CoE: A Clue of Emotion Framework for Emotion Recognition in Conversations

Zhiyu Shen[1], **Yunhe Pang**[1], **Yanghui Rao**[1]*, **Jianxing Yu**[2]
[1]School of Computer Science and Engineering, Guangdong Key Laboratory of Big Data Analysis and Processing, Sun Yat-sen University, Guangzhou, China
[2]School of Artificial Intelligence, Sun Yat-sen University, Zhuhai, China

{shenzhy23, pangyh8}@mail2.sysu.edu.cn, {raoyangh, yujx26}@mail.sysu.edu.cn

Abstract

Emotion Recognition in Conversations (ERC) is crucial for machines to understand dynamic human emotions. While Large Language Models (LLMs) show promise, their performance is often limited by challenges in interpreting complex conversational streams. We introduce a Clue of Emotion (CoE) framework, which progressively integrates key conversational clues to enhance the ERC task. Building on CoE, we implement a multi-stage auxiliary learning strategy that incorporates role-playing, speaker identification, and emotion reasoning tasks, each targeting different aspects of conversational emotion understanding and enhancing the model's ability to interpret emotional contexts. Our experiments on EmoryNLP, MELD, and IEMOCAP demonstrate that CoE consistently outperforms state-of-the-art methods, achieving a 2.92% improvement on EmoryNLP. These results underscore the effectiveness of clues and multi-stage auxiliary learning for ERC, offering valuable insights for future research.

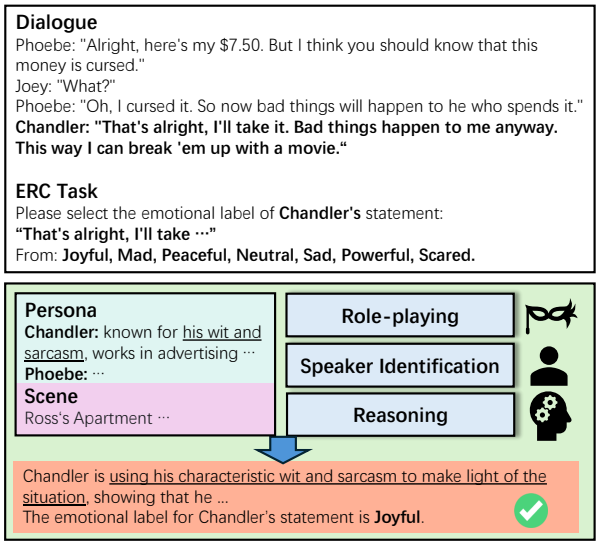

Figure 1: This figure illustrates the CoE framework in the ERC task, where contextual information (e.g., Persona and Scene) on the left and auxiliary tasks (e.g., role-playing, speaker identification, and reasoning) on the right are integrated to infer Chandler's emotional label. These combined components enhance the model's ability to reason through emotional clues and determine that Chandler's emotional state is Joyful.

1 Introduction

Emotion Recognition in Conversations (ERC) (Poria et al., 2019; Zahiri and Choi, 2018) has emerged as a significant research direction, driven by its potential applications to improve human-computer interactions (Brave and Nass, 2007), virtual assistants (Chatterjee et al., 2021), and mental health monitoring (Monteith et al., 2022). Accurately identifying and interpreting emotions within conversational contexts is challenging due to the complexity and subtlety of emotional expressions.

Previous works have used recurrent-based models (Hu et al., 2023) and transformer-based discriminative models (Qin et al., 2023) to learn contextual representations, with some methods further employing graph neural networks to model interactions and influences between individuals in conversations (Ghosal et al., 2019). Recently, generative language models (Brown, 2020) have demonstrated powerful zero-shot capabilities, introducing a new paradigm to various natural language processing tasks. In ERC, some approaches have transformed classification tasks into generation tasks by generating class tokens. InstructERC (Lei et al., 2023) is the first to construct an ERC method based on LLMs, utilizing similarity-based retrieval to build context-learning instructions and incorporating multi-task learning to fine-tune the model.

However, previous LLM-based methods (Zhang et al., 2023b; Fu, 2024; Lei et al., 2023) have overlooked the crucial role of dialogue-specific context

*The corresponding author.

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 23548–23563

and history, limiting their ability to extract deeper insights from conversations and leading to an incomplete understanding of the emotions conveyed. This highlights the need for a model that can effectively capture and analyze the nuances of dialogue context and character information to provide a more comprehensive understanding of emotions.

We believe that previous research on the ERC task still requires further exploration of global dialogue scene information, a crucial component that affects emotion analysis in conversations (Poria et al., 2017). Despite the remarkable performance of current LLMs like o1 (OpenAI et al., 2025; Patil, 2025) and DeepSeek-R1 (DeepSeek-AI et al., 2025) in logical reasoning, mathematics, and programming tasks, their potential for deep information extraction in computational emotion analysis remains largely unexplored.

