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

Xingle Xu, Shi Feng(✉), Daling Wang, Yifei Zhang, and Xiaocui Yang

School of Computer Science and Engineering, Northeastern University, Shenyang, China
2490264@stu.neu.edu.cn,
{fengshi, wangdaling, zhangyifei, yangxiaocui}@cse.neu.edu.cn

Abstract. An ideal emotional dialogue system should have emotion perception capability in conversation across various new scenarios, thereby extracting user needs to guide dialogue generation. However, due to the lack of sufficient training data in new scenarios, improving the model's zero-shot emotion perception capability in conversation has become a new challenge. Moreover, current research mostly focuses on single tasks, which cannot comprehensively reflect the model's emotion perception ability. In this paper, we propose an Internal-to-External Chain-of-Thought method (IoECoT) for the emotion perception in conversation. First, the personality information of target user are extracted from the dialogue history as internal factors, while the polarity of utterances serves as external factors. Guided by internal factors, the model perceives emotions based on external factors. By evaluating the model's performance on both Emotion Recognition in Conversation (ERC) and Emotion Inference in Conversation (EIC), we comprehensively assess its emotion perception capabilities. We conduct extensive experiments, and IoECoT outperforms other baselines on multiple Large Language Models (LLMs) and four datasets, demonstrating that our approach enhances LLMs' emotion perception performance in zero-shot setting. Code is released at: https://github.com/betterfly123/IoECoT.

Keywords: Emotion Perception · Zero-shot · Chain-of-Thought · Emotion State · Personality.

1 Introduction

The use of emotional information can improve the interaction effect of dialogues and enhance emotional resonance, which is key to building high-quality dialogue systems [17,19]. To acquire emotional information, emotion perception in conversation is first required. Emotion Recognition in Conversation (ERC) [**?**] and Emotion Inference in Conversation (EIC) [13] are two emotion-related tasks in the dialogue domain. Although both involve emotion perception, the model capabilities they assess and the perspectives of emotion perception differ

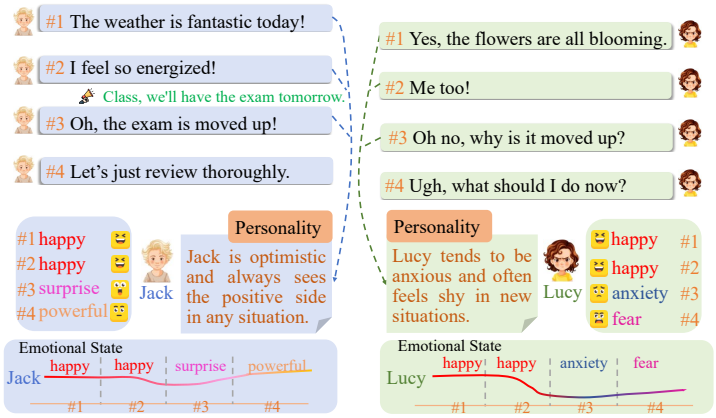


Fig. 1: The example of the role of emotion characteristics and personality.

greatly. ERC focuses on the user's past and current emotional states, requiring the model to have the ability to understand expressed emotions, while EIC predicts the user's future emotional states, emphasizing emotion reasoning abilities. Due to the complementary nature of these two tasks, we study both ERC and EIC to comprehensively improve the model's emotion perception capabilities in conversation.

As demand increases and new scenarios continuously emerge, data acquisition and training costs make it difficult for models to acquire emotion perception capabilities through training. Therefore, it becomes necessary to achieve emotion perception under zero-shot setting. Large language models (LLMs), with their vast parameters and extensive pre-training, possess far greater generalization and reasoning potential than smaller models [35,18]. Therefore, we commit to exploring how to unlock their potential in zero-shot setting to enhance the performance of emotion perception in conversation.

In zero-shot setting, we cannot rely on statistical patterns in the data to improve model performance; instead, we must extract useful information from the dialogue to assist the model in perception. This leads to our consideration: **Which information is effective for emotion perception? How can we correctly obtain this information?**

We first explore which information is effective for emotion perception. As shown in Figure 1, when the exam notification is broadcast, the optimistic Jack is briefly surprised but then begins to prepare positively, while the introverted Lucy experiences noticeable anxiety, though her mood improves under Jack's influence. The emotional changes of both Jack and Lucy are based on their original emotions, which demonstrates the persistence of emotions [20]. Lucy's mood improvement reflects the contagiousness of emotions [4]. This indicates that historical emotional states as the external factor influence emotions. Additionally, personality differences lead to different emotional responses to the

exam notification, showing that personality information as the internal factor also affects emotions. Therefore, historical emotional states and personality information may play a positive role in emotion perception. To accurately obtain historical emotional states and personality information, we extract the easily accessible polarity in the dialogue history as the historical emotional state, and use LLMs to express personality through natural language, avoiding classification errors caused by fixed personality categories.

