Enhancing Zero-shot Emotion Perception in Conversation through the Internal-to-External Chain-of-Thought

Anonymous ACL submission

Abstract

An excellent emotional dialogue model needs to rapidly adapt to new scenarios and perform emotion analysis to meet rapidly changing demands. Therefore, enhancing the model's zero-shot emotion-related capabilities in the dialogue domain has become a new challenge. However, current research shows that large language models (LLMs) perform poorly in zero-shot emotion-related tasks and the Emotion Recognition in Conversations (ERC) task alone doesn't comprehensively reflect the model's emotion understanding capabilities. In this paper, we propose an Emotion Perception in Conversation (EPC) task, which includes both ERC and Emotion Inference in Conversations (EIC), to evaluate the model's emotion perception capabilities in dialogue comprehensively. We propose an Internal-to-External Chain-of-Thought (IoECoT) method for the EPC task. This is a plug-and-play method that first extracts personality information of the dialogue participants from the dialogue history as internal factors influencing emotions, and then uses the sentiment polarity of the historical utterances as external factors. Finally, emotions are perceived by combining internal and external factors. Additionally, we conduct extensive experiments, and the results show that IoECoT significantly outperforms other baselines across multiple models and datasets, demonstrating that IoECoT effectively enhances the emotion perception capabilities of LLMs in zero-shot scenarios.

1 Introduction

The use of emotional information can effectively improve the interaction effect of dialogues and enhance emotional resonance, playing a crucial role in guiding the construction of high-quality dialogue systems (Liu et al., 2021; Ma et al., 2020). To utilize emotional information to enhance dialogue systems, it is first necessary to analyze the emotions involved in the dialogue. Researchers have conducted extensive explorations in emotion analysis, mainly focusing on two tasks: emotion recognition in conversation (ERC) (Poria et al., 2018) and emotion inference in conversations (EIC) (Li et al., 2021a).

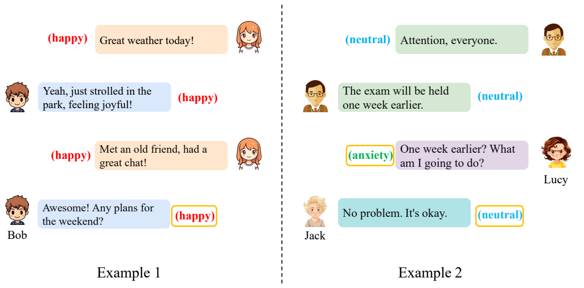

(happy)
Great weather today!
Yeah, just strolled in the park, feeling joyful!
(happy)
(happy)
Met an old friend, had a great chat!
Awesome! Any plans for the weekend?
(happy)
Bob
Example 1
(neutral)
Attention, everyone.
The exam will be held one week earlier.
(neutral)
(anxiety)
One week earlier? What am I going to do?
Lucy
No problem. It's okay.
(neutral)
Jack
Example 2

Figure 1: Two examples showing that emotion feature and personality.

ERC is a traditional task in the field of Natural Language Processing (NLP), aimed at identifying the emotions of known utterances in dialogue, focusing on current emotional states. In contrast, EIC is an emerging task that aims to infer the emotional reactions of dialogue participants to the utterances, focusing on future emotional states. Previous research (Song et al., 2022; Li et al., 2021b) mostly studied these two tasks separately, concentrating on improving the model's capability in a single aspect, which could lead to a loss of capability in another task. However, a high-quality dialogue system requires the model to consider both the current and future emotional states of the users to better serve them. Therefore, drawing on the definition method by (Zhao et al., 2024; Mayer et al., 2001), we refined the Emotion Perception task. We merged ERC and EIC into an Emotion Perception in Conversation (EPC) task to study how to comprehensively enhance the model's ability of emotion.

Recently, Large Language Models (LLMs) have demonstrated significant capabilities in a variety of tasks (Zhao et al., 2023) within Natural Language Processing (NLP). However, when dealing with more complex dialogue texts, researchers have found through a series of evaluations (Amin et al., 2023; Lian et al., 2024) that LLMs perform poorly in emotion analysis and recognition in zero-shot settings. Therefore, improving the performance of LLMs in zero-shot EPC task has become a new challenge. In EPC task, the reasons behind emotions are diverse (ROLLS, 2005), and solely relying on the dialogue context itself is insufficient to fully perceive the emotions within it. This brings us to our key questions: **What information is effective for EPC task? How can this information be accurately obtained?**

Emotion is characterized by two fundamental features: persistence (Mitchell, 2022) and contagiousness (Dimitroff et al., 2017). Persistence pertains to the continuance of emotional states, while contagion involves the transmission of an emotional state among individuals. In dialogue, participants' emotional states may remain unchanged or be influenced by others' emotions. This historical emotional state is an external factor in the generation of emotions. Therefore, we can obtain information about the historical emotional states through the dialogue context and use these two characteristics of emotions to assist in EPC task.

Additionally, individuals with different personalities have varying sensitivities to external factor stimuli (Genova and Gazzillo, 2018). People with different personalities may exhibit different emotional reactions to the same emotional stimulus (Resseguier et al., 2016). Therefore, personality information represents an individual's sensitivity to emotional stimuli and is an internal factor in the generation of emotions. In Figure 1, in the first example, we can clearly see that Bob's emotion and historical emotional states both belong to positive emotions, which is consistent with the characteristics of emotion. Conversely, in the second example, individual Jack demonstrates greater emotional stability when faced with stress and challenges, while individual Lucy is more inclined to experience emotions such as anxiety, nervousness, and agitation. It is evident that people with different personalities have varying sensitivities to emotions, and the emotions generated in the same scenario also differ. Therefore, for the EPC task, historical emotional state information and personality information can both positively contribute to its effectiveness.