In light of these considerations, we propose the Clue of Emotion (CoE) framework to deeply explore the diverse textual clues of characters and scenes within dialogues. These clues enable the model to gain a profound understanding of the emotions expressed in the dialogue. As shown in Figure 1, we incorporate persona and scene as global textual clues to enrich the context of the dialogue and strengthen the model's understanding of conversational dynamics.

Additionally, we introduce auxiliary tasks, including role-playing, speaker identification, and rationale reasoning, coupled with a multi-stage training strategy to integrate these components effectively. The role-playing and speaker identification tasks serve as complementary role-enhancement tasks, enabling the LLM to understand character personalities and scene characteristics in depth. The rationale reasoning task employs a self-taught emotion reasoning approach, allowing the model to interpret characters' emotions in specific scenarios. Through this progressive training framework, we enhance the model's ability to accurately infer emotions in complex conversational contexts while effectively utilizing available dialogue clues.

Our experimental findings highlight important implications for the broader field of computational emotion analysis. The consistent performance improvements across multiple benchmarks suggest that exploring textual clues in conversations opens new avenues for enhancing LLMs' emotional intelligence. This approach points toward the development of more sophisticated conversational frameworks that better capture the nuanced relationship between dialogue context and emotional expression. To the best of our knowledge, current models still underperform in emotion recognition tasks despite their well-established efficacy in other complex reasoning tasks, highlighting a notable research gap in conversational emotion modeling.

The main contributions of this study are summarized as follows:

- We propose the CoE framework, which simultaneously introduces persona and scene-based textual augmentations for dialogue scenarios and enhances the capabilities of LLMs through textual and parametric optimizations.
- We systematically design three auxiliary tasks within the CoE framework to train the LLM, enabling it to deeply understand and extract character traits and the underlying reasons for emotions in dialogue scenarios.
- We conduct comprehensive experiments on the entire method using three ERC benchmark datasets, and the outstanding results demonstrate its effectiveness. Additionally, we carry out in-depth experiments on multiple auxiliary tasks and ERC training strategies to explore the optimal training path for ERC.

2 Related Work

2.1 Emotion Recognition in Conversation

ERC has evolved from feature-based methods to deep learning approaches and recent LLM innovations. Traditional approaches include context-dependent emotion analysis with hand-crafted features (Poria et al., 2017), DialogueRNN's speaker state tracking (Majumder et al., 2019), and Graph Neural Network (GNN)-based like DialogueGCN (Ghosal et al., 2019) that models speaker interactions. While effective, these methods struggled with complex conversational dynamics.

Recent advances position LLMs as powerful tools for ERC. InstructERC reformulates ERC as a retrieval-based instruction-following task, achieving state-of-the-art zero-shot results on multiple datasets that rival previous baselines (Lei et al., 2023). General purpose LLMs such as ChatGPT show emergent zero-shot emotion recognition capabilities (Yang et al., 2023). Subsequent studies further highlight LLMs' promise for emotion cognition while noting their dependence on rich conversational context (Chen and Xiao, 2024; Sabour

et al., 2024). Nevertheless, the full potential of these high-capacity generative models for ERC, especially when cast as a classification problem, remains underexplored, motivating the development of our CoE framework.

2.2 Modeling Conversational Context

Prior studies have modeled speaker characteristics to capture conversational dynamics. For instance, DialogueGCN (Ghosal et al., 2019) uses speaker embeddings to represent inter-speaker interactions, while HeterMPC (Gu et al., 2022) employs heterogeneous GNNs for speaker representations. Recent work further infuses persona information into cross-task GNNs to enhance multimodal ERC (Tu et al., 2024). These methods resonate with our persona clues, focusing on character traits and roles, though not explicitly categorized as such in earlier studies.

Works like the EmoSeC framework (Thuseethan et al., 2022) leverage scene context for emotion recognition, while datasets such as KD-EmoR (Pant et al., 2022) incorporate scene descriptions with Korean drama transcripts to enrich contextual representations, aiding emotion disambiguation. These approaches align with our scene clues concept, though prior work typically integrates such information within broader contextual frameworks without systematically separating these elements as our framework does.

2.3 Auxiliary Learning for ERC

Auxiliary tasks provide richer supervision signals for ERC. Multi-stage frameworks for emotion-cause pair extraction enhance model understanding of emotional triggers (Ding et al., 2020; Li et al., 2023). Role-playing and speaker identification, which we incorporate into CoE, improve speaker-aware emotion modeling (Li et al., 2020).

