Enhancing Large Language Model with Decomposed Reasoning for Emotion Cause Pair Extraction

Anonymous ACL submission

Abstract

Emotion-Cause Pair Extraction (ECPE) involves extracting clause pairs representing emotions and their causes in a document. Existing methods tend to overfit spurious correlations, such as positional bias in existing benchmark datasets, rather than capturing semantic features. Inspired by recent work, we explore leveraging large language model (LLM) to address ECPE task without additional training. Despite strong capabilities, LLMs suffer from uncontrollable outputs, resulting in mediocre performance. To address this, we introduce chain-of-thought to mimic human cognitive process and propose the *Decomposed Emotion-Cause Chain (DECC)* framework. Combining inducing inference and logical pruning, DECC effectively guides LLMs to tackle ECPE task. We further enhance the framework by incorporating in-context learning. Experiment results demonstrate DECC's strength compared to state-of-the-art supervised fine-tuning methods. Finally we analyses each component's effectiveness and method's robustness in various scenarios including different LLM base, rebalanced datasets, and multi-pair extraction.

1 Introduction

Analyzing the causes behind emotions is an interesting research direction in the field of affective computing, attracting increasing attention in recent years (Su et al., 2023). Understanding the reason for occurence of emotions holds commercial value in various areas, including product review mining (Craciun and Moore, 2019), social media mining (Soong et al., 2019), and cognitive-behavioral therapy in psychology (Berking et al., 2013).

Early studies on emotion cause analysis primarily focused on clause-level emotion cause extraction (ECE), which aims to extract cause clauses for given emotion clauses. However, the need of emotion annotation limits ECE's application. As a result, the emotion-cause pair extraction (ECPE) task is introduced (Xia and Ding, 2019). This task aims to extract potential clause pairs of emotions and their corresponding causes in a document. Figure 1 shows an example of ECPE task, where the input document contains several clauses. Both c2 and c5 exhibit emotions. The cause clause for emotion clause c2 is itself, while for emotion clause c5, its associated cause clause is c4. Thus, the output emotion-cause pairs are: (c2, c2), (c5, c4).

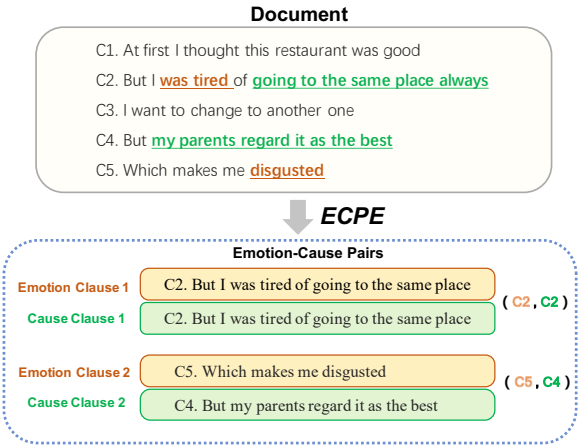

Figure 1: An Example of ECPE task

Although researchers have resorted to various deep-learning approaches (Su et al., 2023) to tackle ECPE task, most methods suffer from position bias problem (Zheng et al., 2022; Ding and Kejriwal, 2020). This issue arises from the relative position information introduced in most existing methods, making ECPE models sensitive to data distribution and lacking robustness. Specifically, given that cause clauses often immediately precede their emotion clause, models may obtain high performance without capturing actual contextual information.

Recently, the rapidly evolving field of natural language processing witnesses the rise of large language model (LLM), such as ChatGPT [1] and LLAMA (Touvron et al., 2023), which have

[1] https://chat.openai.com/

achieved impressive performance in various NLP tasks (Sun et al., 2023; Qin et al., 2023) under zero-shot or few-shot scenarios, even without updating parameters. Due to their powerful capabilities and significant potential, we make effort to leverage LLMs to address the ECPE task in this paper.

However, due to the generation paradigm's nature, directly utilizing large models for information extraction tasks often leads to answers with high recall but low precision (Han et al., 2023a). In Wang et al. (2023), a constraint is imposed in the prompt, requiring the model to output only one emotion-cause pair. This approach not only fails to achieve satisfactory performance but also conflicts the requirements of practical application scenarios.

Since directly instructing a LLM to output all emotion-cause pairs at once yields unsatisfactory results, a question arises: 'Is there a way to decompose this task and utilize the LLM to get the answer step by step?' Fortunately, the Chain-Of-Thought (COT) technique provides a suitable solution.

The core idea of COT involves generating a series of intermediate reasoning steps, significantly enhancing LLMs' ability to perform complex reasoning. Inspired from existing COT techniques and insights from human cognitive processes (Fiske and Taylor, 1991), we have custom-designed a Decomposed Emotion Cause Chain (DECC) framework. Specifically, We decompose the solving process of the ECPE problem into four steps: **Recognizing**, **Locating**, **Analyzing**, **Summarizing**. Each of these steps corresponds to a prompt used for interaction with the LLM, and the output from each step becomes part of the prompt for the next step. The entire framework executes in a pipeline manner.

Besides zero-shot decomposed COT, we also employ In-Context Learning (ICL) with automatic selected demonstrations to enhance the framework. Experiment results verify the effectiveness of our method. Our primary contributions are as follows:

- We propose a DECC framework to address ECPE task, treating the task as a multi-step reasoning problem. Each step corresponds to a sub-problem tailored to the characteristics of ECPE. The framework then solves the sub-problems sequentially following the decomposed chain-of-thought paradigm.

- We employ reliable in-context learning with automatically established demonstrations to further improve the practical performance.

- We conduct experiments both on traditional benchmark datasets and improved ones. Our DECC framework can achieve comparable results to previous supervised fine-tuning methods, including state-of-the-art ones. Further analysis indicates that our approach can effectively alleviate the inherent issues of ECPE that challenge traditional methods.