Secondly, how can we accurately obtain both the historical emotional state information and personality information? Evidence from previous studies (Nguyen et al., 2023) has confirmed that LLMs face fewer challenges in addressing coarse-grained tasks. Therefore, we consider the sentiment polarity of the historical utterance as the historical emotional states, thus obtaining relatively accurate information about the historical emotional states. In comparison to utilizing fixed personality categories for categorization, incorporating natural language to express personality in LLMs not only minimizes classification errors but also enhances the precision of dialogue information.

Human emotions are generated under the combined influence of internal and external factors (Imbir, 2013; Young and Suri, 2019). To achieve this, we utilize the Chain-of-Thought (COT) (Wei et al., 2022) to gradually extract these two types of information under zero-shot settings. Following the rules of emotion generation, we primarily consider internal factors and supplement with external factors, combining these two types of information from internal to external. In this way, we can simulate the emotional changes made in response to emotional stimuli represented by the historical emotional state (external factors) under the sensitivity dominated by the individual information of the dialogue participants (internal factors). This approach allows us not only to perceive emotions in the dialogue from a global perspective but also to ensure that emotion perception has user specificity.

In this work, we explore information that can facilitate LLMs in performing EPC tasks under zero-shot conditions. At the same time, the information is organized using the CoT structure for the EPC task. The contributions of this paper are summarized as follows:

- Through extensive exploratory research, we validate the correlation between the emotion and historical emotional state information. Additionally, we demonstrate the high adaptability of LLMs in the conversational domain for coarse-grained tasks, utilizing an analysis of statistical dialogue data.

- We propose a refined emotion perception task EPC to comprehensively enhance the emotion perception in conversation capabilities of LLMs. Based on this task, we introduce the **Internal-to-External Chain-of-Thought (IoECoT)**, a plug-and-play prompting method. This method combines personality information and historical emotional state

Evaluation of Emotional Realtion						
Dataset	Pervasive	Personal	Proximal	Sum	Total	Proportion
MELD	105	27	17	149	200	0.75
EmoryNLP	40	3	9	52	72	0.72
DailyDialog	532	23	58	613	741	0.83
IEMOCAP	32	5	3	40	51	0.78
Evaluation of Personality						
Dataset	Score:1	Score:2	Score:3	Score:4	Score:5	Average
MELD	7 / 9	23 / 16	130 / 35	34 / 33	6 / 7	3.05 / 3.13
EmoryNLP	8 / 2	11 / 2	16 / 7	36 / 5	1 / 3	3.15 / 3.10
DailyDialog	41 / 6	57 / 15	200 / 24	396 / 30	47 / 25	3.47 / 3.53
IEMOCAP	5 / 3	10 / 2	15 / 4	15 / 5	6 / 5	3.14 / 3.20

Table 1: Above: Sum represents the total number of dialogues containing the three types of relationships, while total represents the total number of dialogues in the test dataset. Below: Evaluation results are indicated by the number of dialogues with the same score. model evaluation results first, followed by human evaluation results.

information in a manner that organizes from internal factors influencing emotion generation to external factors, thereby improving the emotion perception in conversation of LLMs.

- Furthermore, we conduct extensive experiments on multiple datasets and base models, and the results show that our IoECoT can effectively enhance the emotion perception in conversation capabilities of LLMs. We discuss the potential of LLMs in the field of conversational emotions and provide key insights into EPC tasks under zero-shot conditions combined with IoECoT.

2 Related Work

Emotion Recognition in Conversation ERC, as a traditional task related to dialogue emotions, mainly focuses on the emotional state of the current utterance and has achieved many breakthrough advancements through research. When dealing with the complex relationships between characters and the order of dialogue in conversations, graph structures are often used to model the information interactions within the dialogue (Ghosal et al., 2019; Lee and Choi, 2021). Additionally, utilizing commonsense knowledge to understand the dialogue context (Zhong et al., 2019) has become a key focus in the study of ERC, enabling the acquisition of richer contextual information. Recently, with the rise of LLMs, researchers have begun to explore the use of fine-tuning these large models to build generative frameworks (Lei et al., 2023), thereby comprehensively enhancing the performance of ERC.

Emotion Inference in Conversations EIC, as a new task, primarily focuses on the future emotional states of dialogue participants, guiding the generation of dialogue responses that pay more attention to users' emotions. Currently, researchers employ different methods to generate knowledge of varying granularity (Li et al., 2021a,c) to address issues such as consistency in emotional state responses and knowledge integration strategies. Additionally, some studies are based on LLMs to enhance the relevance between knowledge and dialogue (Wang and Feng, 2023), thereby improving the performance of EIC.

LLMs and CoT The introduction of LLMs has provided new approaches to solving problems in zero-shot settings. Models such as GPT-3 (Brown et al., 2020), ChatGLM3 (Zeng et al., 2022), and LLaMA (Touvron et al., 2023) have achieved remarkable results in reasoning (Xi et al., 2023; Wang et al., 2022) problems. Currently, the construction of prompts and the use of CoT techniques are commonly adopted. In particular, CoT technology is widely applied to problem-solving in zero-shot settings. Depending on the problem-solving perspective, various CoT variants such as TreeCoT (Yao et al., 2023), AutoCoT (Zhang et al., 2022), MetaCoT (Zou et al., 2023) and THOR (Fei et al., 2023) have emerged. However, these methods are not well-suited for complex dialogue structures.

3 Pilot Study

3.1 Task Formulation

The EPC task consists of two parts: the ERC task, which perceives current emotions, and the EIC task, which perceives future emotions. Given a multi-turn dialogue $D = [(u_1, p_1), (u_2, p_2), \cdots, (u_n, p_n), p_{n+1}]$, where u_i represents the utterance of the i-th turn, p_i represents the participant of the i-th turn of the dialogue. For the ERC task, what we should do is to predict the emotion label e_i of utterance u_i. For the EIC task, we infer the possible emotion reaction e_{n+1} of the p_{n+1}, given that the utterance u_{n+1} is unknown. We collectively refer to the speaker p_i, whose discourse is to be identified in the ERC task, and the speaker p_{n+1} is to be inferred in the EIC task, as the target individual.

