ECR-Chain: Advancing Generative Language Models to Better Emotion-Cause Reasoners through Reasoning Chains

Zhaopei Huang[1] , **Jinming Zhao**[2*] , **Qin Jin**[1*]

[1]Renmin University of China
[2]Independent Researcher

{huangzhaopei, qjin}@ruc.edu.cn, zhaojinming1@gmail.com

Abstract

Understanding the process of emotion generation is crucial for analyzing the causes behind emotions. Causal Emotion Entailment (CEE), an emotion-understanding task, aims to identify the causal utterances in a conversation that stimulate the emotions expressed in a target utterance. However, current works in CEE mainly focus on modeling semantic and emotional interactions in conversations, neglecting the exploration of the emotion-generation process. This hinders the models from deeply understanding emotions, restricting their ability to produce explainable predictions. In this work, inspired by the emotion generation process of "*stimulus-appraisal-emotion*" in the cognitive appraisal theory, we introduce a step-by-step reasoning method, **E**motion-**C**ause **R**easoning **Chain** (**ECR-Chain**), to infer the stimulus from the target emotional expressions in conversations. Specifically, we first introduce the ECR-Chain to Chat-GPT via few-shot prompting, which significantly improves its performance on the CEE task. We further propose an automated construction process to utilize ChatGPT in building an **ECR-Chain set**, which can enhance the reasoning abilities of smaller models through supervised training and assist the Vicuna-7B model in achieving state-of-the-art CEE performance. Moreover, our methods can enable these generative language models to effectively perform emotion-cause reasoning in an explainable manner. Our code, data and more details are at https:// github.com/hzp3517/ECR-Chain.

1 Introduction

Emotion plays an important role in human communication and there are many dialogue-related research works in the NLP field involving analysis and utilization of emotions. Emotion recognition in conversation, for example, aims to detect the emotional status of a conversation utterance based on the context [Majumder *et al.*, 2019; Ghosal *et al.*, 2019; Liu *et al.*, 2022]. Emotional response generation, on the
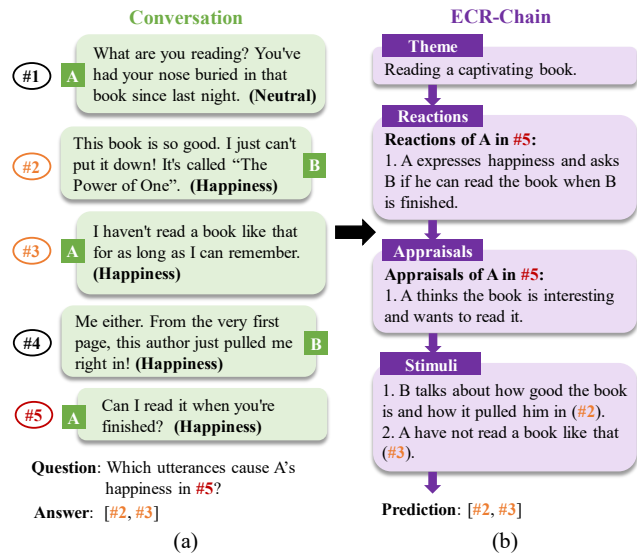
---

*Qin Jin and Jinming Zhao are corresponding authors.

Figure 1: An example of emotion-cause reasoning in a conversation. (**a**): The basic form of the CEE task, requires identifying the causal utterances (#2 & #3) that stimulate the speaker's emotion in the target utterance (#5). (**b**): Our proposed reasoning method, ECR-Chain, begins by summarizing the conversation theme, then describes the reactions and infers the appraisals of the target speaker, and finally deduces emotion-causes.

other hand, aims to incorporate emotional factors into dialog systems to enhance user satisfaction [Wei *et al.*, 2019; Zhong *et al.*, 2021]. However, these tasks lack attention to deeply understanding the causes behind emotions, which limits the application in broader areas like emotional support dialogue systems.

Therefore, a new task called "Causal Emotion Entailment" (CEE) together with a new benchmark dataset "REC-CON" [Poria *et al.*, 2021] has been proposed. As depicted in Figure 1, given a two-person conversation, the corresponding emotional state of the speaker of each sentence, and the target utterance, the goal of CEE is to identify which preceding utterances (including the target itself) are causal utterances containing the stimulus factors that lead to the emotion of the target utterance. Based on this benchmark, various methods that attempt to improve emotion-cause prediction have been explored. Modeling the speaker-aware semantic and emo-

tional interaction in the conversation context has been proven to be effective [Zhang *et al.*, 2022a] while introducing external social commonsense knowledge (CSK) [Hwang *et al.*, 2021] to the contextual interaction could also be helpful [Li *et al.*, 2022a; Zhao *et al.*, 2023a]. These methods treat the CEE task as a supervised classification problem, identifying whether each utterance is a causal utterance or not. Nevertheless, they neglect the exploration and learning of the emotion-generation process, which limits their ability to take a deeper understanding of emotions and restricts them from providing explainable rationales behind the emotion-cause. In contrast to these aforementioned works, we will not only explore to better address the CEE task but also extend it into a more challenging Explainable Emotion-Cause Reasoning task in this paper.