After identifying the effective information and access methods for emotion perception, we integrate the Chain-of-Thought (CoT) technique [28]. In zero-shot setting, LLMs are used to progressively extract internal and external factors. Following the rules of emotion generation [9,31], we integrate these two emotional influencing factors from the internal to external, simulating the sensitivity of personality to emotional stimuli and the impact of historical emotional states on emotional changes. Finally, we consider the emotional fluctuations in the dialogue history in the order in which the dialogue unfolds, inferring the user's current emotional state and predicting potential future emotional state. This method not only enables a global perception of emotions in dialogue but also ensures that the emotion perception is user-specific. The contributions of this paper are summarized as follows:

(i) We simultaneously consider both the ERC and EIC tasks, conducting a more comprehensive study on the model's emotion perception capabilities in conversation. This provides a more complete perspective for emotion perception research in dialogue systems.

(ii) We propose a prompting method, the **Internal-to-External Chain-of-Thought (IoECoT)**. This method combines internal factors and external factors, integrating personality information and historical emotional states to enhance the emotion perception ability of LLMs in the conversation.

(iii) Furthermore, we conduct extensive experiments on multiple datasets and base models. The experimental results show that IoECoT has good generalizability and can effectively enhance the emotion perception capability of LLMs.

2 Related Work

2.1 Emotion Recognition in Conversation

ERC focuses on identifying the emotional state of the current utterance. To capture the relationships between participants and the sequential nature of dialogue, graph structures are employed to model information interactions within dialogues [7]. Moreover, leveraging commonsense knowledge has become essential for understanding dialogue context, providing richer contextual insights [36]. Recently, with the rise of LLMs, fine-tuning these models for generative frameworks has gained attention, further boosting ERC performance [12].

2.2 Emotion Inference in Conversation

EIC is a new task focusing on predicting the future emotional states of dialogue participants to generate emotion-aware responses. Researchers employ

various methods to generate knowledge of different granularity [13,14] to address issues like emotional consistency and knowledge integration. Some studies leverage LLMs to enhance the relevance between knowledge and dialogue [27], improving EIC performance. It is evident that previous studies primarily focus on the investigation of individual tasks related to either EIC or ERC, aiming to improve the performance of a single task. Furthermore, our research considers both tasks simultaneously, providing a more comprehensive exploration of the emotion perception ability of models in conversation.

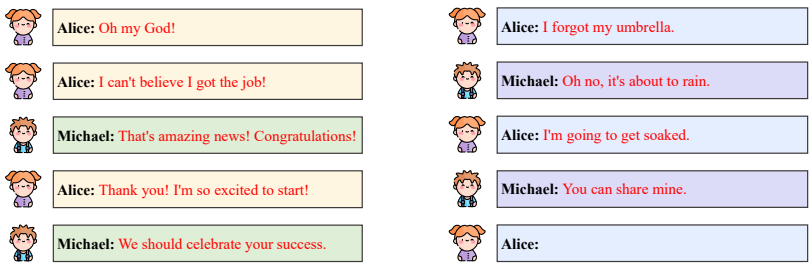
2.3 LLMs and CoT

The introduction of LLMs has revolutionized zero-shot problem-solving, with models like GPT-3 [2], ChatGLM3 [33], and LLaMA [23] achieving notable success in reasoning tasks [29,24]. Techniques such as prompt engineering and CoT reasoning are widely used, particularly in zero-shot settings. Various CoT variants, including TreeCoT [30], AutoCoT [34], Meta-CoT [37], and THOR [6], are proposed to improve the performance of diffrent tasks. However, these methods are not well-suited to complex dialogue structures. In this paper, we focus on how to improve the emotion perception ability of LLMs in zero-shot setting.

3 Pilot Study

3.1 Task Formulation

We evaluate the emotion perception capabilities of LLMs on two tasks: the ERC task, which focuses on perceiving current emotions, and the EIC task, which focuses on perceiving future emotions. Given a multi-turn dialogue $D = [(u_1, p_1), (u_i, p_i), \cdots, (u_n, p_n)]$, 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, we 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. Next, the detailed comparison of the two tasks is provided.



(a) Example of ERC Task. (b) Example of EIC Task.

Fig. 2: Examples of ERC and EIC Task.

ERC tasks primarily aim to identify the emotions of known utterances. As shown in Figure 2a, the goal of the task is to recognize the emotion of Michael's utterance "We should celebrate your success." The model focuses more on the content of the current utterance, perceiving the present emotion, and evaluates the model's ability to understand utterances and emotions.

EIC task mainly involves inferring emotional reactions in the absence of the utterance. As shown in Figure 2b, the goal is to predict the possible emotion that Alice might experience after hearing the utterance "You can share mine." The model focuses more on the impact of the utterance on the individual, perceiving future emotions, and assesses the model's ability to reason about emotions.

From the comparative results, we can observe that, regardless of the model type, the average performance on the ERC task is superior to that on the EIC task. This indicates that the emotional reasoning ability assessed by the EIC task is lacking in the models, making the task more challenging for them.

In summary, the ERC and EIC tasks have differences in task definitions, perceptual directions, ability assessments, and task difficulties. They are two independent tasks. This also demonstrates that considering both tasks simultaneously helps to enhance the model's capabilities more comprehensively.

3.2 Verification Experiment

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, i.e., model score/human score.