3.2 Verification Experiment

In the previous section, we discussed how we can leverage the persistence and contagiousness of emotion, the adaptability of coarse-grained tasks in LLMs, and the ability to extract personality from dialogue to collect the necessary information for emotion inference. In this section, we will present a series of experimental arguments to verify three conjectures.

Experiments on Emotional Features In simpler terms, the persistence and contagiousness of emotions indicate that the dialogue participants' emotion is influenced by the emotional state in the dialogue history. We examine three emotion relationships—Pervasive, Personal, and Proximal—to determine whether the emotional characteristics affect the dialogue participants' emotion in the test datasets. As presented in Table 1, Pervasive denotes that the dialogue participants' emotion aligns with the most frequent emotion in the dialogue history, Personal denotes that the dialogue participants' emotion corresponds to their own highest-frequency emotion in the dialogue history, and Proximal denotes that the dialogue participants' emotion aligns with the emotions of other dialogue participants in close proximity. We analyze the last utterance in the dialogue. When the utterance belongs to multiple relationships, we prioritize selecting a relationship according to the order of Pervasive, Personal, and Proximal. The results in the table clearly demonstrate that dialogues adhering to the three emotional relationships of emotional persistence and contagiousness constitute over 70% of each dataset. This highlights the significant role of emotional state information in the dialogue history, enabling the model to comprehend the dialogue.

Comparison of Task Adaptability We perform experiments on two different granularity tasks, namely coarse-grained polarity classification and fine-grained emotion classification. The polarity classification task involves categorizing utterances into neutral, positive, and negative categories, while the emotion classification task entails categorizing utterances into either 7 or 10 categories depending on the dataset. The accuracy of the coarse-grained task on each dataset is significantly higher than that of the fine-grained task, as depicted in Figure 2. This result demonstrates the strong adaptability of LLMs in classifying dialogues at a coarse-grained level. Consequently, we consider the polarity of dialogue history utterances as the historical emotional state information. This approach helps to reduce errors in emotion perception caused by inaccurate historical emotional states.

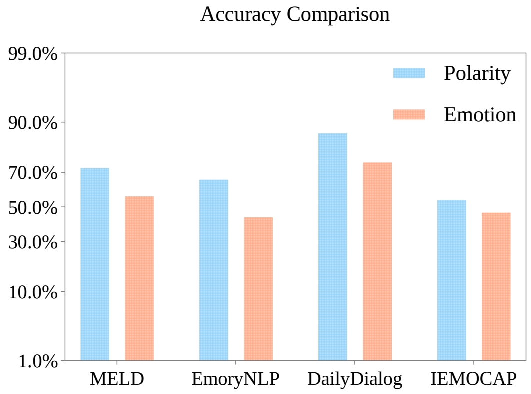
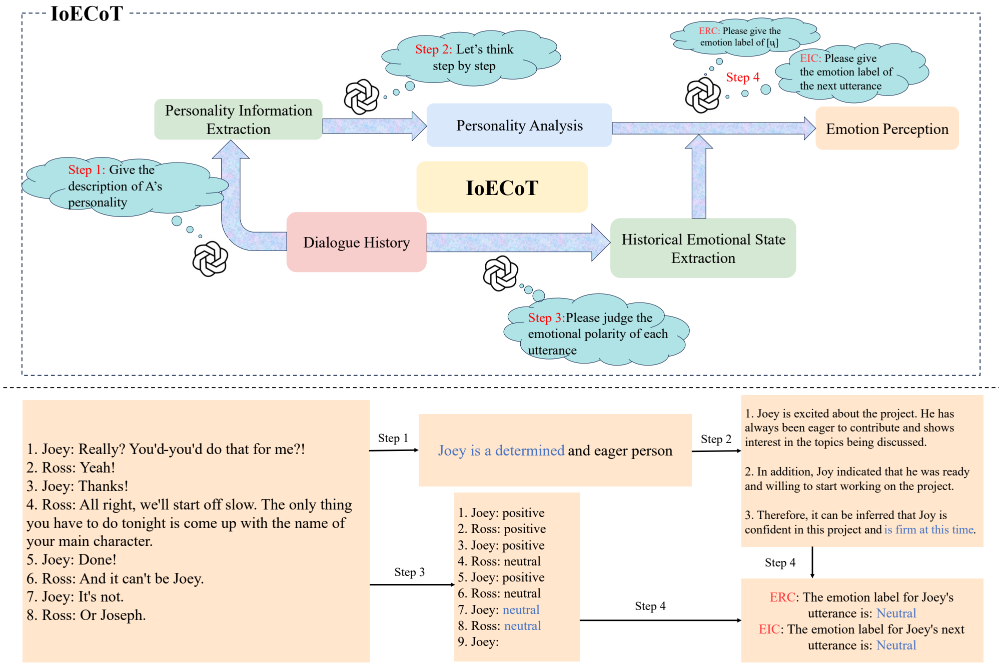


Figure 2: Comparison of Task Adaptability. The experimental framework involves using GPT3.5 as the basis, and conducting experiments on four datasets, with accuracy serving as the evaluation metric. The results of polarity classification are represented in blue, while the results of emotion classification are indicated in orange.