2 Related Work

2.1 Emotion Cause Pair Extraction Task

After Chen et al. (2010) first introduced the task of Emotion Cause Extraction (ECE), Xia and Ding (2019) proposed Emotion-Cause Pair Extraction (ECPE) to extract emotion clause (EC) and cause clause (CC) simultaneously, which aims to reduce the cost of manual labelling in ECE. Early solutions for the ECPE task are pipeline approaches (Xia and Ding, 2019; Xu et al., 2021), where EC and CC are extracted sequentially, which will suffer from error propagation issue. Subsequently, many end-to-end approaches (Ding et al., 2020b; Cheng et al., 2020; Wei et al., 2020; Chen et al., 2020; Liu et al., 2022) have been proposed and achieved better results. Since prompt (Brown et al., 2020) has became popular in the NLP field, Zheng et al. (2022) designed a BERT-based prompting method which achieved state-of-the-art performance. However, Ding and Kejriwal (2020) uncovered severe position bias in traditional ECPE benchmark and proposed a rebalanced dataset, on which the performance of most previous carefully designed models decrease drastically.

2.2 Chain-of-Thought Prompting

To exploit the reasoning ability in LLMs, Wei et al. (2022) proposed Chain-of-Thought (COT) prompting, appending multiple reasoning steps before answering to the input question. With this simple prompting strategy, LLMs are able to perform much better in complex problems. Subsequently, a variety of works (Huang and Chang, 2022) are proposed to further improve COT prompting in different aspects, including prompt selection (Diao et al., 2023; Lu et al., 2022), self-consistency (Wang et al., 2022), problem decomposition (Khot et al., 2022; Zhou et al., 2022), automatic prompting (Shum et al., 2023; Shao et al., 2023), and so on.

The most relevant work to this paper is Wang et al. (2023), which evaluates the performance of chatGPT on many tasks, including ECPE. According to its findings, using chatGPT directly for

the ECPE task does not yield satisfactory results. Therefore, we have chosen to enhance models by integrating COT with the task's inherent characteristics. To the best of our knowledge, this is the first work of leveraging COT in the ECPE task.

3 Methodology

3.1 Task Definition

In the ECPE task, the input is a document D, which consists of several clauses $D = [c_1, c_2, ..., c_N]$. Here, N represents the number of clauses in D. The objective of the ECPE task is to extract all emotion-cause pairs (c_e, c_{ca}) from D at the clause level, where c_e denotes an emotion clause and c_{ca} represents the corresponding cause clause.

3.2 DECC Framework

The Decomposed Emotion-Cause Chain (DECC) is a multi-step reasoning framework based on decomposed chain-of-thought. It involves several interactive prompting steps to optimize the output of large language model. Unlike conventional fine-tuning methods, we aim to utilize the comprehension ability of LLM without additional training cost. As is shown in Figure 2, we design a series of deducing prompts to filter out improper outputs. Inspired by the human cognitive process and attribution theory (Fiske and Taylor, 1991), which state that when human beings perceive something they are inclined to find the cause, the DECC framework is composed of two major phases: Emotion Extraction and Cause Extraction. Each phase also comprises two steps as follows:

3.2.1 Emotion Extraction Phase

The process of emotion extraction consists of the following two steps:

- **Step 1. Recognizing**: We instruct LLM with a prompt[2] to recognize all potential emotions from D, presenting them in the form of emotional keywords. It's worth noting that owing to the powerful understanding ability of LLM, the emotions identified here include both explicit and implicit ones. As shown in Figure 2, the identified emotion *worry* in step 1 belongs to explicit ones (appears in $c6$). Although the emotional keyword *surprise* does not appear in the document, it is distilled by LLM as an implicit emotion based on the content.

[2]Please refer Figure 2 for details of prompts we use.

- **Step 2. Locating**: For each identified emotional keyword, LLM is prompted to find its precise location in document D. If a keyword is successfully located, it proceeds to the next step, and its location is considered as a potential emotion clause. On the other hand, if the emotional keyword cannot be located in D, it is pruned. As presented in Figure 2, the emotional keyword *angry* is pruned since there is no clause that explicitly conveys the emotion of anger. For emotional keyword *surprise*, although it is not literally mentioned in D, it is implicitly expressed in clause $c7$. In other words, its presence can be inferred from the content of the document.

The Emotion Extraction process is expressed by:

$$\mathbb{EC} = \{\mathrm{c} \in C_{loc} \mid \exists e \in \mathrm{E_{rec}} \text{ s.t. } e \subset \mathrm{c}\} \quad (1)$$

Where C_{loc} denotes the clauses located from the passage, E_{rec} denotes the recognized emotions and $\mathbb{EC}$ is the potential emotion clause set. A clause is eligible to be selected as an emotional clause for the next step only if it contains the recognized emotional keyword; otherwise, it will be pruned. This approach ensures that the locating operation prunes mistakenly recognized emotional keywords, and the recognition step limits the output of clauses.

3.2.2 Cause Extraction Phase

When instructing LLM to find a cause clause for each emotion directly, the performance would be disastrous due to the complexity of the ECPE task itself and the uncontrollability of LLM. On the other hand, if we directly ask the LLM to conclude the result pairs, it is very possible for the LLM to obfuscate the granularity of the cause clause. Because every part of reasoning chain could be the cause while only one clause could be the ground truth. Also, because the zero-shot output is unpredictably irregular and contains a large block of tokens, it is challenging to search for final pairs for evaluation. In order to solve the aforementioned problem, we design following two steps:

- **Step 3. Analysing**: Given potential emotion clauses ($\mathbb{EC}$), the LLM is prompted to analyse the underlying reasons behind each emotion clause in a step-by-step manner. For each emotion clause , we instruct the LLM to generate an analysing chain ac, whose beginning is an