To overcome the limitations of previous models and realize explainable emotion-cause reasoning, we explore using generative language models to reason in our task. Considering that inferring the causes from emotional expression can be seen as a reverse process of emotion generation, we can refer to psychological descriptions of how emotions arise to determine the reasoning steps. According to the cognitive appraisal theory [Arnold, 1960; Ellsworth, 1991], the process of emotion generation can be summarized as "*stimulus-appraisal-emotion*". This sequence starts when a person attends to certain aspects of their environment or events, stimulating them to form interpretations or evaluations internally, which ultimately culminate in a specific emotional response. Therefore, in our task, considering our textual conversation scenarios, we can start by identifying behaviors related to the target emotion based on the target utterance and the global context. Subsequently, we can infer the speaker's inner thoughts based on their behaviors and finally deduce the stimuli that led to these thoughts. Based on the above analysis, we introduce an **E**motion-**C**ause **R**easoning **Chain** (**ECR-Chain**), formatted as "*theme → reaction → appraisal → stimulus*", which can be regarded as a step-by-step reasoning process as shown in Figure 1.

Specifically, inspired by the Chain-of-Thought (CoT) prompting method [Wei *et al.*, 2022], we incorporate our ECR-Chain into the in-context learning [Brown *et al.*, 2020] to guide large language model's reasoning step by step and achieve significant performance improvements under the few-shot CEE task setting. Having verified the effectiveness of this chain in emotion-cause reasoning, we further supplement the RECCON dataset with an automatically generated ECR-Chain set that can be used for supervised training. For example, we use the constructed ECR-Chain set for multi-task training on smaller generative language models and effectively enhance their emotional reasoning abilities to provide explainable reasoning paths and cause descriptions that were difficult to achieve with previous CEE models.

The main contributions of this work include: 1) We propose an emotion-cause reasoning method, ECR-Chain, and use it to guide the reasoning process of language models via CoT prompting. 2) We automatically construct an ECR-Chain set to help enhance the emotional reasoning ability of supervised models. 3) Extensive experimental results over various settings demonstrate the effectiveness of our method

for predicting emotion-cause utterances and performing explainable emotion-cause reasoning.

2 Related Works

2.1 Causal Emotion Entailment

Poria et al. [2021] address the task of recognizing emotion-causes in conversations. It includes two sub-tasks: Causal Emotion Entailment (CEE) and Causal Span Extraction (CSE). Recent related works mostly focus on the CEE task. Zhang et al. [2022a] propose a two-stream attention model (TSAM) to incorporate both speaker identities and emotional states into utterance features via contextual interaction modeling. Li et al. [2022a] introduce commonsense knowledge (CSK) related to speakers' emotional interactions and build a knowledge-enhanced conversation graph. Zhao et al. [2023a] further leverage event-centered CSK and social-interaction CSK to construct a knowledge-bridged causal interaction network. These above works all use RoBERTa [Liu *et al.*, 2019] to extract contextual utterance-level features and design specific graph structures to model various interactions between utterances. Although their model structures are suited for making classifications for each utterance in a conversation, the explainability of their models is limited. Recently, Zhao et al. [2023b] evaluate the performance of ChatGPT [1] on various emotion-related dialogue tasks, including the CEE. However, they simply instruct the ChatGPT to provide predictions based on the task input, lacking exploration of reasoning methods, which limits their performance. In our work, we explore employing a reasoning process to enhance the emotion-cause prediction performance of generative language models and enable them to have explainable reasoning capabilities.

2.2 Chain-of-Thought for Reasoning

Wei et al. [2022] first explore adding a few Chain-of-Thought (CoT) demonstrations via few-shot prompting to guide the large language model (LLM) thinking "step-by-step". Following their success, lots of works continue exploring improving the CoT prompting methods, including Self-Consistency [Wang *et al.*, 2022], DiVeRSe [Li *et al.*, 2022c], Auto-CoT [Zhang *et al.*, 2022b], and Active-Prompt [Diao *et al.*, 2023]. On the other hand, though smaller models cannot directly benefit from CoT prompting [Wei *et al.*, 2022], distilling the rationales generated by LLM into smaller models may still be helpful for them to solve the reasoning tasks [Ho *et al.*, 2023]. Li et al. [2022b] first prompt LLM to generate rationales for each question in the dataset. They then propose a multi-task training strategy to utilize the rationales as auxiliary training objectives for smaller models. Hsieh et al. [2023] also apply the multi-task method to distill the reasoning knowledge and improve the smaller models to outperform LLMs on their tasks. Inspired by these works, we first propose a CoT method customized for emotion-cause reasoning, and then improve the smaller models via multi-task training with the rationales automatically constructed by LLMs.

---

[1] https://openai.com/blog/chatgpt

3 Method

3.1 Task Definition

For the *Causal Emotion Entailment (CEE)* task, given a two-party conversation C, which is denoted as $C = [(u_1, k_1, e_1), (u_2, k_2, e_2), ..., (u_t, k_t, e_t)]$, where k_i, u_i and e_i are the speaker identity, the content, and the emotion of the i-th utterance respectively, its goal is to predict which utterances u_i $(i \leq t)$ in the conversation context are responsible for evoking the non-neutral emotion e_t of the target utterance u_t. We can reform it into a question-answer task, where the question is $Q = \{C, (u_t, k_t, e_t)\}$ and the answer is $A = \{u_i\}$. CEE task only evaluates the predicted answer A and does not care how the answer is derived.

To develop explainable models, we further extend the CEE task to an *Explainable Emotion-Cause Reasoning (Explainable ECR)* task, where given a question Q in the same format as mentioned above, the model is required to output the predicted answer A along with its rationale R. Explainable CEE may enhance the credibility of the prediction and benefit other emotion-related tasks.