Evaluation of Personality in Dialogues To assess the level of personality portrayed in the dialogues, we utilize GPT-3.5 as an evaluator to quantify the extent of personality embodiment. This evaluation ensures that the model can extract relevant and significant personality information from the dataset dialogues. The scored rating ranges from one to five, with higher scores indicating a stronger reflection of the speaker's personality. Conversely, lower scores suggest the presence of more meaningless utterances that fail to capture the speaker's personality traits. At the same time, to verify the reliability of the model's evaluation, we select three volunteers to manually score a subset of the dataset in the same manner. According to Table 1, the datasets contain a majority of conversations with scores of 3 and 4. Additionally,


IoECoT
Step 2: Let's think step by step
ERC: Please give the emotion label of $[u_i]$
Step 4
EIC: Please give the emotion label of the next utterance
Personality Information Extraction
Personality Analysis
Emotion Perception
Step 1: Give the description of A's personality
IoECoT
Dialogue History
Historical Emotional State Extraction
Step 3:Please judge the emotional polarity of each utterance
Step 1
Step 2
Step 3
Step 4
Step 4

1. Joey: Really? You'd-you'd do that for me?!
2. Ross: Yeah!
3. Joey: Thanks!
4. Ross: All right, we'll start off slow. The only thing you have to do tonight is come up with the name of your main character.
5. Joey: Done!
6. Ross: And it can't be Joey.
7. Joey: It's not.
8. Ross: Or Joseph.

Joey is a determined and eager person

1. Joey is excited about the project. He has always been eager to contribute and shows interest in the topics being discussed.

2. In addition, Joy indicated that he was ready and willing to start working on the project.

3. Therefore, it can be inferred that Joy is confident in this project and is firm at this time.

1. Joey: positive
2. Ross: positive
3. Joey: positive
4. Ross: neutral
5. Joey: positive
6. Ross: neutral
7. Joey: neutral
8. Ross: neutral
9. Joey:

ERC: The emotion label for Joey's utterance is: Neutral
EIC: The emotion label for Joey's next utterance is: Neutral

Figure 3: Framework of IoECoT and an examples of IoECoT.

the average evaluation scores for each dataset are above 3.0. These evaluation results indicate that the dialogues in the dataset contain a relatively rich amount of personality information, which aligns with our goal of extracting personality traits. Furthermore, the model evaluation results are close to the manual evaluation results, proving the validity of the model evaluation results.

4 Methodology

In this section, we will introduce the Internal-to-External Chain-of-Thought framework. The framework, as depicted in Figure 3, comprises four steps: personality information extraction, personality analysis, historical emotional state extraction, and emotion perception. We will now provide a detailed description of how IoECoT facilitates emotion perception from internal to external.

4.1 Personality Information Extraction

Dialogues typically involve multiple participants, and if these participants are not clearly distinguished, the model will struggle to accurately identify the information relevant to the target individual. To address this issue, the dialogue history is standardized in the form of "speaker name: utterance". This standardization ensures that the model can effectively locate the utterances pertaining to the target individual. Research has demonstrated that LLMs are more susceptible to errors when generating lengthy output (Huang et al., 2023). To address this issue, we employ a restriction that requires generating "the most accurate, one-sentence short description". This approach promotes the generation of concise and accurate personality expressions. LLMs receive a uniformly formatted dialogue history as input. Prompts containing the target individual's name and task requirements are provided, allowing the model to generate a natural language representation of the target individual's personality based on the dialogue history.

4.2 Personality Analysis

To leverage the usefulness of personality information, we conduct an analysis of the acquired personality representation. Direct utilization of the personality representation cannot fully capture the target individual's emotional sensitivity in the context of the ongoing dialogue scenario. Thus, we draw inspiration from previous work (Kojima et al., 2022) and employ the strategy of "Let's think step by step" to interpret the personality information. The model receives a combination of dialogue and personality information, enabling the LLMs to provide a step-by-step explanation of how the addressee, with a particular personality, is influenced by the events unfolding in the dialogue context. Through a two-step process of personality information extraction and interpretation, the model gains an in-depth understanding of the internal factors that generate

Method		MELD	EmoryNLP	DailyDialog	IEMOCAP
		w-F1 / m-F1	w-F1 / m-F1	w-F1 / m-F1	w-F1 / m-F1
ChatGLM3	Direct Prompt	38.97 / **26.03**	22.04 / 16.70	20.69 / 17.78	15.34 / 13.25
	CoT	28.54 / 16.14	12.76 / 11.04	18.07 / 13.80	7.64 / 6.75
	Plan-and-Solve	30.37 / 16.23	6.57 / 4.66	12.85 / 6.90	10.33 / 9.91
	IoECoT	**40.25 /** 22.78	**23.42 / 17.20**	**29.97 / 20.23**	**21.69 / 18.10**
GPT3.5	Direct Prompt	43.80 / 38.98	31.93 / 25.55	29.55 / 16.95	21.35 / 19.28
	CoT	39.32 / 30.34	25.45 / 21.67	42.62 / 19.10	13.53 / 11.38
	Plan-and-Solve	39.65 / 30.77	25.96 / 21.54	40.48 / 20.68	11.84 / 9.76
	IoECoT	**57.15 / 50.91**	**34.26 / 27.38**	**45.33 / 22.64**	**23.61 / 20.77**
Claude-3	Direct Prompt	41.31 / 48.24	25.01 / 19.59	30.65 / 13.87	16.27 / 14.82
	CoT	27.97 / 30.74	27.29 / 23.91	23.46 / 16.21	13.76 / 13.49
	Plan-and-Solve	32.37 / 27.55	22.78 / 17.68	18.15 / 17.87	5.72 / 5.68
	IoECoT	**54.61 / 45.35**	**31.03 / 25.00**	**39.82 / 21.58**	**20.65 / 18.73**
Mixtral	Direct Prompt	42.03 / 33.19	22.96 / 20.23	46.83 / 24.27	18.04 / 16.33
	CoT	43.96 / 36.07	22.59 / 21.87	37.40 / 27.66	7.00 / 7.14
	Plan-and-Solve	34.22 / 26.12	28.45 / 25.24	31.64 / 28.08	12.82 / 12.99
	IoECoT	**59.01 / 48.05**	**32.69 / 27.79**	**59.76 / 37.55**	**22.48 / 20.92**

Table 2: The main results of IoECoT performing the ERC task on the test sets of four datasets. In the results, w-F1 represents the weighted F1 score, and m-F1 represents the macro F1 score. The best results are highlighted in bold.

emotions in the target individuals and obtains the sensitivity of the target individuals to emotional influences in each dialogue history scenario.