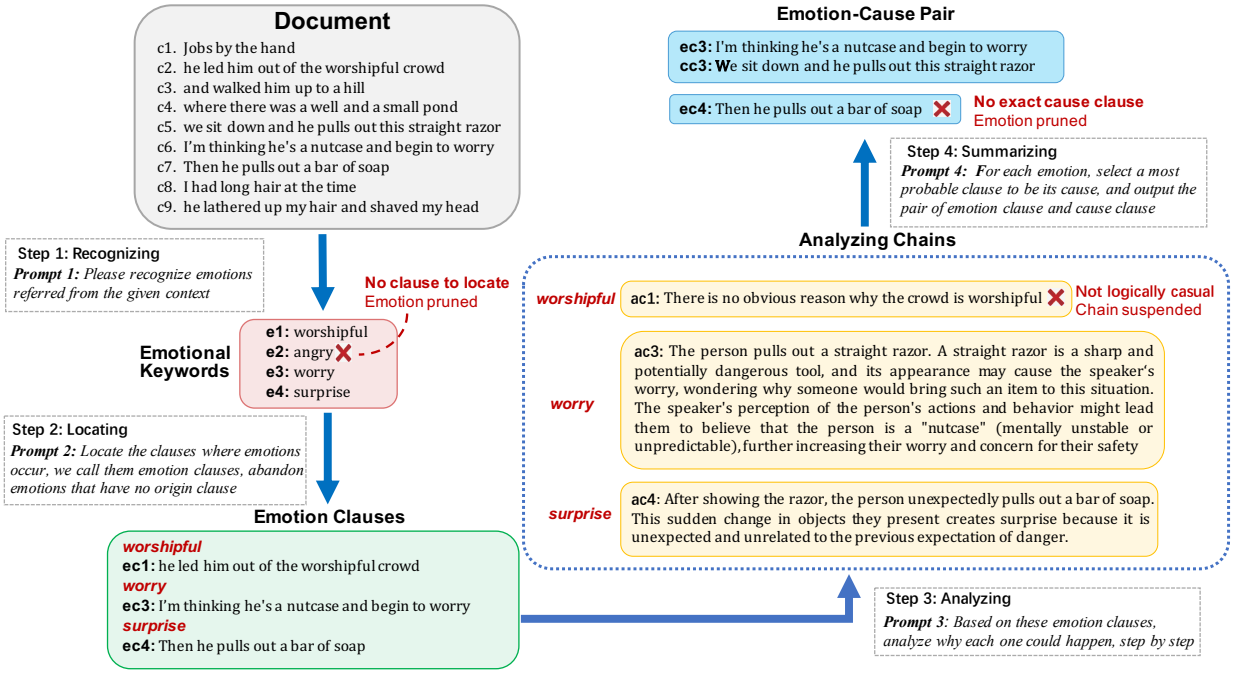

Figure 2: Illustration of DECC framework. At locating step, emotional keyword *angry* is pruned. The analyzing chain corresponding to *worshipful* is pruned at step 3. The analyzing chain of *surprise* is pruned at the last step. *e1-e4* denote the id of specific emotional keyword. *ec* and *cc* denote emotion clause and cause clause, respectively. *ac* stands for the corresponding analyzing chain.

emotion clause ec and content is the reasoning process based on the overall document D.

$$\mathbb{AC} = \{Analyze\,(ec, D) \mid ec \in EC\} \quad (2)$$

If the reason is successfully deduced and can be attributed to a specific clause in the document, the corresponding rationale (i.e., analyzing chain) generated at this step will be passed to the next. Otherwise, the chain would stop analyzing automatically. As illustrated in Figure 2, analyzing chain $ac1$ is suspended because $ec1$ is not logically casual according to the document content and has no obvious reason. Consequently, the branch corresponding to $ec1$ and $ac1$ is pruned at this step.

- **Step 4. Summarizing**: At this step, the LLM is instructed to identify the most probable cause clause cc for each emotion and summarize final emotion-cause pairs (ec, cc).

$$\mathbb{P} = \{(ec, \widehat{cc}) \mid \widehat{cc} = argmaxP\,(cc \mid ec, ac)\} \quad (3)$$

Here, some pruning operations will also be carried out. As depicted in Figure 2, the branch of $ac4$ is pruned because the cause of *surprise* cannot be simply attributed to a specific clause. Meanwhile, this step is also designed to ensure that the output follows a standardized format, making the generated answer more convenient for further evaluation.

3.3 DECC with In-context Learning

Besides zero-shot prompting, in-context learning (ICL), which incorporates examples within the prompt, is also widely employed in recent works (Zhao et al., 2023) to further enhance larger language models. Since ICL is very sensitive to the included demonstration examples (Dong et al., 2022), their selection is a central challenge to address.

Inspired by previous studies (Shao et al., 2023; Fu et al., 2022; Qiao et al., 2022), we select demonstrations with the criterion of ensuring diversity. Specifically, we cluster the documents in the training set in a semantic embedding space, using Sentence-BERT (Reimers and Gurevych, 2019) as the encoder. The number of clusters is equal to the number of demonstrations used for inference. We then choose the documents closest to the centroid of each cluster as demonstration candidates. For each candidate document, we follow the DECC framework to generate its rationale step by step. After manually modify these rationales to ensure their accuracy, we put the selected documents along with their corresponding rationales into the prompt, thus completing the construction of the demonstrations

for ICL. With selected demonstrations, the LLM will be elicit to generate more intermediate reasoning rationales with high quality which is helpful to the final prediction. More details about ICL-based DECC are elaborated in following sections.

4 Experimental Settings

4.1 Datasets

We evaluate our method on the following three published ECPE datasets [3]:

The Chinese ECPE dataset (Xia and Ding, 2019). It is the most widely used benchmark dataset.

The English ECPE dataset (Singh et al., 2021). It is mainly composed of an English novel corpus.

The rebalanced Chinese ECPE dataset (Ding and Kejriwal, 2020). Ding and Kejriwal (2020) have identified extreme imbalance of emotion-cause position in the Chinese ECPE dataset. They further uncovered that the performance of current state-of-the-art methods will dramatically degrade without this distribution bias. To address this issue, they proposed this new rebalanced ECPE dataset, which is constructed based on the Chinese ECPE dataset.

Since our method doesn't require any training data, we maintain the same dataset split as previous studies and conduct experiments directly on the test sets [4].