3.2 Overview

Based on the cognitive appraisal theory in psychology, we introduce a reasoning process of "*theme → reaction → appraisal → stimulus*", which we call it "**E**motion-**C**ause **R**easoning **Chain** (**ECR-Chain**)". Tailored for emotion-cause reasoning in textual conversations, this serialized reasoning scheme is also consistent with the generation mode of generative language models, as they always predict the following words based on the preceding content. Therefore, we consider taking advantage of the ECR-Chain to improve the emotion-cause reasoning of generative language models.

For the two tasks defined in Section 3.1, we explore both the few-shot learning with large language models (LLMs) and the supervised learning with smaller generative language models [2]. For LLMs larger than 100B, inspired by the Chain-of-Thought (CoT) method [Wei *et al.*, 2022], we design a few-shot prompt to instruct them to reason step-by-step following our proposed ECR-Chain and then infer the final answer. But for smaller models, due to their limited parameter scales, the effect of directly applying the CoT prompting is not evident [Wei *et al.*, 2022]. So we consider improving their reasoning abilities through supervised training. As no such ECR-Chain data is available, we employ LLMs to automatically construct an ECR-Chain set to enable supervised training for smaller models to enhance their reasoning capability. Ultimately, through multi-task training, we build a model capable of performing the two tasks.

3.3 Few-shot Prompting with ECR-Chain

For generative language models, the most intuitive solution to a question-answer task is to generate an answer A directly based on the question Q, that is, $P(A|Q)$. However, when performing a reasoning task, the indirect correlation between Q and A may limit the performance. For LLMs, we

---

[2] Models with a size smaller than 100B are considered to lack evident CoT ability [Wei *et al.*, 2022].

Please understand the emotion-cause for the target utterance in a given conversation.
[Task Description]: ...
[Description of Reasoning Process]:
1. Output the "Theme": ...
2. List "Reactions" items of target utterance: ...
3. List "Appraisals" items of target utterance: ...
4. List "Stimuli" items with their corresponding utterance id: ...
5. Based on these "Stimuli" items, output the index number of the causal utterances in the form of Python list without any other content.

I will show you some examples:
[Example]: ...

— To be solved —
[Question]: ...

Figure 2: Illustration of our designed few-shot prompt for reasoning along the ECR-Chain. Detailed prompt examples are presented in the Appendix.

can employ the CoT prompting method to first guide them to generate a rationale R for the question Q, and then derive the final answer A based on R, forming in the format $P(R, A|Q)$. This approach often results in more accurate predictions. Considering that inferring emotion causes from emotional expressions also requires step-by-step reasoning, we design a few-shot prompt, denoted as <reasoning>, to guide the LLMs in reasoning along the ECR-Chain. Figure 2 is a diagram of our prompt, which mainly contains four parts:

- **Task Description**: Contains the goals of the emotion-cause reasoning task and the input and output formats.

- **Description of Reasoning Process**: Describe the meaning of each component in the ECR-Chain and specific requirements for each of the 5-step reasoning steps. We instruct the model to list the "*Reactions*", "*Appraisals*", "*Stimuli*" in items to make the reasoning process clearer. Besides, we require that each generated "*Stimulus*" item be followed by its corresponding utterance indexes so that the model can directly summarize all predicted utterance indexes and provide the final answer.

- **Example**: We manually write expected rationales for a few of the questions in the training set to serve as exemplars in few-shot prompting. Each exemplar consists of the conversation C, the target utterance u_t, and a manually written rationale r. Each utterance with its emotion and speaker label is simply written in this format: #[utterance index]: [speaker] ([emotion]): "[utterance content]".

- **Question**: Present the conversation and target utterance for each sample in the test data and instruct the model to perform 5-step reasoning.

3.4 Supervised Learning with ECR-Chain

Since the reasoning capability of smaller models is limited, it is hard to teach them to reason along our introduced ECR-Chain simply through few-shot prompting. However, recent
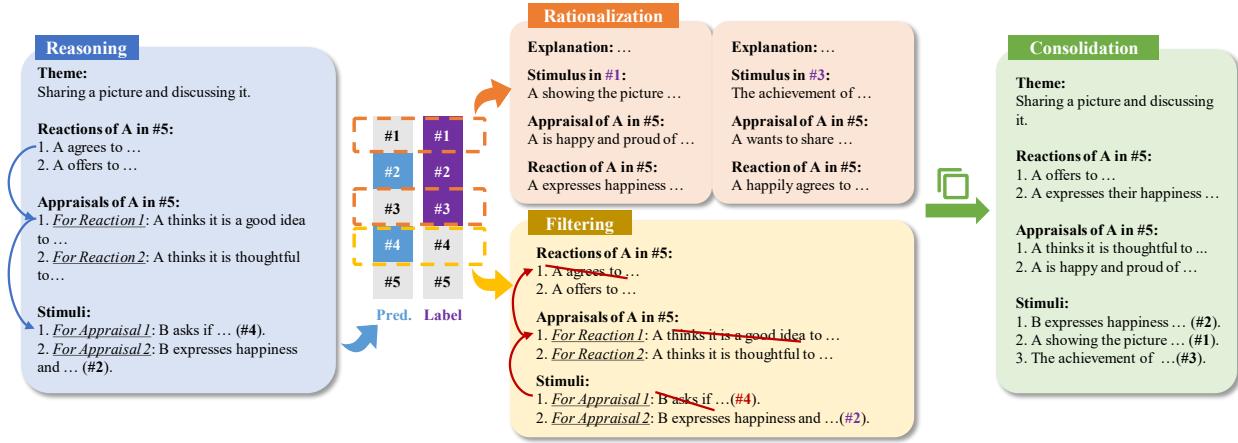
Figure 3: An example of the automated construction process of ECR-Chain set. **1) Reasoning**: Instruct the LLM with a <reasoning> prompt to generate a raw rationale without providing labels; **2) Filtering**: Delete chains which lead to incorrect answers (#4); **3) Rationalization**: For each golden answers not predicted (#1 & #3), supplement an additional chain by instructing the LLM with a <rationalization> prompt; **4) Consolidation**: Merge all retained and supplemented chains and consolidate semantically similar items by instructing the LLM with a <consolidation> prompt.