4.3 Historical Emotional State Extraction

The third step involves extracting the emotional states from the dialogue history. As previously mentioned, the historical utterances are classified based on their coarse-grained polarity. Each utterance is categorized as neutral, positive, or negative, and recorded in the format "speaker name: polarity". As the intensity of emotional affect is influenced by the dialogue interval, the proximity of utterances to each other correlates with the strength of the affect. To ensure an accurate representation of the intensity of affective influence, the sentiment results of all dialogue history utterances were arranged in the order of the conducted dialogues. Through this step, the model obtains the historical emotional states of external factors that generate emotions and acquires information on emotional stimuli in each dialogue history scenario.

4.4 Emotion Perception

Through the previous steps, the model acquires internal and external factors. Starting from the internal factors, LLMs analyze emotional stimuli in historical scenarios, which are external factors, according to the sequence of the dialogue. This analysis is guided by personality information and considers the emotional sensitivity of the target individuals, thus perceiving their emotional changes. Finally, the model derives the results of emotion perception.

In Figure 3, we illustrate an example that demonstrates the enhancement of EPC performance through IoECoT in a real task. We highlight task-relevant information in blue. Initially, we standardize the dataset by converting the eight utterances in the conversation into the format of "speaker name: utterance". Subsequently, we input them into the model. The initial step involves extracting the personality information of the target individual, Joey. The extracted result indicates that Joey is a determined individual. Based on this information, we can preliminarily infer that Joey's emotion is resistant to change and tends to be emotionally persistent. The second step of personality analysis, combinings the dialogue context to determine Joey's sensitivity to emotional stimuli in the historical dialogue scenario. After two intermediate steps, we can conclude that "is firm at this time." Therefore, Joey's emotions are not easily influenced in the historical dialogue scenario. The third step involves extracting the polarity of the utterance as historical emotional state information. The historical emotional state indicates that the emotion of the dialogue has shifted from previously positive to a neutral state, with emotional stimuli being not

Method		MELD	EmoryNLP	DailyDialog	IEMOCAP
		w-F1 / m-F1	w-F1 / m-F1	w-F1 / m-F1	w-F1 / m-F1
ChatGLM3	Direct Prompt	29.75 / 12.64	15.84 / 10.64	18.00 / 11.52	6.00 / 6.47
	CoT	30.01 / 11.22	11.29 / 6.96	14.08 / 14.10	7.35 / 9.04
	Plan-and-Solve	31.43 / 13.67	12.30 / 7.60	17.92 / 16.33	9.03 / 10.07
	IoECoT	**35.11 / 16.70**	**17.29 / 11.75**	**55.51 / 21.64**	**12.70 / 13.69**
GPT3.5	Direct Prompt	26.85 / 14.59	12.17 / 8.76	40.16 / 13.05	11.68 / 9.02
	CoT	33.44 / 13.86	8.99 / 5.44	43.22 / 15.24	7.91 / 7.14
	Plan-and-Solve	34.12 / 12.96	12.43 / 7.84	41.35 / 15.47	7.42 / 7.28
	IoECoT	**35.44 / 21.58**	**14.56 / 11.35**	**48.69 / 17.92**	**12.98 / 12.20**
Claude-3	Direct Prompt	16.77 / 13.60	15.91 / 10.68	26.52 / 17.48	17.53 / 15.70
	CoT	18.92 / 18.07	14.74 / 11.52	18.03 / 12.16	11.41 / 10.57
	Plan-and-Solve	18.36 / 21.64	18.05 / 14.08	22.15 / 21.27	19.00 / 15.71
	IoECoT	**20.35 / 22.26**	**20.31 / 16.09**	**30.58 / 21.62**	**27.84 / 24.72**
Mixtral	Direct Prompt	32.92 / 18.76	18.58 / 15.00	29.86 / 20.22	11.57 / 8.91
	CoT	26.69 / 18.70	18.67 / 13.57	18.21 / 19.30	9.58 / 7.65
	Plan-and-Solve	28.45 / 18.85	17.08 / 13.49	15.93 / 18.22	13.37 / 12.17
	IoECoT	**33.33 / 23.62**	**24.47 / 23.13**	**44.06 / 27.80**	**22.82 / 17.28**

Table 3: The main results of IoECoT performing the EIC task on the test sets of four datasets. In the results, w-F1 represents the weighted F1 score, and m-F1 represents the macro F1 score. The best results are highlighted in bold.

strong. Combined with personality information, Joey tends to maintain his own emotions. After obtaining the key information, the model analyzes the two tasks of EPC. For the ERC task, it determines the emotion of Joey's utterance as neutral based on his utterance. For the EIC task, in the unknown of an utterance, it judges Joey's possible emotional reaction to be neutral.

5 Experiment

5.1 Datasets

We mainly evaluate our model on four commonly used public dialogue datasets. **MELD** (Poria et al., 2018) is a multimodal dialogue dataset collected from Friends, containing seven emotions. Each dialogue involves multiple participants. **EmoryNLP** (Zahiri and Choi, 2018), also collected from Friends, focuses on pure text dialogues and uses a different emotion annotation method than the MELD dataset, containing seven emotions. **IEMOCAP** (Busso et al., 2008) is a multimodal dialogue dataset with a large number of dialogue turns, including nine emotions. Each dialogue involves two speakers. **DailyDialog** (Li et al., 2017) is a multi-round dialogue text dataset collected from various English dialogue practice content on English learning websites, including seven emotions. On average, each dialogue consists of eight turns.