4.2 Evaluation Metrics

We compute the Precision, Recall and F1 score of emotion-cause clause pair for evaluation. However, due to the nature of generative models, evaluation for extraction task such as ECPE is non-trivial (Han et al., 2023b). LLMs often generate reasonable answers that are correct under manual review but fail to match the ground truth (Wadhwa et al., 2023). Consider the following example, suppose there is a ground truth cause clause *" my father takes away the toys"*, while the output of LLM might be *"it is author's father takes his toys away that most probably triggers the emotion"*. The answer indicates exactly the same information with the target, so it should be regarded as a correct output. Since the conventional exact matching evaluation for ECPE task is unfair for LLMs, we follow the methods introduced by Wang et al. (2023) and conduct human evaluation, enlisting expert annotators to judge whether the model outputs convey the same information as the ground truth.

[3] Datasets statistics are listed in Appendix A.

[4] Implementation details are reported in our Appendix B

4.3 Baselines

We compare our method with following baselines. **ECPE-2D** (Ding et al., 2020a) is the most outstanding two-step method. End-to-end methods include **UTOS** (Cheng et al., 2021), **MTST** (Fan et al., 2021) and **ECPE-MLL** (Ding et al., 2020b) which convert ECPE to a sequence labelling task or multi-label classification task, and **RankCP** (Wei et al., 2020) uses BERT+GCN as the clause encoder. **UECA-Prompt** (Zheng et al., 2022) is the state-of-the-art supervised fine-tuning method on both Chinese dataset and rebalanced dataset. We also employ **GPT3.5** with naive prompting for ECPE task (Wang et al., 2023) as a LLM baseline.

5 Results and Analysis

5.1 Zero-short Result

Since ChatGPT is the most popular LLM, we mainly evaluate DECC based on GPT3.5-turbo in the experiment. As reported in Table 1, DECC with GPT3.5-turbo yields a significant improvement, particularly in terms of precision, because many emotion-cause pairs with low probabilities are pruned at step 2 and 3. Despite of some over-pruning operations leads to a certain level of recall degradation, the overall F1 score is still remarkably higher compared to the naive prompting method with GPT3.5-turbo (Wang et al., 2023). Nevertheless, DECC under zero-shot settings still lags behind most fully-supervised fine-tuning methods.

5.2 ICL Few-short Result

After enhancing with ICL (4-shot), the performance of DECC further improves. The performance on both Chinese dataset and English dataset are very close to the SOTA fine-tuning based method. Compare to the zero-shot version of DECC, the recall improves significantly and the precision stay stable on Chinese dataset. It seems that GPT3.5-turbo has grasped the specific rule of the task and the impact of over pruning is alleviated after provided with demonstrations. However, the opposite phenomenon occurred on the English dataset, the recall stays stable and the precision improves a lot. We speculate that the reason for this might be due to annotation errors in the dataset. Specifically, for each document in the English ECPE dataset, only one emotion-cause pair is labelled as ground truth target while other potential correction pairs are simply ignored. Therefore, the demonstrations extracted from the English dataset

Method	Chinese dataset			English dataset			Rebalanced CN dataset			De-bias drop
	P	R	F1	P	R	F1	P	R	F1	F1-ratio
ECPE-2D	72.92	65.44	68.89	**60.49**	43.84	50.73	47.22	37.38	41.73	-39.43%
UTOS	73.89	70.62	72.03	55.69	48.03	**51.53**	42.76	28.95	34.14	-52.60%
MTST-ECPE	**77.46**	71.99	74.63	52.37	43.54	47.47	51.99	40.34	44.93	-39.80%
RankCP	71.19	76.30	73.60	44.00	45.35	44.63	43.22	39.16	41.22	-43.99%
ECPE-MLL	77.00	72.35	74.52	52.96	45.30	51.21	**61.53**	36.39	45.57	-38.85%
UECA-Prompt	71.82	77.99	**74.70**	-	-	-	46.30	53.22	49.37	-33.91%
GPT3.5 prompt (0-shot)	40.74	67.54	50.82	42.11	39.34	40.68	40.72	38.14	39.39	-22.49%
GPT3.5 DECC (0-shot)	61.54	49.76	55.03	34.60	**59.84**	43.84	45.82	49.15	47.42	-13.83%
GPT3.5 DECC (4-shot)	61.23	**81.56**	69.95	46.89	54.42	50.35	50.00	**79.45**	**61.38**	-12.25%

Table 1: Results of different methods on three datasets. The results of traditional baselines are from corresponding original papers. "-" means not available in the original paper. The results of GPT3.5 prompt (zero-shot) are reproduced based on the prompt provided in (Wang et al., 2023).'De-bias drop' indicates the *F1* reducing ratio of the methods from traditional Chinese dataset to rebalanced Chinese dataset. Best results are marked in **bold**.

only elicit LLM to generate the most confident outputs, rather than attempting to extract all possible emotion-cause pairs. This makes LLM relatively conservative, resulting in an increase in precision but no improvement in recall.

5.3 De-bias Result

In 80% of the Chinese ECPE dataset samples, emotions and their corresponding causes either appear in adjacent clauses or belong to the same clause, causing a significant position bias (Ding and Kejriwal, 2020). Therefore, we also examine different methods on the rebalanced dataset. The results indicate that the performance of all previous work decreased drastically by an average of 40% on F1 score. In comparison, the F1 score decrease of DECC is significantly smaller than other models on the rebalanced dataset and the DECC 4-shot with GPT outperforms the state-of-the-art method with a large margin on F1 score.

Previous methods badly over-fit the spurious correlation introduced by the position bias instead of capturing the semantic information from the context, which causes the deterioration in performance. On the contrary, our approach relies solely on the understanding capability of LLM itself and the reasoning ability provided by the decomposed COT. It is less sensitive to the distribution of dataset. Moreover, we do not build ICL demonstrations specifically for experiment on the rebalanced dataset. Although equipped with the same shots as the normal dataset, DECC still yields an outstanding performance, displaying its excellent generalizing ability.