studies have shown that distilling the CoT rationales from LLM to smaller ones can help to enhance the reasoning capabilities of the smaller models on specific tasks [Li *et al.*, 2023; Ho *et al.*, 2023]. To this end, we utilize LLMs to automatically construct an ECR-Chain set, and perform supervised training with ECR-Chain on smaller models.

Automated Construction of ECR-Chain Set

The generation of reasoning chains can be approached in two ways: 1) reasoning through CoT prompting which derives an answer, or 2) generating explanations by rationalization prompting conditioned on labels [Marasović *et al.*, 2022]. The first approach may generate higher-quality rationales if the derived prediction is correct since the rationales have been proven to effectively guide the language model to output our expected answer. However, when the prediction is wrong, the reasoning approach may provide some misleading rationales. Zelikman et al. [2022] and Li et al. [2022b] propose to generate a new rationale by rationalization prompting to supplement for questions that can not be reasoned to a correct answer. However, our ECR-Chain rationales are more complex compared to their works, as our rationales may contain multiple items in the *Reaction*, *Appraisal*, and *Stimulus* part, and different stimulus items may lead to different causal utterance predictions. Therefore, we propose a 4-stage processing procedure to generate and further revise our rationales, which involves reasoning, filtering, rationalization, and consolidation, as illustrated in Figure 3.

a) Reasoning: We first execute a reasoning process using the <reasoning> prompt illustrated in Figure 2 and obtain raw rationales. Here we denote the *Reaction*, *Appraisal*, and *Stimulus* part of our rationals as n, l, s, respectively. Each of them may contain multiple items. Here, we require that for each appraisal item l_j generated, the index number of its corresponding reaction item n_i be specified. For each stimulus item s_k, its corresponding l_j needs to be specified similarly. From this, we can determine the corresponding relationship from n_i to l_j and to s_k, linking these three items to a chain,

as shown in Figure 3. Besides, we can find the set of all predicted answers a^{pred} corresponding to generated stimulus items s.

b) Filtering: Since the ground-truth label is unknown to the LLM in the reasoning stage, the raw rationales may include some chains that lead to wrong answers. With the help of the connections we built in the reasoning stage, we can remove an entire chain $\{n_i, l_j, s_k\}$ corresponding to a wrong predicted answer a_w^{pred}. We also remove those items in each part that could not lead to any subsequent items.

c) Rationalization: If there are some answers in the ground-truth label that do not appear in a^{pred}, we then need to supplement their corresponding rationales. We perform a rationalization process separately for each missing labeled answer a_m^{GT}. Intuitively, rationalization should be the opposite process of reasoning, that is, deducing the target speaker's appraisal of the stimulus factors corresponding to the specified causal utterance, and then further deriving the corresponding reaction. However, extracting the exact stimulus factors from a given causal utterance may not be as straightforward as locating the causal utterance based on already inferred stimulus factors. Because the stimulus factors are often only a part of the full semantic content of its corresponding causal utterance. Therefore, it is necessary to consider the entire process of the emotion generation process first. In our designed <rationalization> prompt, we first add an "explanation" step. In this step, we ask the LLM to explain in free text why the specified causal utterance would trigger the target speaker's target emotion, and what the evoking process is like, to fully consider the connection between the target utterance and the specified causal utterance. Following this, the model can extract a stimulus factor from the causal utterance more precisely and then deduce its subsequent appraisal and response. This approach results in clearer and more precise rationales, especially in the stimulus part.

d) Consolidation: For a conversation sample, we need to merge the reasoning chains after filtering and rationalization.

Statistics	Train	Valid	Test
Positive Causal Pairs	7,027	328	1,767
Negative Causal Pairs	20,646	838	5,330
Unique Conversations	834	47	225
Samples	4,562	200	1,099

Table 1: Dataset statistics. We consider each target utterance as a sample. A conversation may contain several target utterances, forming several samples.

However, the merged rationale may contain many semantically similar items in each part. To make it more concise and clear, we design a <consolidation> prompt to instruct the LLM to condense the items in the *Reaction*, *Appraisal*, and *Stimulus* parts if needed. We require the LLM to merge items with similar semantics in each part while avoiding excessive modifications to the original phrasing.