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 / 15	23 / 32	130 / 74	34 / 65	6 / 14	3.05 / 3.13
EmoryNLP	8 / 7	11 / 8	16 / 27	36 / 20	1 / 10	3.15 / 3.10
DailyDialog	41 / 44	57 / 121	200 / 197	396 / 219	47 / 160	3.47 / 3.53
IEMOCAP	5 / 8	10 / 8	15 / 12	15 / 14	6 / 8	3.14 / 3.20

To validate the effectiveness of the proposed method, we investigate three issues: whether the persistence and contagiousness of emotions truly affect emotional changes, whether polarity extraction under zero-shot setting is more ac-

curate, and whether personality information can be extracted from dialogues. Based on these considerations, this section designs three validation experiments.

Experiments on Emotional Features We examine three emotion relationships—Pervasive, Personal, and Proximal—to determine whether the emotional characteristics affect the emotion. Pervasive denotes that the emotion of target user aligns with the most frequent emotion in the dialogue history, Personal denotes that the target user's emotion corresponds to their own highest-frequency emotion in the dialogue history, and Proximal denotes that the target user's emotion aligns with the emotions of the nearest users. Personal represents persistence, Proximal represents contagiousness, and Pervasive represents a combination of persistence and contagiousness. We analyze the test dataset and select its relationship according to the priority order of Pervasive, Personal, and Proximal when an utterance belongs to multiple relationships. As presented in Table 1, the results clearly demonstrate that dialogues adhering to the three emotional relationships of emotional features constitute over 70% of each dataset. This highlights the important role of emotional state information in the dialogue, enabling the model to comprehend the dialogue.

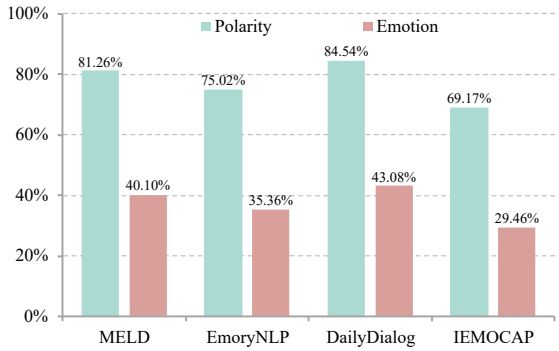


Fig. 3: Comparison of Task Adaptability.

Comparison of Task Adaptability We conduct experiments on both coarse-grained polarity classification tasks and fine-grained emotion classification tasks under zero-shot setting to verify whether polarity classification is more accurate under zero-shot setting. The base model is Mixtral-8x7b [10]. The polarity classification task categorizes utterances into neutral, positive, and negative, while the emotion classification task divides utterances into 7 or 10 categories based on different datasets. As shown in Figure 3, the accuracy of the polarity classification task across all datasets is higher than that of the emotion classification task, demonstrating the high adaptability of LLMs in polarity classification

tasks. Therefore, we incorporate the polarity of utterances as historical emotional state information to assist in emotion perception, reducing perception errors caused by inaccurate historical emotional states.

Evaluation of Personality in Dialogues To ensure that the model can extract meaningful personality information from dialogues, we use GPT-3.5 [2] to assess the degree of personality in the dialogues from the test set. The scoring range is from 1 to 5. Score 1 indicates a complete lack of personality, with mechanical and generic dialogue; score 2 indicates minimal personality, with only slight personalized expressions; score 3 indicates the presence of the user's personality, but still includes some generic expressions; score 4 indicates a strong display of personality, with the dialogue effectively reflecting the user's personality; score 5 indicates that the dialogue is almost entirely composed of personalized expressions. The evaluation prompts are shown in Figure 4. In addition, we also invite three experts specializing in dialogue systems to perform manual scoring in the same manner. The experts not only have strong English reading skills but also possess an in-depth understanding of the field, ensuring an accurate assessment of whether the dialogue contains personality. Each dialogue will be evaluated by three experts, and the final score will be determined by the most frequent score.

[Evaluation rules]

Please score the degree of user personality reflected in the following dialogue on a scale of 1 to 5, where the scores are defined as follows:

Score 1: Completely lacks personality, dialogue is mechanical and generic

Score 2: Minimal personality is reflected, with only slight personalized expressions

Score 3: User's personality is reflected, but still includes some generic expressions

Score 4: Strong personality is reflected, the dialogue well reflects the user's personality

Score 5: The dialogue is almost entirely made up of personalized expressions.

[Dialogue]

Dialogue content

Fig. 4: Prompt of evaluation.

As shown in Table 1, dialogues with scores of 3 and 4 make up the majority in the dataset, with the average scores across datasets exceeding 3.5. The evaluation results indicate that the dialogues in the dataset contain rich personality information, sufficient to support personality extraction.In each dataset, the average scores assigned by the large model and those by human evaluators differ by no more than 0.1, further confirming the reliability of the evaluation results.