5.4 Performance on Multi-pair Document

There are only less than 30% documents in the datasets containing more than one emotion-cause pair. Most conventional methods pay little attention to this particular samples, since the performance on multi-pair documents does not affect much on the overall performance. In fact, multi-pair document is very common in practical scenarios and it should not be simply neglected. UECA-Prompt (Zheng et al., 2022) is the only existing method which has taken account of multi-pairs extraction and achieved the current state-of-the-art performance. As reported in Table 2, DECC 4-shot has surpassed UECA-Prompt with a large margin and achieve new state-of-the-art performance on multi-pair subsets. This result demonstrates that DECC framework is compatible with both multi-pair and single-pair extraction scenarios.

Methods	P	R	F1
UTOS	55.45	46.76	50.35
RankCP	**75.08**	43.90	55.31
ECPE-MLL	70.45	47.76	56.88
UECA-Prompt	69.52	54.66	61.14
UECA-Prompt (m2m)	73.92	56.30	63.45
GPT3.5 DECC $_{4-shot}$	63.93	**73.39**	**68.34**

Table 2: Results on multi-pair extraction scenarios. UECA(m2m) is the special-designed UECA for multi-pair extraction. Best results are marked in **bold**.

5.5 Ablation study

To reveal the specific effect of each step in DECC, we experimented different variants of DECC on GPT3.5 by removing each step respectively [5]. The experimental results are shown in Table 3, from which we see that all the components make positive

[5]Necessary changes are added to corresponding prompt for process connection when removing each step

contributions to the overall performance. The role of each step is as following:

- The performance drop caused by removing the recognizing step indicates its necessity. This is because the subsequent steps introduce many noisy clauses without the initial restriction of emotional keyword in the first step.
- Removing the locating step will result in unrelated emotional keywords entering into the analysis step, which might cause explosive emission of reasoning chains. This would produce many redundant pairs in output, which caused the decline of the precision.
- DECC suffers a loss of 7.37 F1 score without the analyzing step. The reason is that LLM directly extracts the cause clauses in an end-to-end manner, thereby losing the reasoning ability from COT.
- After removing the summarizing step from DECC, the performance remained almost unchanged. But the output of model will be difficult for human evaluation due to the lack of formatting process.

Methods	P	R	F1
DECC	61.54	49.76	55.03
w/o extracting	50.41	54.46	52.36 $\downarrow_{2.67}$
w/o locating	47.06	53.85	50.22 $\downarrow_{4.81}$
w/o analysing	51.52	44.35	47.66 $\downarrow_{7.37}$
w/o summarizing	59.63	49.62	54.17 $\downarrow_{0.86}$

Table 3: Ablation experiment results on Chinese dataset

5.6 Performance on Different LLMs

In this subsection, we test our proposed DECC framework on different LLMs. Besides GPT3.5, we also conduct experiments on both Chinese and English datasets using LLAMA (Touvron et al., 2023) and ChatGLM[6] respectively. As outlined in Table 4, we can see that DECC outperforms naive prompt method significantly across all LLMs with different parameter scales, which illustrates the robustness and universality of our approach.

5.7 Scaling the Demonstration Number

We investigate the scaling effect of demonstration number for DECC with GPT3.5-turbo. The experiments are conducted on Chinese dataset (Xia and Ding, 2019) and English dataset by varying the demonstrations numbers under the input token limit. We should select the appropriate configurations based on the trade-off between performance and computational cost.

[6]https://chatglm.cn/

	Prompt	DECC	Gain
GPT3.5 (Ch.)	50.82	56.73	+11.63%
ChatGLM (Ch.)	44.73	50.51	+12.92%
GPT3.5 (En.)	40.68	43.84	+7.77%
LLaMA (En.)	39.72	42.55	+7.12%

Table 4: F1 results of naive prompt and DECC on different LLM bases. 'En.' and 'Ch.' refer to the English and Chinese dataset. 'Gain' indicates the increase ratio.

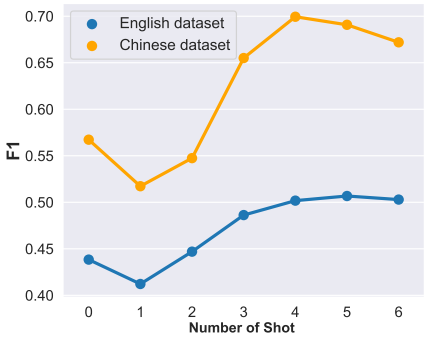

Figure 3: Results under different ICL shot numbers on Chinese dataset and English dataset

As depicted in Figure 3, the trends of the two curves are generally consistent. There is a sudden drop of performance at the point of one-shot. A possible reason behind is that LLM could generate output solely based on its comprehending ability under zero-shot settings, while in the ICL one-shot scenario, LLM might be misled by the only demonstration. For instance, if this demonstration belongs the type of single-pair extraction, then LLM will tend to forcibly extract only one pair for each document, similarly, a multi-pair demonstration may lead to an all-multi-pair extracting phenomenon. As a result, the performance degrades severely. As the shot-number increases, the performance improves significantly and reached the top at four shots. After the shot-number exceeds five, the performance begins to gradually deteriorate. This phenomenon arises may due to the misguidance caused by too many demonstrations, leading the output content to become increasingly unstable.

Document: *1. Talking about the scene of parents clearing snow for her* *2. Susan's eyes have tears twinkling* *3. her parents are more than 60 years old* *4. 30 miles of the mountain road* *5. she said* *6. when she stepped on the back road swept by them* *7. the heart is not only moved* *8. more is regret for not doing filial piety at home*	**Document:** *1. To prevent Ellie from changing her mind* *2. Mike picks her up every day night after school* *3. for this kind of entanglement* *4. Susan had regarded it as a totally trouble* *5. so she resolutely broke up with jack* *6. and asked her father to pick her up every day* *7. His girlfriend just broke up with him* *8. this fact made Mike feel very bitter*
Ground truth: Emotion: *more is regret for not doing filial piety at home* Cause: *more is regret for not doing filial piety at home* Clause Pair: $(8, 8)$ **Output:** *1.Emotion - the heart is not only moved* *Reason - parents clearing snow for her* *2.Emotion - more is regret for not doing filial piety at home* *Reason - more is regret for not doing filial piety at home*	**Ground truth:** Emotion: *this fact made Mike feel very bitter* Cause: *His girlfriend just broke up with him* Clause Pair: $(8, 7)$ **Output:** *Emotion - made Jack feel very bitter (clause 8)* *Reason - so she resolutely broke up with jack (clause 5)* (clause 5 and clause 7 both refer to the cause of 'breaking up')

Table 5: Two typical label error examples in English dataset.