Other complementary tasks proposed in previous research like emotion shift analysis (Poria et al., 2019; Gao et al., 2022) and sarcasm recognition (Castro et al., 2019) improve robustness by handling complex emotional transitions and nuanced expressions that often confuse standard ERC models. Semi-supervised methods with context-augmented auxiliary tasks (Kang et al., 2021) further enhance performance in low-resource settings. Our multi-stage auxiliary learning strategy builds on these approaches, sequentially incorporating role-playing, speaker identification, and emotion reasoning to provide comprehensive supervision for our CoE framework.

3 Preliminaries

In the ERC task, our goal is to identify the speaker's emotion label at each turn of the conversation. Given a set of speakers S and an emotion label set $\mathcal{E}$, each N-turn conversation C is represented as $\{(s_1, u_1), (s_2, u_2), \ldots, (s_N, u_N)\}$, where $s_i \in S$ is the speaker and u_i is the utterance of the i-th turn. The model can only utilize previous dialogue utterances $\{(s_1, u_1), (s_2, u_2), \ldots, (s_{t-1}, u_{t-1})\}$ as input to predict the emotion label y_t for the t-th dialogue utterance. The emotion labels are from a predefined set $\mathcal{E} = \{e_1, e_2, \ldots, e_o\}$, where o is the number of emotion categories.

4 Clue of Emotion

In this section, we first introduce the overall CoE framework (Section 4.1). Next, we present the auxiliary tasks (Section 4.2), encompassing three distinct domains of emotion-related capabilities. Finally, we discuss the design of the training strategy (Section 4.3).

4.1 CoE Framework

To effectively utilize critical conversational clues in the ERC task, we propose a CoE framework that synthesizes essential dialogue components. Our approach not only analyzes the utterance itself but also integrates the dialogue scenes, speaker personas, and dialogue history, providing a more comprehensive and structured understanding of the conversation. By combining these elements, the CoE framework can be applied to various tasks related to the ERC task. It can also be equipped with auxiliary tasks to enhance the model's acquisition of emotion-related domain-specific skills.

To demonstrate the advantages of the CoE framework, we consider a conversational scenario where multiple participants with distinct personas engage in a discussion. In such interactions, capturing emotional dynamics requires understanding not only the utterances but also the underlying speaker personas, the evolving dialogue context, and the scene. Each speaker may express emotions differently based on their persona, their role in the conversation, and the environment where they are situated. The CoE framework effectively captures subtle emotional shifts by incorporating these key conversational clues. Detailed descriptions of clues can be found in Section 5.1.1 and Appendix A.

To formalize the proposed CoE framework, we define the following components: the speaker set

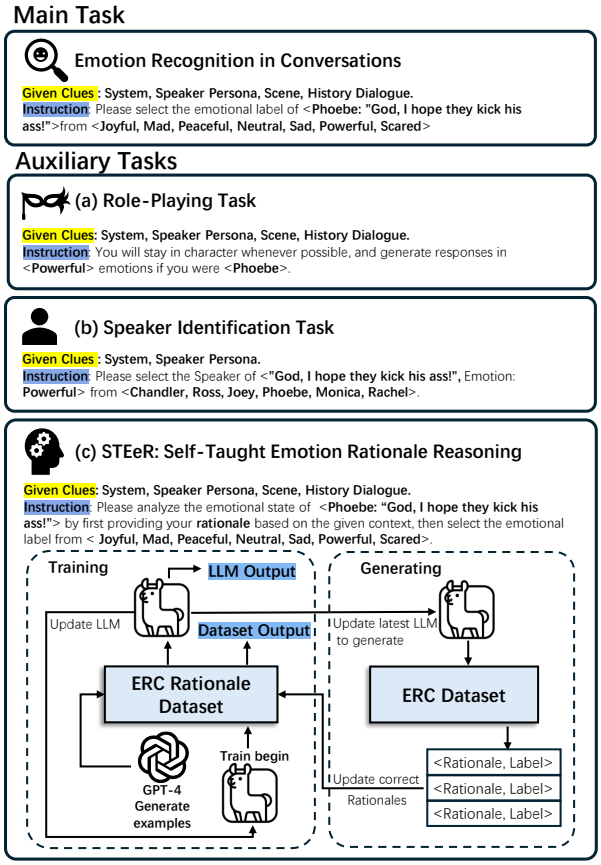

Figure 2: The CoE Framework

$S = \{s_1, s_2, \ldots, s_M\}$, where M represents the number of speakers; the dialogue scene D; the speaker persona set $P = \{p_1, p_2, \ldots, p_M\}$, where each p_i corresponds to a speaker in S; and the historical context set $H = \{h_1, h_2, \ldots, h_{t-1}\}$, capturing previous dialogue history.

The emotion prediction task using CoE can be expressed as a function f_{emotion}, which combines all these elements to predict the emotion label:

$$e_i = \arg\max_{e \in \mathcal{E}} f_{\text{emotion}}(u_i, S, D, P, H, \mathcal{E}) \quad (1)$$

where e_i is the predicted emotion label for the utterance u_i. Other illustrations of the CoE can be seen in Appendix A.