4 Methodology

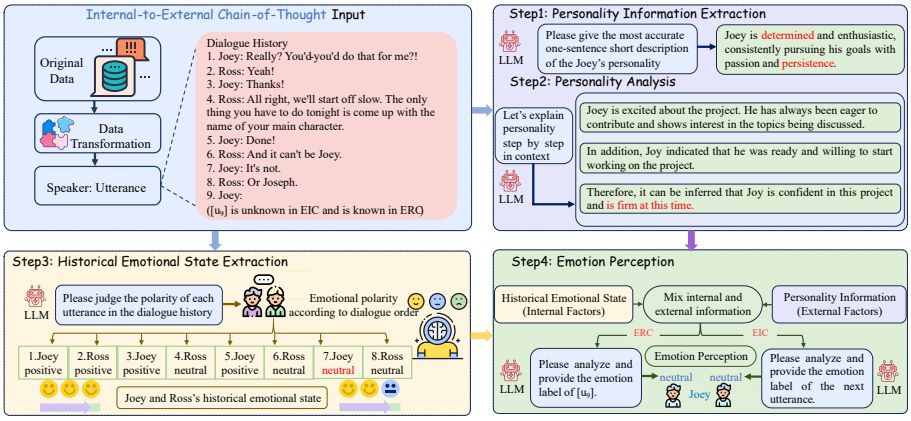


Fig. 5: Framework of IoECoT and an examples of IoECoT.

In this section, we propose the Internal-to-External Chain-of-Thought (IoE-CoT) framework. As shown in Figure 5, the framework sequentially complete four key steps of the chain in a zero-shot setting: personality information extraction, personality analysis, historical emotional state extraction, and emotional perception. First, we illustrate how IoECoT accurately performs emotion perception in real tasks through a multi-step chain of thought utilizing an example.

The objective of this task is to predict Joey's emotion and emotion reaction. Initially, we reorganize the eight sentences from the dialogue into a unified "speaker name: utterance" format. The first step involves extracting Joey's personality information. Based on traits such as "determined" and "persistence," we preliminarily infer that Joey's emotions are not easily altered. The second step is to analyze Joey's personality information by considering the dialogue context. This leads to the conclusion that he is "firm at this time," indicating that his emotional state is stable and not easily influenced. The third step entails extracting the polarity of the utterances to serve as historical emotional states. The dialogue's polarity shifts from positive to neutral, with weak emotional stimuli. Combined with personality information, this suggests that Joey tends to maintain his emotional state. Based on this information, the model conducts analyses for the ERC and EIC tasks separately. For the ERC task, given Joey's ninth utterance, the model determines that the emotion of this utterance is neutral. For the EIC task, in the absence of the ninth utterance, the model predicts that Joey's possible emotional reaction is neutral.

Next, we will provide a detailed explanation of the technical details of each step in the IoECoT, clarifying the implementation process.

4.1 Personality Information Extraction

The "Internal-to-External Chain-of-Thought Input" in Figure 5 indicates that there are usually multiple users in a dialogue. Without distinction, the model will encounter difficulties in positioning and will be unable to accurately capture information related to the target user. Therefore, the dialogue history is standardized in the form of "speaker name: utterance" to ensure that the model can accurately locate utterances associated with a target user. After constructing the data, we proceed to Figure 5 Step1: Personality Information Extraction. Research [8] shows that LLMs are more prone to errors when their outputs are too long. We impose a limit by using "the most accurate one-sentence short description" to generate concise and accurate personality descriptions. Specifically, the standardized dialogue history is input into the LLMs, and a prompt containing the target user's name and task requirements is set, allowing the model to output a natural language representation of the target user's personality based on the dialogue history.

4.2 Personality Analysis

In Figure 5 Step2: Personality Analysis, we further interpret the obtained personality information to fully utilize the personality information. Directly using the personality information cannot comprehensively reflect the target user's emotional sensitivity in the current dialogue context. Drawing from the previous work [11], we adopt the "Let's think step by step" approach to interpret the personality information. We combine the dialogue history with personality information and feed them into the model, enabling LLMs to progressively explain how the target user with this personality is affected by events in the dialogue context. The analyzed personality information already includes the target user's sensitivity to emotional impacts in the current scenario, providing an internal basis for whether emotions continue or transfer to the next step. Through two steps of information extraction and interpretation, the model obtains the target user's personality information as an internal factor for emotions.

4.3 Historical Emotional State Extraction

We utilize Step3 in Figure 5 to obtain the historical emotional state as an external factor. As previously mentioned, LLMs have higher accuracy in classifying the polarity. Therefore, we utilise LLMs to categorize each utterance as neutral, positive, or negative, and record them in the format "speaker name: polarity." Since the intensity of emotional impact is related to the temporal distance between utterances and the closer the utterances are, the stronger the emotional impact. We arrange all polarity results of the historical utterances in the order of the dialogue to ensure the accuracy of emotional impact intensity. This method captures the emotional changes that occur during the progression of the dialogue, providing historical emotional states as external factors for emotion perception.

4.4 Emotion Perception

After the previous steps, the model has acquired both internal and external factors. We proceed to the final step in Figure 5, Step4: Emotion Perception. Guided by internal factors such as personality information, the LLMs analyze the historical emotional states, which serve as an external factor, based on the emotional sensitivity of the dialogue participants and the order of the dialogue. This multi-dimensional analysis ensures both the coherence of emotion perception and captures the emotional shifts of participants during the dialogue. Through this approach, the model produces an accurate result of emotion perception.

5 Experiment

5.1 Datasets

We mainly evaluate our model on four commonly used public dialogue datasets. **MELD** [21] is a multimodal dialogue dataset collected from Friends, containing seven emotions. **EmoryNLP** [32] is collected from Friends, containing seven emotions. **IEMOCAP** [3] is a multimodal dialogue dataset, including nine emotions. **DailyDialog** [15] is a multi-round dialogue dataset collected from the English learning websites, including seven emotions.