5.8 Case Study[7]

Table 6 presents a sample from the English dataset, illustrating ground truth labels and two answers generated by DECC 4-shot. While Output A may not be an exact match, it aligns with the practical requirements of our application. Although Output B does not align with the standard format, it still accurately identifies the cause clause. Therefore, both of them could be considered correct, highlighting the importance of introducing human evaluation.

During the evaluation, we also discover some kinds of label error in the datasets. Here we display two most common ones, as presented in Table 5. In the left column, there are actually two obvious emotion-cause pairs in the document, whereas the gold truth only contains one pair. LLM with DECC tends to extract both pairs, resulting in an inevitable drop in precision. For the right column, Clause 4 and Clause 7 convey almost identical meanings. While either could be chosen as the cause clause, in most cases, only the one closer to the emotion clause tends to be labeled as the cause clause. This further highlights how traditional methods can benefit from position bias. These examples illustrate the limitations of the current benchmark datasets. On the other hand, whey also underscore the significant potential of our proposed DECC.

Document: *1.Hertzfeld took great pride that* *2.they achieved the functionality solely using software* *3.he replied* *4.We don't have any special hardware for it* *5.Gates insisted special hardware necessary that* *6.it can move the cursor that way*
Ground tructh: Emotion clause: *Hertzfeld took great pride that* Cause clause: *they achieved the functionality using software* Clause Pair: $(1, 2)$ **Output A:** *Emotion clause -"took great pride"* *Cause clause -"achieve the functionality using software"* **Output B:** *Emotion: pride (Clause 1)* *Explanation: The reason for the emotion of pride is the achievement of being able to achieve their functionality solely using software, as stated in clause 2*

Table 6: Examples of LLM producing "also correct but not exactly matching" results.

6 Conclusion

The primary strategy of DECC framework is to logically break down the ECPE task into a series of subtasks and solve them in sequence. The resolution of each subtask is facilitated by the answers of previous subtasks. With the decomposed chain-of-thought paradigm and meticulously crafted prompts, DECC is able to significantly enhance the reasoning ability of large language models on the ECPE task, while concurrently mitigating challenges faced by traditional methods. Experiments conducted on three benchmark datasets have demonstrated the effectiveness and robustness of DECC. In the future, we aim to explore more efficient and robust demonstration selection methods to further enhance its performance.

[7]Please refer to our Appendix C for more cases.

7 Limitations

Since DECC decomposes the ECPE task into multiple sub-problems, it will increase the inference time of large language models. Besides, we did not test DECC on GPT4 due to the cost constraint.

References

Matthias Berking, David Ebert, Pim Cuijpers, and Stefan G Hofmann. 2013. Emotion regulation skills training enhances the efficacy of inpatient cognitive behavioral therapy for major depressive disorder: a randomized controlled trial. *Psychotherapy and psychosomatics*, 82(4):234–245.

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

Ying Chen, Wenjun Hou, Shoushan Li, Caicong Wu, and Xiaoqiang Zhang. 2020. End-to-end emotion-cause pair extraction with graph convolutional network. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 198–207.

Ying Chen, Sophia Yat Mei Lee, Shoushan Li, and Chu-Ren Huang. 2010. Emotion cause detection with linguistic constructions. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 179–187.

Zifeng Cheng, Zhiwei Jiang, Yafeng Yin, Na Li, and Qing Gu. 2021. A unified target-oriented sequence-to-sequence model for emotion-cause pair extraction. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:2779–2791.

Zifeng Cheng, Zhiwei Jiang, Yafeng Yin, Hua Yu, and Qing Gu. 2020. A symmetric local search network for emotion-cause pair extraction. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 139–149.

Georgiana Craciun and Kelly Moore. 2019. Credibility of negative online product reviews: Reviewer gender, reputation and emotion effects. *Computers in Human Behavior*, 97:104–115.

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

Jiayuan Ding and Mayank Kejriwal. 2020. An experimental study of the effects of position bias on emotion cause extraction. *arXiv preprint arXiv:2007.15066*.

Zixiang Ding, Rui Xia, and Jianfei Yu. 2020a. Ecpe-2d: Emotion-cause pair extraction based on joint two-dimensional representation, interaction and prediction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3161–3170.

Zixiang Ding, Rui Xia, and Jianfei Yu. 2020b. End-to-end emotion-cause pair extraction based on sliding window multi-label learning. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, pages 3574–3583.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, and Zhifang Sui. 2022. A survey for in-context learning. *arXiv preprint arXiv:2301.00234*.

Chuang Fan, Chaofa Yuan, Lin Gui, Yue Zhang, and Ruifeng Xu. 2021. Multi-task sequence tagging for emotion-cause pair extraction via tag distribution refinement. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:2339–2350.

Susan T Fiske and Shelley E Taylor. 1991. *Social cognition*. Mcgraw-Hill Book Company.

Yao Fu, Hao Peng, Ashish Sabharwal, Peter Clark, and Tushar Khot. 2022. Complexity-based prompting for multi-step reasoning. *arXiv preprint arXiv:2210.00720*.

Ridong Han, Tao Peng, Chaohao Yang, Benyou Wang, Lu Liu, and Xiang Wan. 2023a. Is information extraction solved by chatgpt? an analysis of performance, evaluation criteria, robustness and errors. *arXiv preprint arXiv:2305.14450*.

Ridong Han, Tao Peng, Chaohao Yang, Benyou Wang, Lu Liu, and Xiang Wan. 2023b. Is information extraction solved by chatgpt? an analysis of performance, evaluation criteria, robustness and errors. *arXiv preprint arXiv:2305.14450*.

Jie Huang and Kevin Chen-Chuan Chang. 2022. Towards reasoning in large language models: A survey. *arXiv preprint arXiv:2212.10403*.