Multi-task Training with ECR-Chain Set

After constructing the ECR-Chain set, we can utilize it to assist the supervised training of smaller models. We adopt the MT-CoT method proposed by [Li *et al.*, 2022b] to train on both the golden answer and the collected rationales. Specifically, we utilize two types of instructions. The one uses a simple <answer> prompt that guides the model to directly output the answer based on the question, namely $P(A|Q)$. And the other applies the <reasoning> prompt to instruct the model performing $P(R, A|Q)$. During the training process, we construct each question sample into these two types of instructional inputs separately and calculate the $\mathcal{L}_{reasoning}$ and $\mathcal{L}_{answer}$ with the referenced rationale constructed by LLM and the golden causal utterance label, respectively. The overall loss function is simply a mixture of these two losses:

$$\mathcal{L}_{MT} = \mathcal{L}_{reasoning} + \mathcal{L}_{answer} \qquad (1)$$

This approach enables our trained model to perform inference in different manners. If we only need to identify the causal utterances in applications like the CEE task, we can use the <answer> instruction to predict directly. If we expect an explainable generation, we can use the <reasoning> instruction to provide step-by-step descriptions of the reasoning process.

4 Experiment

4.1 Dataset

We conduct experiments on the RECCON-DD dataset [Poria *et al.*, 2021]. This dataset supplements causal utterance annotations for each non-neutral utterance in the conversations of the DailyDialog dataset [Li *et al.*, 2017]. For a target utterance, its corresponding causal utterances form positive causal pairs with it, while the remaining utterances in the conversation history form negative pairs with it. Following [Zhao *et al.*, 2023a], we only consider the potential causes within the conversation history and remove repetitive positive causal pairs in the original dataset. The statistical details of the dataset are presented in Table 1. Following Poria et al. [2021], we report the F1 scores of both negative and positive causal pairs and the macro F1 scores of them when evaluating performance on the CEE task.

	Method	Neg. F1	Pos. F1	Macro F1
1-shot	Zhao et al.[†]	82.10	52.84	67.47
	Answer	80.41	47.00	63.71
	Reasoning	**85.92**	**55.05**	**70.48**
4-shot	Answer	77.40	53.46	65.43
	Reasoning	**87.40**	**58.97**	**73.19**

Table 2: Results of Few-shot CEE. [†] denotes the results referred from [Zhao *et al.*, 2023b]. "Answer" means instructing the LLM to directly predict the answer, whereas "Reasoning" means guiding the LLM to follow the ECR-Chain to derive the answer.

T	R	A	S	Neg. F1	Pos. F1	Macro F1
✓	✓	✓	✓	**87.40**	**58.97**	**73.19**
×	✓	✓	✓	87.35	57.71	72.53
✓	×	✓	✓	85.54	55.32	70.43
×	×	✓	✓	85.93	58.00	71.97
×	×	×	✓	78.28	47.31	62.79

Table 3: Ablation study on the individual parts of the ECR-Chain. "T", "R", "A", and "S" correspond to theme, reaction, appraisal, and stimulus, respectively.

4.2 Implementation Details

We utilize ChatGPT (gpt-3.5-turbo-0613) as our LLM and set the temperature to 0 to reduce the randomness. During the construction of the ECR-Chain set, we employed 4 hand-written exemplars for the <reasoning> prompt, 4 exemplars for the <rationalization> prompt, and 3 for the <consolidation> prompt to enhance the LLM's understanding of our specific requirements. The details about the examplars and prompts are presented in the Appendix. For the smaller language model, we opt for Vicuna-7B-v1.3 [3] which is based on the LLaMA [Touvron *et al.*, 2023]. Due to the computation limit, we applied LoRA fine-tuning [Hu *et al.*, 2021] for the supervised training. Our total training batch is set to 256 (with gradient accumulation) and the learning rate is set to 1e-3. We train 10 epochs and pick the model that performed best on the validation set to evaluate on the test set. We report the average results of 3 runs for the supervised model.

4.3 CEE with Few-shot Learning

We first report the results of using LLMs to perform the CEE task in a few-shot manner, as shown in Table 2. Zhao et al. [Zhao *et al.*, 2023b] have evaluated the performance of ChatGPT on various emotion-related tasks, including the CEE task. They construct task-specific prompts and ask the model to predict the answer without attempting to guide the model through reasoning. Since they do not provide prompt details in their paper, we directly reference the results they reported. We then compare the CEE performance of directly predicting the answers with reasoning through our ECR-Chain, both under our input format. In the 1-shot scenario, the performance of "Answer" in our experiments is inferior to that of [Zhao *et al.*, 2023b], which might be due to discrepancies in input format, ChatGPT version, or the chosen ex-

---

[3]https://lmsys.org/blog/2023-03-30-vicuna/

Method	Expl.	Neg. F1	Pos. F1	Macro F1
RoBERTa-B	×	88.74	64.28	76.51
RoBERTa-L	×	87.89	66.23	77.06
KEC[†]	×	88.85	66.55	77.70
KBCIN	×	89.65	68.59	79.12
TSAM	×	90.48	70.00	80.24
Answer	×	89.78	68.66	79.22
Reasoning	✓	89.70	63.36	76.53
Multi-task	×	**90.82**	**70.84**	**80.83**
	✓	90.33	65.89	78.11

Table 4: Results of Supervised CEE. "Expl." indicates whether the answer is predicted in an explainable manner. Our multi-task trained model is capable of both directly predicting answers and providing explainable rationales during inference, resulting in two separate lines in this table. [†] denotes that we reference the results from [Zhao *et al.*, 2023a], as the original KEC paper treated neutral utterance as target utterance as well, which differs from our setting.