4.2 Auxiliary Learning Tasks

Previous work (Chen et al., 2024; Kung et al., 2021) has demonstrated that identifying and leveraging interdependent skills can enhance model performance by making the learning process more efficient. Similarly, we integrate auxiliary tasks related to the ERC task into the CoE framework to improve the ERC task performance by providing additional insights into speaker behavior and conversational dynamics. We focus on three tasks: the role-playing task, the speaker identification task, and the self-taught emotion rationale reasoning (STEeR) task. These tasks capture key conversational elements, enabling the model to understand the emotion shift in dialogue better. By leveraging these auxiliary tasks, the model can enrich its understanding and boost performance in the ERC task. Detailed examples of the prompt templates used for each task are provided in Appendix E.

4.2.1 Role-Playing Task

Role-specific dialogue generation has been explored in personalizing dialogue agents (Zhang et al., 2018; Madotto et al., 2019), where conditioning on speaker profiles improves response prediction and dialogue coherence by adapting to speaker identity and style. More recently, LLMs have been trained to simulate specific personas by integrating profiles and emotional states (Shao et al., 2023; Wang et al., 2024), further enhancing the quality of generated dialogues. Similarly, role prediction is critical for ERC, as roles shape emotional expression, language style, and intent. The role-playing task enhances the model's ability to generate plausible dialogue by understanding how context, speaker identity, and emotional state influence communication. This ability to predict roles and infer intent is crucial for improving ERC performance within the CoE framework.

In this study, we simulate a role-playing task where the model generates the current utterance u_t^* based on historical context $H = \{h_1, h_2, \ldots, h_{t-1}\}$ and contextual clues (particularly the persona of speaker s). This task can be formulated as:

$$u_t^* = f_{\text{role}}(s, e, D, P, H) \quad (2)$$

where s represents the specified speaker, e denotes the speaker's emotion when speaking. The function $f_{\text{role}}(s, e, D, P, H)$ generates the most logically coherent utterance u_t^* based on the context H.

To process dialogue segments from various scenarios, we apply a series of filtering and transformation steps. The detailed process is provided in Appendix B.

4.2.2 Speaker Identification Task

Different speakers express emotions in distinct ways. Learning each speaker's characteristics helps the model understand the link between personal

traits and emotional expression. Previous models using hierarchical attention networks and GRUs (Liu et al., 2024) failed to connect emotional nuances directly to speaker personas, while InstructERC (Lei et al., 2023) struggled with missing persona information.

Our CoE incorporates speaker personas as key conversational elements and works complementarily with the role-playing task. Through joint learning, the model develops a deeper understanding of how character roles and personalities influence emotional expression, resulting in more accurate predictions of both identity and intent.

The speaker identification task aims to predict the specific speaker based on the utterance and expressed emotion. The task can be formulated as follows:

$$s^* = \arg\max_{s \in S} f_{\text{speaker}}(u, e, S, P) \tag{3}$$

where u represents the current utterance, e denotes the emotion of the speaker when speaking, and $f_{\text{speaker}}(u, e, S, P)$ is an evaluation function.

4.2.3 Self-Taught Emotion Rationale Reasoning Task

When dealing with limited data samples, the model's self-learning mechanism can enhance performance. Inspired by the Self-Taught Reasoning (STaR) method (Zelikman et al., 2022), we adapted this approach for emotion rationale reasoning, resulting in the development of the **S**elf-**T**aught **E**motion Rationa**l**e **R**easoning (STEeR) task. This method iteratively generates rationale data by itself to improve the model's ability to reason through emotional contexts.

Specifically, STEeR begins by generating a small number of initial rationales using GPT-4o [1] due to the large training dataset. However, GPT-4o's relatively low accuracy results in a higher proportion of incorrect answers. To address this, we use these initial rationales to generate further rationales in subsequent iterations, with feedback guiding the process. Over time, as the model generates more correct rationales, its reasoning accuracy improves progressively through each iteration. This iterative process strengthens the model's emotion reasoning capabilities by continuously refining its ability to generate accurate rationales. This is visually represented in Figure 2, where the rationale generation and feedback loop are depicted.

[1]We use the gpt-4o-2024-08-06.