The detailed emotion categories for the different datasets are as follows: MELD contains **seven** types of emotions: joyful, sad, neutral, angry, surprise, fear, and disgust. EmoryNLP also includes **seven** emotions: joyful, sad, neutral, mad, powerful, frustrated, and peaceful. DailyDialog has **seven** emotions: anger, happiness, sadness, fear, disgust, surprise, and others. IEMOCAP contains **six** emotions: angry, happy, sad, neutral, frustrated, and excited.

5.2 Baselines and Experimental Setup

We compare with the current traditional zero-shot approaches, CoT methods in the field of NLP and the models trained on the data. **Direct Prompt**: The use of natural language as a prompt for LLMs to accomplish tasks. **CoT** [28]: "Let's think step by step" served as guidance for the LLMs. **Plan-and-Solve** [26]: It instructs LLMs to develop a problem-solving plan. **ECoT** [16]: Improving the performance of LLMs in various emotion tasks by aligning them with human emotional intelligence guidelines. **Cue-CoT** [25]: It enhances LLM reasoning through intermediate reasoning steps to identify clues presented in the dialogue. **SPCL** [22]: A supervised prototypical contrastive learning loss for the ERC task. **InstructERC** [12]: It reformulate the ERC task from a discriminative framework to a generative framework based on LLMs. **DialogInfer-(S+G)+K** [14] It integrates COMET commonsense, sequence, graph modeling for the EIC task. **DialogueGLP** [27] It uses InstructGPT and RNN for the EIC task.

We utilize ChatGLM3-6B [5], Claude-3 [1], and Mixtral 8x7B [10] as base model. 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.

Table 2: The main results of IoECoT performing the ERC task, w-F1 represents the Weighted-F1 score, and m-F1 represents the Macro-F1 score.

Method		MELD	EmoryNLP	DailyDialog	IEMOCAP
		w-F1 / m-F1	w-F1 / m-F1	w-F1 / m-F1	w-F1 / m-F1
ChatGLM3 (zero-shot)	Direct Prompt	38.97 / 26.03	22.04 / 15.70	20.69 / 17.78	15.34 / 13.25
	CoT	28.54 / 16.14	12.76 / 11.04	18.07 / 13.80	07.64 / 06.75
	Plan-and-Solve	30.37 / 16.23	06.57 / 04.66	12.85 / 06.90	10.33 / 09.91
	ECoT	29.60 / 20.06	21.52 / 15.98	14.52 / 13.60	17.47 / 16.81
	Cue-CoT	32.12 / 27.62	16.08 / 12.81	19.22 / 18.75	12.35 / 09.56
	IoECoT	**40.25 / 28.78**	**23.42 / 16.20**	**29.97 / 20.23**	**21.69 / 18.10**
Claude-3 (zero-shot)	Direct Prompt	41.31 / 42.24	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	05.72 / 05.68
	ECoT	44.48 / 39.59	26.33 / 21.81	33.24 / 20.32	10.07 / 09.48
	Cue-CoT	25.22 / 18.81	30.66 / 23.30	27.83 / 21.26	18.33 / 17.61
	IoECoT	**54.61 / 45.35**	**31.03 / 25.00**	**39.82 / 21.58**	**20.65 / 18.73**
Mixtral-8x7b (zero-shot)	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	07.00 / 07.14
	Plan-and-Solve	34.22 / 26.12	28.45 / 25.24	31.64 / 28.08	12.82 / 12.99
	ECoT	38.38 / 24.49	20.24 / 14.82	44.48 / 22.60	15.83 / 15.62
	Cue-CoT	50.29 / 35.90	31.55 / 28.99	30.05 / 21.78	16.18 / 14.64
	IoECoT	**59.01 / 48.05**	**32.69 / 27.79**	**59.76 / 37.55**	**22.48 / 20.92**
Fine-Tuning (full data)	SPCL	63.71 / 46.21	40.70 / 35.29	75.23 / 54.55	67.78 / 66.38
	InstructERC	69.15 / 67.32	41.37 / 38.29	82.21 / 64.33	71.39 / 70.35

5.3 Main Results

Tables 2 and Table 3 present the performance of IoECoT on two emotion perception in conversations tasks. The evaluation results show that IoECoT achieves the State-of-the-Art (SOTA). Compared with traditional zero-shot methods and chain-of-thought approaches in the NLP field, IoECoT demonstrates stronger emotion perception capabilities, not only accurately capturing current emotions but also effectively predicting future emotions. It efficiently extracts information that aids in understanding the dialogue and correctly organizes and applies this information to assist in emotion perception. It is obvious that IoECoT enhances model performance across the three base models, proving its excellent adaptability and portability, and can flexibly adapt to different base models.