Method	Macro F1	GPT4 Score	Claude3 Score
Vanilla-Vicuna	53.41	6.63	7.22
MT-Vicuna	**78.11**	7.68	7.56
ChatGPT	70.48	**8.16**	**7.99**

Table 5: Results of Explainable ECR. The "GPT4 Score" and "Claude3 Score" evaluates the quality of rationales, with the score ranging from 1 to 10.

emplar. But our primary focus here is on the effectiveness of reasoning using the ECR-Chain. According to Table 2, guiding the LLM to reason following the ECR-Chain significantly improves causal utterance prediction over the two aforementioned baselines. In the 4-shot scenario, our proposed reasoning method can bring more than a 7% improvement in the Macro F1 metric over the baseline, which demonstrates the benefit of the reasoning process in correctly identifying emotion-causes.

We also conduct an ablation study to analyze the impact of each part of our proposed reasoning chain, as shown in Table 3. The stimulus part, which is directly associated with the final causal utterances, is considered a detailed description of the cause and thus was retained in all experiments to assess the influence of the presence or absence of other components. When the LLM is instructed to directly provide a stimulus description based on the input, the predicted causal utterances corresponding to the stimulus are often inaccurate. In contrast, the accuracy of tracing the cause improves significantly if the model first deduces the speaker's appraisal associated with the target emotion before inferring the stimulus that evoked that emotion. This indicates that "appraisal" is central to the entire process of emotion generation, and understanding appraisal is crucial for comprehending emotions. Additionally, summarizing the conversation's theme and the speaker's emotional reaction is beneficial. It helps the LLM gain a fuller understanding of the dialogue context and the speaker's state, thereby enhancing the causal tracing.

4.4 CEE with Supervised Learning

As shown in Table 4, we compare our multi-task model against models trained on single tasks, namely the "Answer" and "Reasoning" listed in the table. "Answer" refers to only employing the CEE task labels for training the Vicuna. During inference, the model directly outputs the predicted answers, which are not explainable. "Reasoning" denotes fully supervising the Vicuna by ECR-Chains generated by the LLM during training. The trained model outputs rationales in a step-by-step manner during inference, making the results explainable. The multi-task model, as described in

Section 3.4, is trained using both answer-only and reasoning supervision, allowing the model to either directly predict answers or generate rationales to derive the answer, depending on the form of the input instruction. Besides, we report the results of some previous methods based on RoBERTa: **KEC** [Li *et al.*, 2022a], **KBCIN** [Zhao *et al.*, 2023a], **TSAM** [Zhang *et al.*, 2022a], along with the results of vanilla RoBERTa. These previous works extract features from utterances, model interactions between different features within the conversation, and then perform binary classification on each utterance feature to obtain a prediction of either positive or negative. Hence, the predictions from these works are unexplainable.

Compared to the answer-only training strategy, incorporating the ECR-Chain set for multi-task training can bring an obvious performance gain when utilizing the <answer> instruction for inference. Our multi-task model, when directly predicting answers, also surpasses the previous baselines. This indicates that our ECR-Chain set, providing more detailed and comprehensive supervision, can improve the supervised models to capture emotion-causes more precisely.

When comparing the results of "Answer" and "Reasoning", we observe that performing reasoning in training and inference is less effective than directly predicting the answer for the CEE task. We think there are two possible reasons. Firstly, the long length of our rationales may lead to insufficient optimization of the model for the final answer prediction part, which constitutes only a small portion of an entire rationale. Secondly, the reasoning ability of the smaller model may be limited by its scale. We consider this an open problem, as similar phenomena were also observed in the one-stage fine-tuning experiments of [Zhang *et al.*, 2023]. The multi-task training method can mitigate the first issue, and our experiments show that the predictions derived through reasoning in multi-task models indeed bring improvement. Moreover, the explainable outcomes achieved by our multi-task model via <reasoning> instruction are comparable to those unexplainable baselines specifically training for the CEE task from previous works, further highlighting the advantages of our model.

In summary, our multi-task model can be regarded as a versatile model, capable of providing precise predictions directly via the <answer> instruction, as well as offering explainable reasoning of the predictions through the <reasoning> instruction.

4.5 Explainable Emotion-Cause Reasoning

Here we evaluate the performance of models when performing emotion-cause reasoning in an explainable manner, as shown in Table 5. "Vanilla-Vicuna" refers to the original Vicuna model without our fine-tuning, while "MT-Vicuna" is

Conversation:	[Vanilla-Vicuna]	[MT-Vicuna]	[ChatGPT]
#1: A (neutral): "Is there anything I can do for you ?" #2: B (neutral): "Would you please show me that mural ?" #3: A (happiness): "Certainly . Here you are ." #4: B (happiness): "It looks so beautiful . How much is it ?" #5: A (neutral): "Four hundred yuan ." #6: B (neutral): "That's a bit more than I wanted to pay . Can you cut the price down a bit ?" #7: A (neutral): "I'm sorry . Ours is a fixed price shop and I can assure you this is the best price you can get for a mural like this ." #8: B (neutral): "Maybe you are right . OK , I'll take it ." #9: A (happiness): "Well , I'll have it wrapped for you ." #10: B (happiness): "Thank you . Here is four hundred yuan ." #11: A (happiness): "Thank you . Please come again . " **Question**: Which utterances cause A's happiness in #11? **Label**: [**#10**]	**Theme:** The conversation is about buying a mural. **Reactions of A in #11:** 1. A is happy to help B and show the mural. 2. A is happy that B likes the mural and wants to buy it. **Appraisals of A in #11:** 1. A thinks B is a nice customer. 2. A thinks B is willing to pay the fixed price for the mural. **Stimuli:** 1. B asks A to show the mural (**#2**). 2. B is interested in buying the mural (**#4**, **#6**). 3. A thinks B is a nice customer (**#11**). **GPT4 Score**: 6.5	**Theme:** Selling a mural and expressing gratitude. **Reactions of A in #11:** 1. A expresses happiness and gratitude towards B for making a purchase. **Appraisals of A in #11:** 1. A thinks it is a good thing that B made a purchase and is satisfied with the service. **Stimuli:** 1. B made a purchase and gave A four hundred yuan (**#10**). **GPT4 Score**: 8.0	**Theme:** Buying a mural in a shop. **Reactions of A in #11:** 1. A is happy to receive the payment from B. 2. A is happy that B expressed gratitude. **Appraisals of A in #11:** 1. A thinks it is a positive interaction with B. 2. A feels satisfied with the transaction. **Stimuli:** 1. B thanked A and paid for the mural (**#10**). 2. B agreed to purchase the mural (**#8**). 3. A wrapped the mural for B (**#9**). **GPT4 Score**: 9.0
(a)	(b)	(c)	(d)