Algorithm 1: The Main Training Process of STEeR

Input: Pretrained LLM M, Dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{D}$, Training Dataset $\mathcal{D}_{\text{train}}$

$M_0 \leftarrow M$, $D_0 \leftarrow D \setminus \mathcal{D}_{\text{train}}$ // Copy the original model and reasoning dataset (pre-generated)

for $n \in 1 \dots N$ **do**

  $M_n \leftarrow \text{train}(M_{n-1}, \mathcal{D}_{\text{train}})$ // Finetune the model on the updated memory dataset

  for $(x_i, y_i) \in \mathcal{D}_{n-1}$ **do**

    $(\hat{r}_i, \hat{y}_i) \leftarrow M_n(x_i)$ // Perform rationale generation

    $\mathcal{D}_{\text{train}} \leftarrow \mathcal{D}_{\text{train}} \cup \{(x_i, \hat{r}_i, y_i) \mid \hat{y}_i = y_i\}$ // Add correct rationales

  end

  $\mathcal{D}_n \leftarrow \mathcal{D}_{n-1} \setminus \mathcal{D}_{\text{train}}$ // Update dataset by removing generated samples

end

The STEeR method operates in two stages: the initial data generation phase and the main STEeR training phase (see Algorithm 1). In the training phase, the model is fine-tuned iteratively using the generated rationales, with each cycle refining its emotional reasoning capacity. This integration of rationale generation and model fine-tuning enables continuous improvement in emotional reasoning tasks, helping the model better align with human standards for the ERC task. The data generation process is detailed in Appendix C.

4.3 Designing Multi-Stage Training Strategies

To effectively leverage each auxiliary task for ERC improvement, we designed a multi-stage training strategy. Each task contributes uniquely: **Role-playing** enhances the model's ability to predict character roles and understand varied language styles and speaker intents. **Speaker identification** builds upon this by enabling the model to differentiate speakers and capture their unique emotional expressions. Finally, **STEeR** further strengthens the model's emotional intelligence, fostering its capacity for emotion reasoning and generating justifiable rationales. These synergistic tasks collectively form a robust foundation for our CoE framework.

Recognizing the challenge of determining the optimal training approach, we investigated both curriculum-based and mixed training strategies. Our experimental findings (Section 5.4) demonstrate that a multi-stage, mixed auxiliary training strategy provides the most significant performance improvements for the ERC task.

5 Experiments

5.1 Experimental Setup

5.1.1 Datasets

We conduct experiments on three ERC datasets: MELD (Poria et al., 2019), EmoryNLP (Zahiri and Choi, 2018), and IEMOCAP (Busso et al., 2008).

MELD (Poria et al., 2019): This dataset is derived from the Friends TV series, including over 1400 dialogues and 13000 utterances. Each dialogue involves multiple speakers and is annotated with seven emotions. We manually set persona information based on the six main characters from the Friends TV series to improve dialogue understanding. We also extract the scene context for each dialogue opening in the dataset.

EmoryNLP (Zahiri and Choi, 2018): This dataset contains 97 episodes and 12,606 utterances, with each utterance being annotated with seven emotions adapted from the Willcox's feeling wheel (Willcox, 1982). We extract scene information to provide clues about the setting of each dialogue. Similar to MELD, we also manually set persona information based on the six main characters.

IEMOCAP (Busso et al., 2008): This dataset contains video recordings of two-person interactions, capturing natural emotional exchanges between pairs of actors. It includes 151 sessions, with each session involving two speakers, resulting in 302 dialogues. Since the dataset consists of segmented video recordings, we rely on historical dialogue information to assist with the ERC task.

5.1.2 Metrics

To maintain consistency with previous methods, we used weighted F1 (W-F1) as the evaluation metric, as it better reflects the model's performance given the severe class imbalance in the tested datasets (Lei et al., 2023; Song et al., 2022). Furthermore, we ensured consistency in training/validation/test splits with both InstructERC (Lei et al., 2023) and COSMIC (Ghosal et al., 2020).

5.1.3 Training Setup

For our experiments, we kept the LLM parameter count around 7B. Considering computational constraints and time limitations, we employed the LoRA method for model fine-tuning. The rank of the LoRA adapter was set to 16, the learning rate was configured to 2e-4, and the LoRA scaling factor was set to 0.1. Our experiments were conducted on an 8×32 GB Nvidia V100 GPU cluster, utilizing FP16 precision for training. The final results were averaged across three independent runs.

5.1.4 Baselines

We select several ERC baselines to compare with our CoE. SPCL-CL (Song et al., 2022), COSMIC (Ghosal et al., 2020), and EACL (Yu et al., 2024) use transformer-based models, where SPCL-CL introduces a contrastive learning approach and EACL employs an emotion-anchored learning framework. UniMSE (Hu et al., 2022) and TelME (Yun et al., 2024) leverage multimodal data through teacher-student architectures. SACL (Hu et al., 2023) and EmotionIC (Liu et al., 2024) use recurrent-based methods. DualGATs (Zhang et al., 2023a) employs GNN-based methods. InstructERC (Lei et al., 2023) and BiosERC (Xue et al., 2024) use LLM-based methods. We also test the performance of GPT-4o (Hurst et al., 2024), and DeepSeek-R1 (DeepSeek-AI et al., 2025), a leading reasoning model, as strong baselines.