Tushar Khot, Harsh Trivedi, Matthew Finlayson, Yao Fu, Kyle Richardson, Peter Clark, and Ashish Sabharwal. 2022. Decomposed prompting: A modular approach for solving complex tasks. *arXiv preprint arXiv:2210.02406*.

Junlong Liu, Xichen Shang, and Qianli Ma. 2022. Pair-based joint encoding with relational graph convolutional networks for emotion-cause pair extraction. *arXiv preprint arXiv:2212.01844*.

Pan Lu, Liang Qiu, Kai-Wei Chang, Ying Nian Wu, Song-Chun Zhu, Tanmay Rajpurohit, Peter Clark, and Ashwin Kalyan. 2022. Dynamic prompt learning via policy gradient for semi-structured mathematical reasoning. In *The Eleventh International Conference on Learning Representations*.

Shuofei Qiao, Yixin Ou, Ningyu Zhang, Xiang Chen, Yunzhi Yao, Shumin Deng, Chuanqi Tan, Fei Huang, and Huajun Chen. 2022. Reasoning with language model prompting: A survey. *arXiv preprint arXiv:2212.09597*.

Chengwei Qin, Aston Zhang, Zhuosheng Zhang, Jiaao Chen, Michihiro Yasunaga, and Diyi Yang. 2023. Is chatgpt a general-purpose natural language processing task solver? *arXiv preprint arXiv:2302.06476*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. 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 3982–3992.

Zhihong Shao, Yeyun Gong, Yelong Shen, Minlie Huang, Nan Duan, and Weizhu Chen. 2023. Synthetic prompting: Generating chain-of-thought demonstrations for large language models. *arXiv preprint arXiv:2302.00618*.

KaShun Shum, Shizhe Diao, and Tong Zhang. 2023. Automatic prompt augmentation and selection with chain-of-thought from labeled data. *arXiv preprint arXiv:2302.12822*.

Aaditya Singh, Shreeshail Hingane, Saim Wani, and Ashutosh Modi. 2021. An end-to-end network for emotion-cause pair extraction. *arXiv preprint arXiv:2103.01544*.

Hoong-Cheng Soong, Norazira Binti A Jalil, Ramesh Kumar Ayyasamy, and Rehan Akbar. 2019. The essential of sentiment analysis and opinion mining in social media: Introduction and survey of the recent approaches and techniques. In *2019 IEEE 9th symposium on computer applications & industrial electronics (ISCAIE)*, pages 272–277. IEEE.

Xinxin Su, Zhen Huang, Yunxiang Zhao, Yifan Chen, Yong Dou, and Hengyue Pan. 2023. Recent trends in deep learning based textual emotion cause extraction. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.

Xiaofei Sun, Linfeng Dong, Xiaoya Li, Zhen Wan, Shuhe Wang, Tianwei Zhang, Jiwei Li, Fei Cheng, Lingjuan Lyu, Fei Wu, et al. 2023. Pushing the limits of chatgpt on nlp tasks. *arXiv preprint arXiv:2306.09719*.

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

Somin Wadhwa, Silvio Amir, and et al. 2023. Revisiting relation extraction in the era of large language models. *arXiv preprint arXiv:2305.05003*.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.

Zengzhi Wang, Qiming Xie, Zixiang Ding, Yi Feng, and Rui Xia. 2023. Is chatgpt a good sentiment analyzer? a preliminary study. *arXiv preprint arXiv:2304.04339*.

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

Penghui Wei, Jiahao Zhao, and Wenji Mao. 2020. Effective inter-clause modeling for end-to-end emotion-cause pair extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3171–3181.

Rui Xia and Zixiang Ding. 2019. Emotion-cause pair extraction: A new task to emotion analysis in texts. *arXiv preprint arXiv:1906.01267*.

Bo Xu, Hongfei Lin, Yuan Lin, and Kan Xu. 2021. Two-stage supervised ranking for emotion cause extraction. *Knowledge-Based Systems*, 228:107225.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*.

Xiaopeng Zheng, Zhiyue Liu, Zizhen Zhang, Zhaoyang Wang, and Jiahai Wang. 2022. Ueca-prompt: Universal prompt for emotion cause analysis. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 7031–7041.

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc Le, et al. 2022. Least-to-most prompting enables complex reasoning in large language models. *arXiv preprint arXiv:2205.10625*.

A Dataset statistics

Statistics	Chinese	English	Rebalanced
Total documents	1,945	2,843	756
One pair doc	1,746	2,537	733
Multi-pairs doc	199	306	23
Total pairs	2,167	3,215	780

Table 7: The statistics of three datasets

B Implementation Details

For GPT3.5-turbo with naive prompt, we keep exactly the same settings as reported in (Wang et al., 2023). For DECC, the tunable hyperparameters include temperature, top-K, penalty factor. These hyperparameters are manually tuned on each LLM with hold-out validation. The results reported in our experiments are all based on the average F1 score of 5 random runs on the test set. During all the experiments. We employed 5 annotators for human evaluation. For each instance, we consider it as correct only when 3 or more annotators agreed that the answer generated by LLM is correct.

For GPT3.5-turbo and ChatGLM, we obtain corresponding answers through official API. Since LLAMA 2 has not provided an official API, we utilized a third-party API[8]. The details about model vision, price and cost are reported in Table 10.

Method	Model	Temperature	Top_K	Penalty
Prompt	GPT3.5	0	1^*	1^*
	ChatGLM	0	1^*	-
	LLaMA	0	1^*	1.15^*
DECC	GPT-3.5	0.7	1	0.3
	ChatGLM	0.7	0.7	-
	LLaMA-2	0.7	1	0.3

Table 8: Hyperparameters of LLMs in our experiments. "*" denotes default setting. "-" means that ChatGLM API does not provide penalty adjustment.)