Figure 4: Case Study on Explainable ECR. (a) shows the question and the labeled causal utterances. (b), (c), and (d) show rationales generated by three models. According to the analysis of GPT-4, the output of our multi-task model is considered to be "more concise and straightforward". The other two models, however, are thought to "contain inaccuracies or assumptions", as indicated by the orange background content in (b) and (d). Nonetheless, GPT-4 states that ChatGPT "provides a more detailed analysis of the stimuli" and "exhibits a more comprehensive understanding of the situation", as indicated by the green background content in (d). This is the reason for the higher score awarded to ChatGPT.

our multi-task trained model. All models apply the same 1-shot <reasoning> instruction as input. "Macro F1" evaluates the final reasoned causal utterance indexes against golden labels. The "GPT-4/Claude3 Score", on the other hand, employs advanced LLMs, GPT-4 and Claude3-Opus, to assess the quality of the generated rationales. Specifically, we follow the evaluation template provided in FairEval [Wang *et al.*, 2023] to design our own evaluating prompt. We instruct it to focus on three aspects of the rationale: logical coherence within the rationale, the relevance of the rationale to the original conversation, and the plausibility of the final stimulus description. It is required to give an overall score for each rationale, together with the scoring explanations. More detailed evaluation settings and evaluation prompts are presented in the Appendix.

For the "Macro F1" score of causal utterance prediction in Table 5, our multi-task model performs the best, even surpassing the results of the teacher model ChatGPT, which indicates that our ECR-Chain set refined based on the golden labels can effectively aid supervised models in deriving the expected answers. And for the metric of rationale quality, "GPT-4/Claude3 Score", our multi-task model also significantly outperforms the vanilla vicuna, proving that our constructed ECR-Chain set can effectively enhance the reasoning abilities and the explainability for emotion causes via supervised training. However, there is still a certain gap between our multi-task model and ChatGPT. On the one hand, this may be due to ChatGPT providing rationales richer in logical coherence and more aligned with the conversation context, which reflects the powerful logical and explanatory capabilities of ChatGPT brought about by its large parameter scale. On the other hand, the reason could also be that ChatGPT offered some causal explanations that led to answers mismatched with the labels, yet were convincing enough for the evaluation LLMs to consider them reasonable. These explanations might also make sense, considering the subjectivity involved in emotion-cause annotations.

We further conduct a case study as shown in Figure 4. In this case, the vanilla Vicuna performs the worst, with its reaction summary not aligning with the target sentence, leading to an unreasonable derivation of the stimulus content. Our multi-task model precisely identifies the direct factor, matching the final answer with the label and presenting a concise and straightforward reasoning chain. The ChatGPT seems to introduce some assumptions in the reaction part, but is considered to provide more detailed stimuli by GPT-4. Actually, in this example, the golden label corresponds to the most direct cause, while the additional two stimulus items provided by ChatGPT can be viewed as indirect background factors. Therefore, if our goal is to extract more precise stimulus factors and acquire more direct reasoning chains, we may opt for the supervised model trained on our ECR-Chain set to avoid excessive irrelevant generated content. If we aim to discover more potential factors, we can directly instruct the LLM with our proposed ECR-Chains to fully leverage its rich knowledge and powerful reasoning capabilities, which may result in some insightful descriptions related to the emotion generation process.

5 Conclusion

In this paper, we explore emotion-cause reasoning with generative language models. We first introduce the Emotion-Cause Reasoning Chain (ECR-Chain), which can guide large language models (LLMs) to infer the emotion causes from emotional expressions with CoT prompting. Considering the limited reasoning abilities of smaller models, we utilize LLMs to automatically construct an ECR-Chain set through a four-stage processing workflow and perform supervised training based on the ECR-Chain set through multi-task learning. Experimental results demonstrate that our method can enhance the prediction of emotion-cause utterances while also enabling generative language models with the ability to perform explainable emotion-cause reasoning.

Ethical Statement

In this paper, we utilize ChatGPT to construct a reasoning chain dataset for supervised training. We notice that during the reasoning process, ChatGPT may generate hallucinations, such as conflating the identities of two speakers within a conversation. We propose revising strategies based on task labels and find that they can effectively remove some items that contain factual errors. However, there may still be some misleading content remaining in the set.

Acknowledgments

References

[Arnold, 1960] Magda B Arnold. Emotion and personality. vol. i. psychological aspects. 1960.