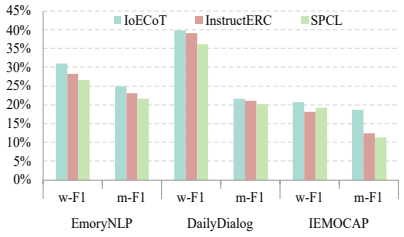
Additionally, IoECoT achieves performance improvements across various datasets. Despite the differences in dialogue scenarios, the number of emotion categories, dialogue turns, and the number of participants in these datasets, IoECoT consistently demonstrates strong adaptability. No matter how complex the dialogue scenarios are, IoECoT effectively performs emotion perception and enhances model performance. This indicates that the model has advantages of maintaining stable performance across a wide range of application scenarios.

Although there is still a performance gap between our model and the trained models, we can efficiently complete emotion perception without any data or training resources, overcoming the limitations of high-resource-demanding models that perform poorly when facing new types of data. As shown in Figure 6, while these models achieve good results on the MELD dataset after training

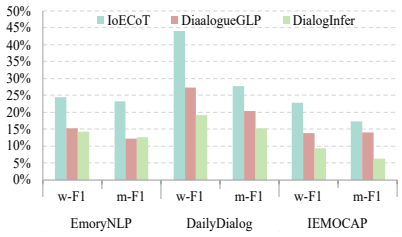
Table 3: The main results of IoECoT performing the EIC task.

Method		MELD	EmoryNLP	DailyDialog	IEMOCAP
		w-F1 / m-F1	w-F1 / m-F1	w-F1 / m-F1	w-F1 / m-F1
ChatGLM3 (zero-shot)	Direct Prompt	29.75 / 12.64	17.88 / 14.51	18.00 / 11.52	06.00 / 06.47
	CoT	30.01 / 11.22	11.29 / 06.96	14.08 / 14.10	07.35 / 09.04
	Plan-and-Solve	31.43 / 13.67	12.30 / 07.60	17.92 / 16.33	09.03 / 10.07
	ECoT	29.96 / 16.26	17.88 / 14.51	12.43 / 11.24	05.11 / 05.68
	Cue-CoT	25.22 / 18.81	18.13 / 15.76	11.00 / 05.71	05.49 / 05.00
	IoECoT	**35.11 / 16.70**	**17.29 / 11.75**	**55.51 / 21.64**	**12.70 / 13.69**
Claude-3 (zero-shot)	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 / 20.27	19.00 / 15.71
	ECoT	14.65 / 10.26	18.78 / 17.50	32.13 / 14.44	16.80 / 17.96
	Cue-CoT	14.52 / 14.09	19.42 / 17.24	13.62 / 17.89	11.17 / 11.85
	IoECoT	**20.35 / 22.26**	**20.31 / 16.09**	**32.58 / 20.36**	**27.84 / 24.72**
Mixtral-8x7b (zero-shot)	Direct Prompt	32.92 / 18.76	18.58 / 15.00	29.86 / 20.22	11.57 / 08.91
	CoT	26.69 / 18.70	18.67 / 13.57	18.21 / 19.30	09.58 / 07.65
	Plan-and-Solve	28.45 / 18.85	17.08 / 13.49	15.93 / 18.22	13.37 / 12.17
	ECoT	15.08 / 10.82	20.31 / 19.17	22.15 / 13.28	07.11 / 06.90
	Cue-CoT	23.42 / 17.79	10.87 / 09.01	25.75 / 19.45	11.22 / 12.45
	IoECoT	**33.33 / 23.62**	**24.47 / 23.13**	**44.06 / 27.80**	**22.82 / 17.28**
Fine-Tuning (full data)	DialogueGLP	38.42 / 20.88	22.08 / 19.13	75.11 / 40.93	26.35 / 23.66
	DialogInfer	39.22 / 21.05	22.35 / 18.27	54.24 / 32.66	23.24 / 20.15

on MELD data, they fail to perform well on the other three datasets. In contrast, our model performs well on the other three datasets without any training, demonstrating its excellent generalization ability. This generalization ability reflects our method's adaptability to different data distributions. Compared to traditional models, our approach meets the diverse needs of different domains and datasets, making the model more robust.



(a) Generalization experiment on the ERC task.



(b) Generalization experiment on the EIC task.

Fig. 6: Generalization experiment. The base model is Mixtral-8x7b.

In the experimental results of Table 2 and Table 3, the performance gap between IoECoT and the trained model on the IEMOCAP dataset is larger compared to other datasets. This is because the long-context nature of the IEMO-

CAP dataset further highlights the inadequacies of LLMs in handling long dialogues. The average number of dialogue turns in the IEMOCAP dataset is higher than in other datasets, reaching 49.2 turns, while the dialogue turns in other datasets range only between 7.9 and 11.5. This indicates that the IEMOCAP dataset contains longer contexts, reflecting more dialogue content and higher complexity. Nevertheless, IoECoT still maximizes the model's potential in long-context scenarios, enhancing its ability of emotion perception.

5.4 Ablation Study

As shown in Table 4, we conduct ablation experiments on four datasets using the Mixtral-8x7b. The results indicate that removing either historical emotional state information or personality information weakens the model's performance. This is because when only historical emotional state is retained, the internal factors contributing to emotion generation are overlooked, and the model's emotion perception becomes entirely dependent on historical emotional states. This leads to a failure in accurately assessing the participants' sensitivity to external stimuli and the magnitude of potential emotional changes, resulting in biased emotion perception. On the other hand, when only personality information is retained, the external factors influencing emotion generation are omitted, weakening the model's understanding of external influences. Emotions become solely tied to personality, independent of historical emotional states, which clearly contradicts human emotional cognition.