5.2 Main Results

Previous methods, such as COSMIC (Ghosal et al., 2020), leveraged commonsense knowledge through knowledge graphs to model mental states and contextual information, enhancing their ability to infer emotions. InstructERC (Lei et al., 2023) was the first to leverage generative large language models for the ERC task, leading to improved performance across all datasets and demonstrating the potential of generation-based modeling for emotion recognition in conversations. While these models incorporated external information sources, our approach directly and systematically models conversational elements such as speaker personas and scenes, embedding them explicitly within the task itself. Additionally, we incorporate auxiliary tasks to further enhance the model's ability to process these clues by utilizing supervised learning to refine its understanding of emotional dynamics.

Our results (Table 1) show that CoE outperformed previous methods across multiple datasets,

Table 1: Main Results for the ERC task

Methods	Backbone	EmoryNLP W-F1	MELD W-F1	IEMOCAP W-F1	Average W-F1
EmotionIC	RB	40.25	66.32	69.61	58.73
SACL	RB	39.65	66.45	69.22	58.44
DualGATs	GB	40.69	66.90	67.68	58.42
COSMIC	TB	38.11	65.21	65.28	56.2
SPCL-CL	TB	40.94	67.25	69.74	59.31
EACL	TB	40.24	67.12	70.41	59.26
UniMSE	TB (MM)	-	65.51	70.66	-
TelME	TB(MM)	-	67.37	70.48	-
GPT-4o	LLM	39.78	59.18	51.68	50.18
DeepSeek-R1	LLM	30.94	59.51	53.64	48.03
BiosERC	LLM	41.68	69.83	71.19	60.9
InstructERC	LLM	41.37	69.15	**71.39**	60.64
CoE	LLM	**44.29**	**70.11**	70.46	**61.62**

Backbone abbreviations: TB = Transformer-based, MM = Multimodal, RB = Recurrent-based, GB = GNN-based, LLM = Large Language Model.

achieving a 2.92% improvement on the EmoryNLP dataset and a 0.96% improvement on MELD, reflecting its effectiveness in various ERC benchmarks. Notably, despite their remarkable capabilities in logical reasoning and programming tasks, strong LLMs like GPT-4o and DeepSeek-R1 show limited performance in zero-shot ERC. CoE achieved the highest average W-F1 score (+ 0.98%) across the tested datasets, indicating its overall superiority in handling a variety of conversational contexts. These improvements stem from the CoE's ability to progressively integrate conversational clues, such as speaker personas and scenes, which are embedded directly within the task through the multi-stage auxiliary learning strategy. By enhancing its understanding of emotional subtleties and speaker-specific behaviors, CoE surpasses prior methods that rely on external knowledge sources. However, on the IEMOCAP dataset, which consists of scripted, discrete conversations between temporary actors, our method faced challenges due to the absence of persona and scene information, which are essential components for effectively capturing the full emotional context.

5.3 Ablation Studies

5.3.1 Effect of Different LLMs

To further evaluate the performance of LLMs in the ERC task, we conducted experiments using different LLMs, as shown in Figure 3. The experiments focused on the ERC task using the CoE framework without auxiliary tasks, allowing for a direct comparison of the performance of various LLMs.

Among the models tested, Mistral-7B-v0.1 demonstrated the best overall performance across the majority of datasets, leading to its selection as the base model for subsequent ablation studies on multi-stage auxiliary learning.

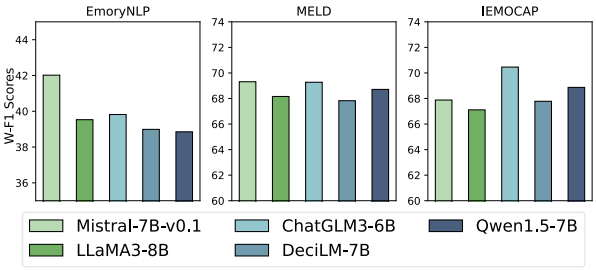

Figure 3: Comparison with Other LLMs

5.3.2 Ablation Study on Textual Clues

In this section, we conduct an ablation study to analyze the impact of textual clues, specifically persona and scene information, on the model's performance within the CoE framework. Importantly, these experiments do not incorporate auxiliary tasks, allowing for a focused analysis of the direct contributions of persona and scene clues. The results are summarized in Table 2, from which we can conclude that the study highlights the significance of persona and scene clues. Removing either persona or scene leads to a noticeable performance drop, indicating their essential role in providing context to the dialogue. The most significant decline occurs when both clues are removed, underscoring their complementary effect on improving the accuracy of the ERC task.

Table 2: Ablation Study on Textual Clues

Methods	EmoryNLP W-F1	MELD W-F1
LLM + Textual Clues	**42.02****	**69.31****
LLM + Textual Clues (w/o Persona)	41.51*	69.13**
LLM + Textual Clues (w/o Scene)	40.71**	-
LLM + Textual Clues (w/o Persona + Scene)	39.95**	-

Note: *p<0.05, **p<0.01 indicate the level of statistical significance.

