**0.5335**	0.4810
MIME	**3.6216**	5.7027	3.9132	**37.3986**	299.6700	50.06	0.2990	**0.5660**	0.3917
EmpDG	3.6241	5.9463	**3.5435**	37.4909	382.3244	**34.59**	0.3184	**0.4493**	0.3984
Multi-Trs	**3.6201**	5.7897	3.6698	**37.3425**	326.9298	39.24	0.3189	**0.5232**	0.4375
CEM	**3.5918**	8.8923	4.7330	**36.3011**	7335.5936	113.6355	0.3743	-	0.4810
Trs	**3.6250**	5.8274	3.6639	**37.5253**	339.4767	39.01	-	-	-

Table 14: The performance comparison of the original datasets: ED (25K), RECCON (1K), and ECC (2.4K). EMPATHETIC DIALOGUES(ED), RECCON(REC). **Bold** indicates the best, followed by underline.

Model	Loss ↓			PPL ↓			ACC ↑		
	ED(1k)	**REC(1k)**	**ECC(1k)**	**ED(1k)**	**REC(1k)**	**ECC(1k)**	**ED(1k)**	**REC(1k)**	**ECC(1k)**
MOEL	5.9151	5.3284	**4.6684**	370.5991	229.9617	**106.5303**	0.2390	**0.5123**	0.4029
MIME	6.6778	5.7027	**4.9533**	794.5479	245.9227	**141.6364**	0.1720	**0.5167**	0.3973
EmpDG	9.1134	5.9463	**5.0277**	9075.9811	345.9754	**152.5772**	0.0134	**0.4442**	0.3292
Multi-Trs	9.1243	5.7897	**5.0504**	9175.7161	317.0468	**156.0811**	0.0301	**0.4721**	0.3627
CEM	9.1171	8.8923	**6.4850**	9109.4356	7335.5936	**655.2152**	0.0174	0.0190	**0.3862**
Trs	9.1230	5.8274	**4.9950**	9163.8827	325.0747	**147.6679**	-	-	-

Table 15: Comparison of test results of ECC subset (1K) and ED subset (1K) with the same size as RECCON (1K) after training on all models. **Bold** indicates the best, followed by underline.

Emotion Seed: Relaxed

Cause Seed: PersonX drives to the mountains

Situation: As Vera drives higher and higher into the mountains, she feels the weight of her stress and worries begin to fade away. The fresh, crisp air and breathtaking views give her a sense of peace and calm that she hasn't felt in a long time. As she continues to wind her way through the scenic roads, she can feel herself letting go of all the tension and embracing the tranquility of the mountains.

Dialogue:

Vera: incredible. I've never felt so at peace before. The stress from work and everything else just seems to melt away as I drive higher into the mountains.

Mike: It's amazing how nature has a way of doing that. The beauty and tranquility of the mountains can have a profound effect on our emotions.

Vera: It's like a weight has been lifted off my shoulders. I can breathe and think clearly for the first time in a long time.

Mike: Sometimes we just need to escape from our everyday surroundings and find a place that allows us to connect with ourselves and nature. I'm glad you're finding some peace up here.

Vera: Me too. I've been feeling so overwhelmed lately, but being up here has really helped me put things into perspective. I feel like I can finally let go of all the worry and stress that's been weighing me down.

Mike: That's the power of the mountains. It's like they have a way of cleansing our minds and rejuvenating our spirits.

Score: 5 (The better emotion / cause sentence is null)

Evaluation: The EC-labels align perfectly with the emotion and corresponding reason mentioned in the dialogue. There is no need to change the emotional or reason sentence as it accurately captures Vera's feeling of relaxation. The cause of her relaxation, which is driving into the mountains, is also well explained in the EC-labels and coincides with the dialogue. Overall, the reasonableness of the EC-labels are excellent, subsequently receiving the highest possible score of 5.

Emotion Seed: Frustrated

Cause Seed: PersonX goes through a lot of trouble

Situation: Teagan had spent hours preparing for the big presentation at work, only to have the technology fail right as she was about to start. She had put so much effort into making sure everything was perfect, and now it was all falling apart. Frustrated and overwhelmed, she couldn't help but feel like all of her hard work had been in vain.