Table 4: Ablation study. We utilize the Mixtral-8x7b as the base model.

	Method	MELD	EmoryNLP	DailyDialog	IEMOCAP
		weighted-F1	weighted-F1	weighted-F1	weighted-F1
ERC	IoECoT	**59.01**	**32.69**	**59.76**	**22.48**
	w/o personality	51.25	29.84	49.61	20.12
	w/o emotional state	52.65	29.24	50.47	19.33
EIC	IoECoT	**33.33**	**24.47**	**44.06**	**22.82**
	w/o personality	30.18	18.96	41.03	19.24
	w/o emotional state	29.06	14.23	42.66	13.45

Furthermore, the results of the ablation experiments further validate the positive impact of historical emotional state information and personality information obtained through LLM on the model's performance. Including either of these two types of information leads to an improvement in the model's performance. This finding indicates that the information generated by guiding the LLM has high quality, effectively enhancing the model's emotional perception capabilities, thus ensuring the accuracy and reliability of subsequent reasoning processes.

6 Case Study

As shown in Figure 7, Ross feels confused and anxious after unexpectedly using a British accent, and he urgently seeks help. According to the personality information extracted from the dialogue history, Ross exhibits strong self-consciousness and insecurity, particularly concerning others' opinions, and he tends to seek external support under emotional stress. Further personality analysis reveals that Ross's emotions are easily influenced by others' perceptions, leading to fluctuations in his emotional state. Consequently, his emotional state is prone to change during dialogue. By extracting the historical emotional state, it is evident that while Ross's polarity remains negative, his friends' polarity is neutral, and they continuously provide guidance. In the ERC task, combining Ross's line "Monica, you really think I should try this phasing out thing?" it can be concluded that his emotions have begun to stabilize, with the emotion identified as neutral. In the EIC task, Ross is encouraged by his friends, and based on his personality, it is evident that his emotions are influenced by them, resulting in a neutral emotional response.

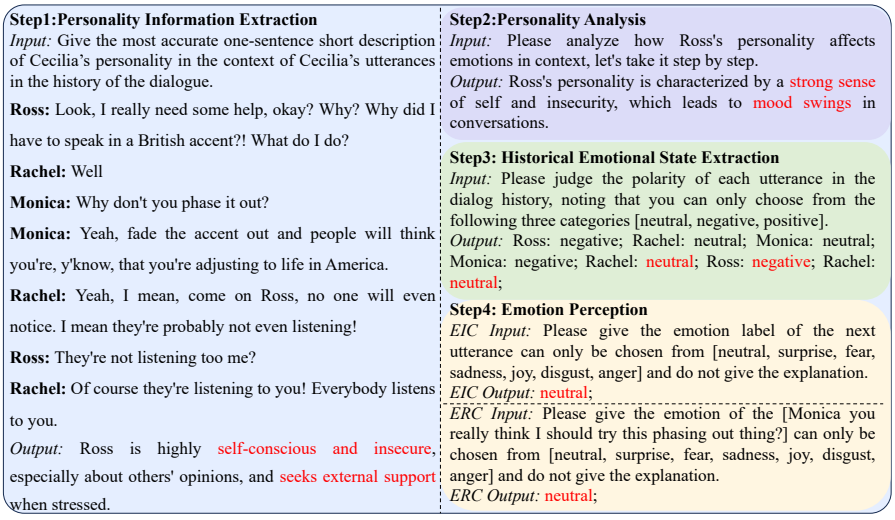


Fig. 7: Case Study. The case of the MELD dataset with Mixtral-8x7b as the base model.

7 Conclusion

In this paper, we introduce a novel chain-of-thought framework called IoECoT. Our framework aims to integrate and leverage historical emotional states with

personality information, using an internal-to-external approach for information integration. Experimental results demonstrate that our proposed framework enhances the ability of emotion perception in conversation, particularly in zero-shot setting. Furthermore, we simultaneously consider both the ERC and EIC tasks, conducting a more comprehensive study on the emotion perception ability of models in conversation. 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 dialogue understanding.

Acknowledgments. This work is supported by the National Natural Science Foundation of China (62272092, 62172086) and the Fundamental Research Funds for the Central Universities of China (No. N2116008).