5.2 Baselines and Models

We compared our proposed method with the existing zero-shot chain-of-thought approach. **Direct Prompt**: The use of natural language as a direct prompt for LLMs to accomplish specific tasks. **CoT** (Wei et al., 2022): The phrase "Let's think step by step" served as guidance for the LLMs to generate a sequence of intermediate steps automatically. This process enabled them to ultimately accomplish the intended task using the provided reasoning steps. **Plan-and-Solve** (Wang et al., 2023): It instructs LLMs to develop a problem-solving plan by using the prompt "Let's first understand the problem and devise a plan to solve it." Subsequently, LLMs are guided to execute the plan and solve the problem step by step.

We utilize ChatGLM3-6B (Du et al., 2022), GPT-3.5 [1], Claude-3 [2], and Mixtral 8x7B (Jiang et al., 2024) as baseline models. In our study, we utilize weighted F1 and Macro F1 as evaluation metrics. We set the temperature to 0 to ensure deterministic output. The experimental results are reported by computing the mean values over five runs.

5.3 Main Results

Table 2 and Table 3 present the performance of IoECoT in the two sub-tasks of EPC. The evalua-

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

[2] https://www.anthropic.com/claude

Method		MELD	EmoryNLP	DailyDialog	IEMOCAP
		w-F1 / m-F1	w-F1 / m-F1	w-F1 / m-F1	w-F1 / m-F1
ERC	IoECoT	**57.15 / 50.91**	**34.26 / 27.38**	**45.33 / 22.64**	**23.61 / 20.77**
	w/o personality	48.62 / 46.33	30.36 / 29.15	40.25 / 19.32	21.69 / 19.60
	w/o emotional state	50.32 / 46.22	31.44 / 29.36	39.03 / 20.12	21.57 / 18.13
EIC	IoECoT	**35.44 / 21.58**	**14.56 / 11.35**	**48.69** / 17.92	**12.98 / 12.20**
	w/o personality	33.21 / 19.57	11.47 / 11.26	47.28 / 19.30	10.24 / 11.02
	w/o emotional state	34.64 / 20.90	11.55 / 7.08	48.03 / **20.44**	11.95 / 11.35

Table 4: Ablation study on the four datasets. For the ablation studies, we selected complete samples from four datasets. We utilized the GPT-3.5 as baseline model. The best results are in bold.

tion results across four datasets show that IoECoT achieves the State-of-the-Art (SOTA) in weighted-F1 metrics across all datasets. Notably, it outperforms the strongest baseline method, Plan-and-solve, by a minimum of two percentage points on each dataset and exhibits a robust zero-shot capability. In the meantime, IoECoT demonstrates its ability to outperform other CoT methods on all models. This indicates that IoECoT exhibits a robust generalization capability across both the datasets and the models. The experimental results showcase the effectiveness of our proposed IoECoT framework in extracting information that enhances dialogue comprehension. Moreover, our framework adeptly organizes and utilizes this information to facilitate accurate EPC.

Moreover, it can be observed that IoECoT has achieved varying degrees of improvement across different datasets. This is because we employ a diverse range of dialogue datasets, which exhibit significant differences in dialogue scenarios, the number of emotion categories, dialogue turns, and the number of participants. These variations have led to different degrees of improvement. This also demonstrates that our IoECoT can adapt to complex and changing demands, showcasing good generalizability.

5.4 Ablation Study

Table 4 showcases the results of our ablation experiments on four datasets, utilizing GPT-3.5 as the underlying model. Through these experiments, we observe that the model's capacity is significantly diminished when both historical emotional state information and personality information are removed. This is because when only historical emotional states are considered, it results in the lack of internal factors that generate emotions. Consequently, the model's perception of emotions is governed by historical emotional states, making it impossible to determine the target individual's sensitivity to external factors, leading to errors in perception. Similarly, When only personality information is considered, it results in the lack of external factors that generate emotions. This weakens the model's understanding of external influences, binding emotions to personality and making them independent of historical emotional states, which is clearly inconsistent with human cognition. Therefore, only by allowing historical emotional state information and personality information to work together, enabling the model to gradually analyze the target individual's emotional sensitivity and emotional stimuli in the order of the dialogue, and reasoning the evolution of emotions from internal factors to external factors based on the information, can the performance of the model's emotion perception be improved.

6 Conclusion

In this paper, we introduce a novel chain-of-thought framework called IoECoT. Our framework aims to integrate and leverage emotional state information from dialogue history in combination with personality information, using an internal-to-external approach for information integration. Experimental results demonstrate that our proposed framework significantly enhances the ability of emotion perception in conversation, particularly in zero-shot scenario. We conduct a series of validation experiments to investigate the properties related to emotions and to showcase the extensive adaptability of LLMs in coarse-grained tasks. The effectiveness of the IoECoT demonstrates that incorporating historical emotional state information and personality traits contributes to the understanding of dialogue. This finding establishes a robust foundation for further research in the field of dialogue understanding.

7 Limitations

In this section, we acknowledge the following constraints in our study: (1) Our current exploration of the factors influencing emotions remains incomplete, and during the process of reasoning, new scenarios frequently arise, rendering the information we have gathered insufficient to support effective reasoning. Therefore, the next phase of our research aims to study the intrinsic mechanisms of emotion generation. (2) Despite our numerous attempts to mitigate the issue of instruction noncompliance in LLMs, instances still arise where the generated content is irrelevant, posing a hindrance to effective perception. Therefore, resolving the problem of instruction noncompliance in LLMs, alongside addressing phantom issues, will greatly enhance the performance of our model.