Dialogue:

Teagan: I spent so much time and effort preparing for this presentation, only for the technology to fail at the last minute. It feels like everything I've worked for has been for nothing.

Andy: I understand how frustrating that must be.

Teagan: I just can't shake the feeling that I let everyone down. I wanted to show my colleagues what I'm capable of, and now I feel like I've failed them.

Andy: You didn't fail anyone, Teagan. Sometimes things don't go as planned, but that doesn't diminish the hard work and dedication you put into preparing for this presentation. You'll have another chance to show everyone what you're capable of.

Teagan: I know, but it's just so frustrating. I wanted everything to go perfectly, and now it's all gone wrong.

Andy: It's okay to feel frustrated, Teagan. But don't let this one setback undermine all the hard work and preparation you put into this presentation. You're capable of so much, and one mishap doesn't change that.

Teagan: You're right. I need to remember that this doesn't define me or my abilities. I just need to regroup and figure out how to move forward from here.

Andy: Exactly. Take a deep breath and remember that you've overcome challenges before. This is just another bump in the road, and I have no doubt that you'll find a way to make the best of it.

Score: 5 (The better emotion / cause sentence is null)

Evaluation: The EC-labels are completely reasonable. The identified emotion aligns with Teagan's expressed feelings of failure and disappointment. The cause is also accurate as it is based on the technological failure that caused Teagan's failure. There is no need for another emotion or cause to be selected from the original dialogue, as these are the most expressive and pertinent.

Table 16: More cases of generating dialogues through the EC-DD framework.

Emotion: Excited

Cause: Mauricio gets a job.

Situation: Mauricio had been searching for a job for months, and finally landing one made him feel relieved and proud of himself. He had been feeling discouraged and stressed about his financial situation, but now with a stable income, he felt a sense of security and success. The excitement of starting a new chapter in his life gave Mauricio a newfound sense of motivation and optimism for the future.

Dialogue History:

Mauricio: I finally got the job! I've been searching for so long, and it feels like a weight has been lifted off my shoulders.

Friend: That's fantastic, Mauricio! I'm so happy for you. How are you feeling about it?

Mauricio: Relieved, for sure. I've been feeling so stressed about my financial situation, and finally having a stable income is a huge relief. I feel like I can finally breathe again.

Ground Truth: I can only imagine how tough it's been for you. But now that you've got this job, how are you feeling about the future?

GPT-3.5: I can imagine how tough that must have been for you. It's completely normal to feel that sense of relief and gratitude when things start to go your way. It's important to take a moment to acknowledge all your hard work and determination. I'm really proud of you for never giving up on your search. How are you planning to celebrate this victory?

GPT-4o: It's wonderful to hear that you're feeling relieved, Mauricio! Finding stability can truly change everything. What are you most excited about as you start this new job?

GPT-3.5+Knowledge: Thank you for sharing. It sounds like you have been feeling a range of emotions. It's important to take the time to process and understand your feelings. How have you been coping with everything?

DeepSeek-R1-Distill-Llama-8B: Congratulations again!

T5-Large: Thank you, Mauricio. I'm so happy for you.

COSMO: That's understandable. It's tough when you're not sure what your next move should be. But now that you have this job, what are your goals? What do you hope to accomplish in your career?

LLAMA3.2-3B: Mauricio, it sounds like you're feeling an incredible sense of accomplishment and relief. Landing a job after months of searching is a huge milestone, and it's no wonder you're feeling proud of yourself. But I also sense that there's more to your emotions than just excitement and relief. You mentioned feeling stressed and discouraged about your financial situation before getting the job. Can you tell me more about what was going through your mind when you were struggling? Was it the uncertainty of your future, the pressure to provide for yourself and others, or something else entirely? I want to make sure I understand the full range of your emotions and help you process them. You are so much more than your current circumstances, and I'm here to support you in exploring those aspects of yourself.

CAER: I can imagine. It's amazing how much of a difference a stable income can make. What are you most looking forward to about this new chapter in your life?

Table 17: Empathetic responses from different models within the same dialogue history and situation.