C More Case studies

In some cases, the DECC framework, especially under zero-shot settings, tends to output underlying causes of emotions rather than direct causes. Table 9 shows an example. Although these underlying causes are considered as wrong answer under our human evaluation due to the confliction with the ground truth label, this kind of answers is valuable in many practical application scenarios, which also demonstrates the potential of DECC with LLM. A complete example of DECC workflow is shown in Table 11

[8]https://replicate.com/

Document:
1.Last week
2.outside the burn surgery Department of the Hospital
3.the reporter met Mr. Shi
4.He regretted the fire in his home
5.If I didn't leave the gas tank by the door
6.Even if there's an electric fire in the house
7.My wife and my mother will be fine
8.And now they are both burned
7.It's all my fault
Ground tructh:
Emotion clause: *He regretted the fire in his home*
Cause clause: *He regretted the fire in his home*
Clause Pair: $(4, 4)$
Output:
Emotion keyword: *Regret*
Emotion clause: *4.He regretted the fire in his home*
Analysis: *Mr.Shi feels regretful to the fire in his house, if he didn't put the gas tank by the door, the injuries of his wife and mother wouldn't be like that, so the reason for regret is putting the gas tank*
Cause clause: *5.If I didn't leave the gas tank by the door*

Table 9: Underlying cause

D The Specific Content of the Prompts

We propose the DECC , aiming at providing a effective multi-stage framework for leveraging LLM for the ECPE task. However, the prompts of different LLMs settings do not have to be strictly consistent. Since LLMs are pretrained on different corpora and sensitive to different prompts. The performance of identical prompts may vary greatly across different LLMs. We provide our English and Chinese prompt settings used in our experiments in Figure4 and Figure 5, respectively. There are subtle differences between them.

Model	Version	Provider	Pricing	Total cost
GPT-3.5 Turbo	175B - July 6 version	OpenAI	USD 0.0015(in) 0.002(out) / 1K tokens	USD 250
ChatGLM	130B - STD	ZhiPu	CNY 0.005 / 1K tokens	CNY 32
LLaMA-2	70B - chat- HF	Third-Party	USD 0.0032 / second	USD 260

Table 10: Details about LLM API calls.

Document:
1. Mike was angered by this
2. after thinking back to the grievance and pain of this period
3. He quickly took out the knife from his trouser pocket
4. Stabbed Susan four times in the chest
5. When Susan's father woke up from his shock
6. Mike had run a long way
7. although Susan's father sent her to the hospital as quickly as possible
9. Susan suffered a severe chest injury of hemopneumothorax
10. because of the deep wound by knife
Ground tructh:
Emotion clause: *When Susan's father woke up from his shock*
Cause clause: *Stabbed Susan four times in the chest*
Clause Pair: $(5, 4)$

Subtask 1:
System:
Context : + $[Document]$ + According to the given context, each number at the beginning of line represents a clause, complete the following tasks, do not output uncertain answer:
User:
Please recognize emotions referred from the given context
LLM output:
anger, grievance and pain, shock, anxiety

Subtask 2:
User:
Locate the clauses where emotions occur, we call them emotion clauses, abandon those emotions that have no origin clause
LLM output:
Emotion clauses:
"Mike was angered by this"
" after thinking back to the grievance and pain of this period"
" When Susan's father woke up from his shock"

Subtask 3:
User:
Based on these emotion clauses, analyze why each one could happen, step by step
LLM output:
1. Emotion: anger
- Mike feel angry may because of certain matters caused by Susan.
2. Emotion: grivance and pain
- Mike thought back to this period, this leads the change of his emotion.
3. Emotion: shock
- Susan's father got shock because of seeing Mike stabbed Susan four times in the chest, this make him very shocked.

Subtask 4:
User:
For each emotion, select a most probable clause to be its cause, and output the pair of emotion clause and cause clause
LLM output:
1. Emotion: anger
- Emotion clause: 1. Mike was angered by this
- None explicitly mentioned in the given context
2. Emotion: grievance and pain
- Emotion clause: 2. after thinking back to the grievance and pain of this period
- None explicitly mentioned in the given context
3. Emotion: shock
- Emotion clause: 5. When Susan's father woke up from his shock
- Cause clause: 4. Stabbed Susan four times in the chest

Table 11: An example of common DECC workflow on LLaMA-2 70B-chat-HF

English ECPE task

Prompts on LLaMA-2 70B-chat-HF

Step1. please extract most obvious emotions referred from the given context, just output words among given words, do not make up uncertain emotion
Step2. locate the clauses where the emotions literally occur, we call them emotion clauses
Step3. based on those emotions, analyses why each emotion could happen, step by step
Step4. each of the emotion, select a most probable clause to be its explanation , and output the pair of emotion clause and reason clause by number

Prompts on GPT-3.5-turbo

Step1. please extract emotions referred from the given context, just use given words to output, do not output uncertain emotion
Step2. locate the clauses where the emotions literally occur, we call them emotion clauses, if an emotion have no origin clause, then this emotion should be pruned
Step3. based on those emotions, analyze why each emotion could happen, step by step
Step4. for each of the emotion, select a most probable clause to be its reason , and output the pair of emotion clause and reason clause by number, do not output uncertain pairs

Figure 4: The prompt settings for English dataset.

Chinese ECPE task

Prompts on ChatGLM-STD

Step1. 抽取其中明显的情绪情感，只能输出存在于原文的情绪词，禁止输出非情绪词和无关内容
Step2. 找到每个情绪所在的那一个子句，如果子句没有包含情绪词则不能选取
Step3. 一步一步分析所选情绪的产生原因，找到原因子句
Step4. 为每个情绪输出最有可能的情绪子句和原因子句对，禁止输出不确定的内容

Prompts on GPT-3.5-turbo

Step1. 抽取其中表达最明显的情绪，只能输出原文中的内容，禁止概括不存在的情绪
Step2. 每个情绪对应一个情绪子句，找出情绪子句，这个子句必须是包含上述情绪的，不包括情绪词的子句不要输出
Step3. 根据这个情绪子句，一步一步分析其中情感产生原因
Step4. 选出最有可能的原因子句，输出置信度最高的情绪和原因子句对（带序号），在文中找不到具体的原因子句则不输出

Figure 5: The prompt settings for Chinese dataset.