8 Ethics Statement

During the utilization of LLMs, we diligently scrutinize the prompts to safeguard against the generation of discriminatory and biased content. Additionally, while our models display proficiency in reasoning about human emotional responses, they do not actively intervene in human emotional communication.

References

Mostafa M Amin, Rui Mao, Erik Cambria, and Björn W Schuller. 2023. A wide evaluation of chatgpt on affective computing tasks. *arXiv preprint arXiv:2308.13911*.

Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Celikyilmaz, and Yejin Choi. 2019. COMET: Commonsense transformers for automatic knowledge graph construction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4762–4779, Florence, Italy. Association for Computational Linguistics.

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.

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

Yirong Chen, Xiaofen Xing, Jingkai Lin, Huimin Zheng, Zhenyu Wang, Qi Liu, and Xiangmin Xu. 2023. Soulchat: Improving llms' empathy, listening, and comfort abilities through fine-tuning with multi-turn empathy conversations. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1170–1183.

Ernest Davis and Gary Marcus. 2015. Commonsense reasoning and commonsense knowledge in artificial intelligence. *Communications of the ACM*, 58(9):92–103.

Stephanie J Dimitroff, Omid Kardan, Elizabeth A Necka, Jean Decety, Marc G Berman, and Greg J Norman. 2017. Physiological dynamics of stress contagion. *Scientific reports*, 7(1):6168.

Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. 2022. GLM: General language model pretraining with autoregressive blank infilling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 320–335, Dublin, Ireland. Association for Computational Linguistics.

Hao Fei, Bobo Li, Qian Liu, Lidong Bing, Fei Li, and Tat-Seng Chua. 2023. Reasoning implicit sentiment with chain-of-thought prompting. In *The 61st Annual Meeting Of The Association For Computational Linguistics*.

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

Federica Genova and Francesco Gazzillo. 2018. Personality organization, personality styles, and the emotional reactions of treating clinicians. *Psychodynamic psychiatry*, 46(3):357–392.

Deepanway Ghosal, Navonil Majumder, Soujanya Poria, Niyati Chhaya, and Alexander Gelbukh. 2019. Dialoguegcn: A graph convolutional neural network for emotion recognition in conversation. In *Conference on Empirical Methods in Natural Language Processing*.

Chao-Chun Hsu, Sheng-Yeh Chen, Chuan-Chun Kuo, Ting-Hao Huang, and Lun-Wei Ku. 2018. EmotionLines: An emotion corpus of multi-party conversations. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. 2023. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *arXiv preprint arXiv:2311.05232*.

Kamil K Imbir. 2013. Origins and source of emotion as factors that modulate the scope of attention. *Roczniki Psychologiczne*, 16(2):287–310.

Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088*.

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

Bongseok Lee and Yong Suk Choi. 2021. Graph based network with contextualized representations of turns in dialogue. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 443–455, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Shanglin Lei, Guanting Dong, Xiaoping Wang, Keheng Wang, and Sirui Wang. 2023. Instructerc: Reforming emotion recognition in conversation with a retrieval multi-task llms framework. *arXiv preprint arXiv:2309.11911*.

Aitor Lewkowycz, Anders Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay Ramasesh, Ambrose Slone, Cem Anil, Imanol Schlag, Theo Gutman-Solo, et al. 2022. Solving quantitative reasoning problems with language models. *Advances in Neural Information Processing Systems*, 35:3843–3857.

Dayu Li, Xiaodan Zhu, Yang Li, Suge Wang, Deyu Li, Jian Liao, and Jianxing Zheng. 2021a. Emotion inference in multi-turn conversations with addressee-aware module and ensemble strategy. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3935–3941, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Dayu Li, Xiaodan Zhu, Yang Li, Suge Wang, Deyu Li, Jian Liao, and Jianxing Zheng. 2021b. Enhancing emotion inference in conversations with commonsense knowledge. *Knowledge-Based Systems*, 232:107449.

Dayu Li, Xiaodan Zhu, Yang Li, Suge Wang, Deyu Li, Jian Liao, and Jianxing Zheng. 2021c. Enhancing emotion inference in conversations with commonsense knowledge. *Knowledge-Based Systems*, 232:107449.

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

Zaijing Li, Gongwei Chen, Rui Shao, Dongmei Jiang, and Liqiang Nie. 2024. Enhancing the emotional generation capability of large language models via emotional chain-of-thought. *arXiv preprint arXiv:2401.06836*.

Zheng Lian, Licai Sun, Haiyang Sun, Kang Chen, Zhuofan Wen, Hao Gu, Bin Liu, and Jianhua Tao. 2024. Gpt-4v with emotion: A zero-shot benchmark for generalized emotion recognition. *Information Fusion*, 108:102367.

Kevin Hsin-Yih Lin, Changhua Yang, and Hsin-Hsi Chen. 2008. Emotion classification of online news articles from the reader's perspective. In *2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, volume 1, pages 220–226. IEEE.

Jiacheng Liu, Alisa Liu, Ximing Lu, Sean Welleck, Peter West, Ronan Le Bras, Yejin Choi, and Hannaneh Hajishirzi. 2022. Generated knowledge prompting for commonsense reasoning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3154–3169, Dublin, Ireland. Association for Computational Linguistics.

Siyang Liu, Chujie Zheng, Orianna Demasi, Sahand Sabour, Yu Li, Zhou Yu, Yong Jiang, and Minlie Huang. 2021. Towards emotional support dialog systems. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3469–3483, Online. Association for Computational Linguistics.

Yukun Ma, Khanh Linh Nguyen, Frank Z Xing, and Erik Cambria. 2020. A survey on empathetic dialogue systems. *Information Fusion*, 64:50–70.

John D Mayer, Peter Salovey, David R Caruso, and Gill Sitarenios. 2001. Emotional intelligence as a standard intelligence.