[Brown *et al.*, 2020] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[Diao *et al.*, 2023] Shizhe Diao, Pengcheng Wang, Yong Lin, and Tong Zhang. Active prompting with chain-of-thought for large language models. *arXiv preprint arXiv:2302.12246*, 2023.

[Ellsworth, 1991] Phoebe C Ellsworth. Some implications of cognitive appraisal theories of emotion. 1991.

[Ghosal *et al.*, 2019] Deepanway Ghosal, Navonil Majumder, Soujanya Poria, Niyati Chhaya, and Alexander Gelbukh. Dialoguegcn: A graph convolutional neural network for emotion recognition in conversation. 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 154–164, 2019.

[Ho *et al.*, 2023] Namgyu Ho, Laura Schmid, and Se-Young Yun. Large language models are reasoning teachers. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14852–14882, Toronto, Canada, July 2023. Association for Computational Linguistics.

[Hsieh *et al.*, 2023] Cheng-Yu Hsieh, Chun-Liang Li, Chih-kuan Yeh, Hootan Nakhost, Yasuhisa Fujii, Alex Ratner, Ranjay Krishna, Chen-Yu Lee, and Tomas Pfister. Distilling step-by-step! outperforming larger language models with less training data and smaller model sizes. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8003–8017, Toronto, Canada, July 2023. Association for Computational Linguistics.

[Hu *et al.*, 2021] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

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

[Li *et al.*, 2017] Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 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, 2017.

[Li *et al.*, 2022a] Jiangnan Li, Fandong Meng, Zheng Lin, Rui Liu, Peng Fu, Yanan Cao, Weiping Wang, and Jie Zhou. Neutral utterances are also causes: Enhancing conversational causal emotion entailment with social commonsense knowledge. In Lud De Raedt, editor, *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pages 4209–4215. International Joint Conferences on Artificial Intelligence Organization, 7 2022. Main Track.

[Li *et al.*, 2022b] Shiyang Li, Jianshu Chen, Yelong Shen, Zhiyu Chen, Xinlu Zhang, Zekun Li, Hong Wang, Jing Qian, Baolin Peng, Yi Mao, et al. Explanations from large language models make small reasoners better. *arXiv preprint arXiv:2210.06726*, 2022.

[Li *et al.*, 2022c] Yifei Li, Zeqi Lin, Shizhuo Zhang, Qiang Fu, Bei Chen, Jian-Guang Lou, and Weizhu Chen. On the advance of making language models better reasoners. *arXiv preprint arXiv:2206.02336*, 2022.

[Li *et al.*, 2023] Yingcong Li, Kartik Sreenivasan, Angeliki Giannou, Dimitris Papailiopoulos, and Samet Oymak. Dissecting chain-of-thought: Compositionality through in-context filtering and learning. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[Liu *et al.*, 2019] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[Liu *et al.*, 2022] Yuchen Liu, Jinming Zhao, Jingwen Hu, Ruichen Li, and Qin Jin. Dialogueein: emotion interaction network for dialogue affective analysis. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 684–693, 2022.

[Majumder *et al.*, 2019] Navonil Majumder, Soujanya Poria, Devamanyu Hazarika, Rada Mihalcea, Alexander Gelbukh, and Erik Cambria. Dialoguernn: An attentive rnn for emotion detection in conversations. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 6818–6825, 2019.

[Marasović *et al.*, 2022] Ana Marasović, Iz Beltagy, Doug Downey, and Matthew E Peters. Few-shot self-rationalization with natural language prompts. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 410–424, 2022.

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

[Touvron *et al.*, 2023] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[Wang *et al.*, 2022] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*, 2022.

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

[Wei *et al.*, 2019] Wei Wei, Jiayi Liu, Xianling Mao, Guibing Guo, Feida Zhu, Pan Zhou, and Yuchong Hu. Emotion-aware chat machine: Automatic emotional response generation for human-like emotional interaction. In *Proceedings of the 28th ACM international conference on information and knowledge management*, pages 1401–1410, 2019.

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

[Zelikman *et al.*, 2022] Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah Goodman. Star: Bootstrapping reasoning with reasoning. *Advances in Neural Information Processing Systems*, 35:15476–15488, 2022.

[Zhang *et al.*, 2022a] Duzhen Zhang, Zhen Yang, Fandong Meng, Xiuyi Chen, and Jie Zhou. Tsam: A two-stream attention model for causal emotion entailment. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6762–6772, 2022.

[Zhang *et al.*, 2022b] Zhuosheng Zhang, Aston Zhang, Mu Li, and Alex Smola. Automatic chain of thought prompting in large language models. In *The Eleventh International Conference on Learning Representations*, 2022.

[Zhang *et al.*, 2023] Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. Multimodal chain-of-thought reasoning in language models. *arXiv preprint arXiv:2302.00923*, 2023.

[Zhao *et al.*, 2023a] Weixiang Zhao, Yanyan Zhao, Zhuojun Li, and Bing Qin. Knowledge-bridged causal interaction network for causal emotion entailment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 14020–14028, 2023.

[Zhao *et al.*, 2023b] Weixiang Zhao, Yanyan Zhao, Xin Lu, Shilong Wang, Yanpeng Tong, and Bing Qin. Is chatgpt equipped with emotional dialogue capabilities? *arXiv preprint arXiv:2304.09582*, 2023.

[Zhong *et al.*, 2021] Peixiang Zhong, Di Wang, Pengfei Li, Chen Zhang, Hao Wang, and Chunyan Miao. Care: commonsense-aware emotional response generation with latent concepts. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14577–14585, 2021.