5.3.3 Ablation Study on Auxiliary Tasks

We also conduct an in-depth analysis of the auxiliary tasks to evaluate their individual contributions. Table 3 illustrates the contribution of each auxiliary task to the main task's performance across both datasets, demonstrating that each task independently adds value to the model's emotion recognition capabilities.

Based on the results, the **speaker identification task** (+1.08% on EmoryNLP, +0.36% on MELD) demonstrates the strongest effect on model performance, likely because it enables the model to link

emotional clues directly to individual personas, capturing subtle shifts in tone that vary between speakers. Although the **role-playing task** (+0.38% on EmoryNLP, +0.41% on MELD) shows relatively smaller gains, its full potential may be realized when integrated with other auxiliary tasks during joint training, where it can enhance the model's understanding of personas. While **STEeR's improvement** (+1.01% on EmoryNLP, +0.27% on MELD) is a little smaller than **speaker's**, its focus on emotion rationale reasoning still plays a key role in interpreting complex emotional states. The smaller improvement in MELD (+0.36%) may be due to the imbalanced distribution of emotion classes, where the model focuses more on the majority of neutral emotions, resulting in limited gains for rarer emotions.

Table 3: Performance Improvement of the Main Task with Each Individual Auxiliary Task

Tasks	EmoryNLP W-F1	MELD W-F1
Emotion CoE Only	42.02**	69.31**
Emotion CoE + Speaker CoE	43.10**	69.67**
Emotion CoE + Role-Playing CoE	42.40*	69.42*
Emotion CoE + Reasoning CoE	43.03**	69.58*

Note: *p<0.05, **p<0.01 indicate the level of statistical significance.

5.4 Train Strategy Study

In this section, we examine various training strategies designed to enhance model performance on the ERC task. The experiments were conducted in two phases: first, we performed a preliminary evaluation of a subset of the data to compare the effectiveness of mixed auxiliary training and curriculum training. Building on these findings, we then developed and tested two refined strategies for comprehensive evaluation.

5.4.1 Phase 1: Preliminary Evaluation of Training Approaches

In this initial phase, we conducted experiments on a subset of the dataset to assess the effectiveness of two key training strategies: curriculum-based sequential training (Soviany et al., 2022; Wang et al., 2021) and mixed auxiliary training (Chen et al., 2022). Given the high computational cost associated with training large models across various auxiliary task combinations, we focused on role-playing and speaker identification tasks using the EmoryNLP dataset for evaluation. These tasks were selected because they address critical aspects of the ERC task, allowing us to more precisely evaluate the impact of different training strategies.

We compare three training strategies for ERC, as shown in Figure 4:

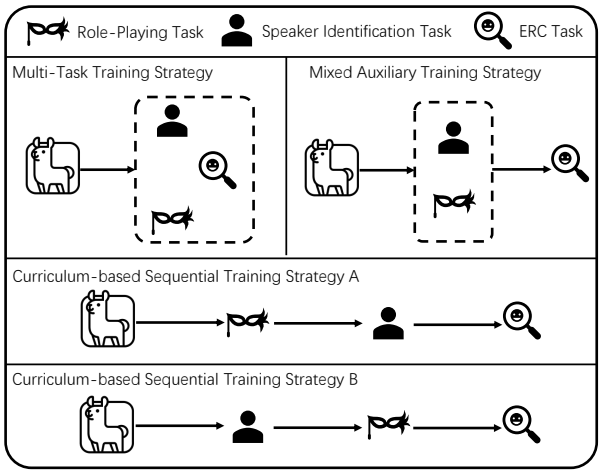

Figure 4: Three Training Strategies for ERC

1. **Curriculum-based Sequential Training Strategy:** Each task is trained in a separate stage, with the ERC task always being the final stage of training.

2. **Mixed Training Strategy:** All auxiliary tasks are merged into a single multi-task dataset for joint training, with a separate stage reserved for fine-tuning the model on the ERC task.

3. **Multi-Task Training Strategy:** All tasks, including the ERC task, are mixed and trained together in a single stage as a multi-task setup.

The results in Table 4 of this small-scale experiment demonstrate that **mixed auxiliary training** consistently outperforms the curriculum-based training approach. This suggests that the mixed training strategy enables more dynamic knowledge sharing between tasks, making it a more suitable approach for the ERC task.

Table 4: Ablation Study on Different Strategies

Models	EmoryNLP W-F1
Fully Mixed Strategy	42.46**
Mixed Auxiliary Strategy	**43.64****
Curriculum-based Strategy A	43.30*
Curriculum-based Strategy B	43.35*

Note: *p<0.05, **p<0.01 indicate the level of statistical significance.