Jonathan Mitchell. 2022. Affective persistence and the normative phenomenology of emotion.

Hoang H Nguyen, Ye Liu, Chenwei Zhang, TAO ZHANG, and S Yu Philip. 2023. Cof-cot: Enhancing large language models with coarse-to-fine chain-of-thought prompting for multi-domain nlu tasks. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2018. Meld: A multimodal multi-party dataset for emotion recognition in conversations. *arXiv preprint arXiv:1810.02508*.

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*.

Beverly Resseguier, Pierre-Majorique Léger, Sylvain Sénécal, Marie-Christine Bastarache-Roberge, and François Courtemanche. 2016. The influence of personality on users' emotional reactions. In *HCI in Business, Government, and Organizations: Information Systems: Third International Conference, HCIBGO 2016, Held as Part of HCI International 2016, Toronto, Canada, July 17-22, 2016, Proceedings, Part II 3*, pages 91–98. Springer.

EDMUND T ROLLS. 2005. What are emotions, why do we have emotions, and what is their computational basis in the brain? *Who Needs Emotions?*, pages 117–146.

Xiaohui Song, Longtao Huang, Hui Xue, and Songlin Hu. 2022. Supervised prototypical contrastive learning for emotion recognition in conversation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5197–5206.

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*.

Lei Wang, Wanyu Xu, Yihuai Lan, Zhiqiang Hu, Yunshi Lan, Roy Ka-Wei Lee, and Ee-Peng Lim. 2023. Plan-and-solve prompting: Improving zero-shot chain-of-thought reasoning by large language models. *arXiv preprint arXiv:2305.04091*.

Renxi Wang and Shi Feng. 2023. Global-local modeling with prompt-based knowledge enhancement for emotion inference in conversation. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 2120–2127, Dubrovnik, Croatia. Association for Computational Linguistics.

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*.

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.

Gloria Willcox. 1982. The feeling wheel: A tool for expanding awareness of emotions and increasing spontaneity and intimacy. *Transactional Analysis Journal*, 12(4):274–276.

Zhiheng Xi, Senjie Jin, Yuhao Zhou, Rui Zheng, Songyang Gao, Tao Gui, Qi Zhang, and Xuanjing Huang. 2023. Self-polish: Enhance reasoning in large language models via problem refinement. *arXiv preprint arXiv:2305.14497*.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models. *arXiv preprint arXiv:2305.10601*.

Gerald Young and Gaurav Suri. 2019. Emotion regulation choice: A broad examination of external factors. *Cognition and Emotion*.

Sayyed M Zahiri and Jinho D Choi. 2018. Emotion detection on tv show transcripts with sequence-based convolutional neural networks. In *Workshops at the thirty-second aaai conference on artificial intelligence*.

Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, et al. 2022. Glm-130b: An open bilingual pre-trained model. *arXiv preprint arXiv:2210.02414*.

Zhuosheng Zhang, Aston Zhang, Mu Li, and Alex Smola. 2022. Automatic chain of thought prompting in large language models. *arXiv preprint arXiv:2210.03493*.

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*.

Weixiang Zhao, Zhuojun Li, Shilong Wang, Yang Wang, Yulin Hu, Yanyan Zhao, Chen Wei, and Bing Qin. 2024. Both matter: Enhancing the emotional intelligence of large language models without compromising the general intelligence. *arXiv preprint arXiv:2402.10073*.

Peixiang Zhong, Di Wang, and Chunyan Miao. 2019. Knowledge-enriched transformer for emotion detection in textual conversations. 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 165–176, Hong Kong, China. Association for Computational Linguistics.

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*.

Anni Zou, Zhuosheng Zhang, Hai Zhao, and Xiangru Tang. 2023. Meta-cot: Generalizable chain-of-thought prompting in mixed-task scenarios with large language models. *arXiv preprint arXiv:2310.06692*.

A Manual Evaluation

The manual evaluation experiments in this study are conducted by two graduate students specializing in dialogue. They possess not only good English reading skills but also an in-depth understanding of the field, ensuring an accurate assessment of whether the dialogues contain personality. Additionally, these graduate students underwent relevant training before the evaluation to standardize the assessment criteria, ensuring the reliability and consistency of the evaluation results. Their professional background and evaluation capabilities provide a solid foundation for this research, guaranteeing the accuracy and credibility of the experimental results.

B Templates

Figure 4 and Figure 5 illustrate the prompt templates used in executing ERC and EIC tasks. The model gradually generates the required key information through these prompts, processes it for memory retention, and ultimately achieves the task results.

[Personality Information Extraction]
Give the most accurate one-sentence short description of [target individual]'s personality in the context of [target individual]'s utterances in the history of the dialogue.

[Personality Analysis]
Recognizing the emotion of [utterance] based on the personality of [target individual]. Let's explain step by step.

[Historical Emotional State Extraction]
Please judge the sentiment polarity of each utterance in the dialog history, noting that you can only choose from the following three categories [neutral, negative, positive].

[Emotion Perception]
Please give the emotion label of the [utterance] can only be chosen from [emotion candidates] and do not give the explanation.

Figure 4: Templates of ERC Task.

[Personality Information Extraction]
Give the most accurate one-sentence short description of [target individual]'s personality in the context of [target individual]'s utterances in the history of the dialogue.

[Personality Analysis]
Complete an emotion inference task to predict the emotion of speaker [target individual] in the next utterance of the dialogue, make inferences based on [target individual]'s personality. Let's explain step by step.

[Historical Emotional State Extraction]
Please judge the sentiment polarity of each utterance in the dialog history, noting that you can only choose from the following three categories [neutral, negative, positive].

[Emotion Perception]
Please give the emotion label of the next utterance can only be chosen from [emotion candidates] and do not give the explanation.

Figure 5: Templates of EIC Task.