References

1. Anthropic: Introducing the next generation of Claude \ Anthropic (2024)
2. Brown, et al.: Language models are few-shot learners. Advances in neural information processing systems **33**, 1877–1901 (2020)
3. Busso, C., et al.: Iemocap: Interactive emotional dyadic motion capture database. Language resources and evaluation **42**, 335–359 (2008)
4. Dimitroff, et al.: Physiological dynamics of stress contagion. Scientific reports **7**(1), 6168 (2017)
5. Du, Z., et al.: GLM: General language model pretraining with autoregressive blank infilling. In: ACL. pp. 320–335. Dublin, Ireland (May 2022)
6. Fei, H., Li, B., Liu, Q., Bing, L., Li, F., Chua, T.S.: Reasoning implicit sentiment with chain-of-thought prompting. In: ACL (2023)
7. Ghosal, D., et al.: Dialoguegcn: A graph convolutional neural network for emotion recognition in conversation. In: EMNLP (2019)
8. Huang, L., et al.: A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. ACM Transactions on Information Systems (2024)
9. Imbir, K.K.: Origins and source of emotion as factors that modulate the scope of attention. Roczniki Psychologiczne **16**(2), 287–310 (2013)
10. Jiang, Albert Q, e.a.: Mixtral of experts. arXiv preprint arXiv:2401.04088 (2024)
11. Kojima, et al.: Large language models are zero-shot reasoners. Advances in neural information processing systems **35**, 22199–22213 (2022)
12. Lei, S., et al.: Instructerc: Reforming emotion recognition in conversation with a retrieval multi-task llms framework. arXiv preprint arXiv:2309.11911 (2023)
13. Li, D., et al.: Emotion inference in multi-turn conversations with addressee-aware module and ensemble strategy. In: EMNLP. pp. 3935–3941. Association for Computational Linguistics, Online and Punta Cana, Dominican Republic (Nov 2021)
14. Li, D., et al.: Enhancing emotion inference in conversations with commonsense knowledge. Knowledge-Based Systems **232**, 107449 (2021)
15. Li, Y., et al.: DailyDialog: A manually labelled multi-turn dialogue dataset. In: IJCNLP (Volume 1: Long Papers). pp. 986–995. Asian Federation of Natural Language Processing, Taipei, Taiwan (Nov 2017)

16. Li, Z., et al.: Enhancing the emotional generation capability of large language models via emotional chain-of-thought. arXiv preprint arXiv:2401.06836 (2024)
17. Liu, S., et al.: Towards emotional support dialog systems. In: ACL. pp. 3469–3483. Online (Aug 2021)
18. Liu, Y., Feng, S., Wang, D., Zhang, Y., Schütze, H.: Chatzero: Zero-shot cross-lingual dialogue generation via pseudo-target language. In: Endriss, U., Melo, F.S., Bach, K., Diz, A.J.B., Alonso-Moral, J.M., Barro, S., Heintz, F. (eds.) ECAI 2024 (PAIS 2024). Frontiers in Artificial Intelligence and Applications, vol. 392, pp. 3867–3874. IOS Press (2024)
19. Ma, Y., et al.: A survey on empathetic dialogue systems. Information Fusion **64**, 50–70 (2020)
20. Mitchell, J.: Affective persistence and the normative phenomenology of emotion. In: Tappolet, C., Deonna, J., Teroni, F. (eds.) A Tribute to Ronald de Sousa (2022)
21. Poria, S., et al.: Meld: A multimodal multi-party dataset for emotion recognition in conversations. In: ACL. pp. 527–536 (2019)
22. Song, X., et al.: Supervised prototypical contrastive learning for emotion recognition in conversation. In: EMNLP. pp. 5197–5206 (2022)
23. Touvron, et al.: Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971 (2023)
24. Wang, et al.: Self-consistency improves chain of thought reasoning in language models. arXiv preprint arXiv:2203.11171 (2022)
25. Wang, H., et al.: Cue-CoT: Chain-of-thought prompting for responding to in-depth dialogue questions with LLMs. In: Findings of EMNLP 2023. pp. 12047–12064. Association for Computational Linguistics, Singapore (Dec 2023)
26. Wang, L., et al.: Plan-and-solve prompting: Improving zero-shot chain-of-thought reasoning by large language models. In: ACL. pp. 2609–2634 (2023)
27. Wang, R., Feng, S.: Global-local modeling with prompt-based knowledge enhancement for emotion inference in conversation. In: EACL 2023 Finding. pp. 2120–2127. Dubrovnik, Croatia (May 2023)
28. Wei, J., et al.: Chain-of-thought prompting elicits reasoning in large language models. Advances in Neural Information Processing Systems **35**, 24824–24837 (2022)
29. Xi, Z., et al.: Self-polish: Enhance reasoning in large language models via problem refinement. In: Findings of EMNLP 2023. pp. 11383–11406 (2023)
30. Yao, S., et al.: Tree of thoughts: Deliberate problem solving with large language models. Advances in Neural Information Processing Systems **36** (2024)
31. Young, et al.: Emotion regulation choice: A broad examination of external factors. Cognition and Emotion (2019)
32. Zahiri, et al.: Emotion detection on tv show transcripts with sequence-based convolutional neural networks. In: Workshops at the thirty-second aaai conference on artificial intelligence (2018)
33. Zeng, et al.: Glm-130b: An open bilingual pre-trained model. arXiv preprint arXiv:2210.02414 (2022)
34. Zhang, Z., Zhang, A., Li, M., Smola, A.: Automatic chain of thought prompting in large language models. arXiv preprint arXiv:2210.03493 (2022)
35. Zhao, W.X., et al.: A survey of large language models. arXiv preprint arXiv:2303.18223 (2023)
36. Zhong, P., et al.: Knowledge-enriched transformer for emotion detection in textual conversations. In: EMNLP-IJCNLP. pp. 165–176. Association for Computational Linguistics, Hong Kong, China (Nov 2019)
37. Zou, A., et al.: Meta-cot: Generalizable chain-of-thought prompting in mixed-task scenarios with large language models. arXiv preprint arXiv:2310.06692 (2023)