5.4.2 Phase 2: Strategy Refinement and Full Evaluation

One key challenge is that the STEeR method cannot be jointly trained with other datasets due to its unique rationale generation process. This means a straightforward mixed training strategy is not effective for integrating STEeR with other tasks. Based on these findings from the small-scale experiments, we developed two refined strategies for full-scale evaluation to address these issues and optimize the training process.

- **Strategy A**: This strategy follows a three-stage training process. Firstly, the model is trained on a consisting of role-playing and speaker identification tasks. The model is then trained using the STEeR method. Finally, the model is trained for the ERC task.

- **Strategy B**: This approach begins by generating a rationale by using the STEeR method. This enriched dataset allows for joint training on all three auxiliary tasks simultaneously. Once the intensive joint training phase is complete, the model is trained for the ERC task.

Analysis of Strategy Refinement: Strategy A follows a structured, stepwise approach, where each training stage builds upon the previous one, enhancing the model's grasp of speaker identities, conversational roles, and emotional reasoning. This hierarchical progression enables the model to specialize in each aspect before progressing to the next, thereby creating a refined understanding at every level. On the other hand, Strategy B begins by generating a rationale dataset using the STEeR method, which enables the simultaneous joint training of all three auxiliary tasks. This joint approach fosters a more integrated prior model, providing a solid foundation for the ERC task.

We conducted experiments to compare the performance of the model under Strategy A and Strategy B. As shown in Table 5, the results indicate that Strategy B consistently outperformed Strategy A in the final ERC task. By pre-generating a rationale dataset, Strategy B facilitated a more effective integration of auxiliary tasks within its multi-stage auxiliary learning framework, enabling joint training that strengthened the model's overall capability for the ERC task. In contrast, segmentation in Strategy A can lead to weaker knowledge retention in earlier stages, potentially limiting the model's capacity to fully leverage learned representations in the final ERC stage. For a detailed analysis of how different task combinations affect the model's performance, see Appendix D. These findings underscore the significance of integrating reasoning processes early in the training pipeline, which plays a crucial role in enhancing the model's performance across various ERC benchmarks by providing it with robust prior knowledge.

Table 5: Comparison of CoE Strategies A and B

Strategies	EmoryNLP W-F1	MELD W-F1
CoE Strategy A	43.68*	69.89*
CoE Strategy B	**44.29****	**70.11***

Note: *p<0.05, **p<0.01 indicate the level of statistical significance.

6 Conclusion

In this work, we present the CoE framework, which enhances ERC performance by utilizing personas and scene information. The framework combines role-playing, speaker identification, and STEeR to enhance emotional reasoning and understanding. Through multi-stage training strategies, we found that combining auxiliary tasks into a unified multi-task training phase outperforms other strategies, achieving state-of-the-art performance on the EmoryNLP and MELD datasets.

Limitations

The CoE framework, while effective for text-based ERC, currently focuses only on textual modality. Future work could explore the integration of multimodal information, such as visual or acoustic cues, to further enhance emotion recognition capabilities. Additionally, the dataset constraints affected our analysis — IEMOCAP's scripted conversations between temporary actors lack the consistent speaker identities needed for speaker identification tasks, while MELD's incomplete scene information prevented certain ablation experiments. These limitations restricted our ability to fully evaluate the impact of contextual elements on ERC performance. To address these dataset limitations, future research could explore the automatic generation of persona and scene information to enrich contextual understanding in dialogues where such information is limited or missing.

Acknowledgements

This work has been supported by the National Natural Science Foundation of China (62372483). The work of Jianxing Yu has been supported by the National Natural Science Foundation of China (62276279), Guangdong Basic and Applied Basic Research Foundation (2024B1515020032).

References

Scott Brave and Cliff Nass. 2007. Emotion in human-computer interaction. In *The Human-Computer Interaction Handbook*, pages 103–118.

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

Carlos Busso, Murtaza Bulut, Chul Min 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(4):335–359.

Santiago Castro, Devamanyu Hazarika, Verónica Pérez-Rosas, Roger Zimmermann, Rada Mihalcea, and Soujanya Poria. 2019. Towards multimodal sarcasm detection (an obviously perfect paper). In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 4619–4629. Association for Computational Linguistics.

Rajdeep Chatterjee, Saptarshi Mazumdar, R Simon Sherratt, Rohit Halder, Tanmoy Maitra, and Debasis Giri. 2021. Real-time speech emotion analysis for smart home assistants. *IEEE Transactions on Consumer Electronics*, 67(1):68–76.

Hong Chen, Xin Wang, Chaoyu Guan, Yue Liu, and Wenwu Zhu. 2022. Auxiliary learning with joint task and data scheduling. In *International Conference on Machine Learning*, pages 3634–3647. PMLR.

Mayee Chen, Nicholas Roberts, Kush Bhatia, Jue Wang, Ce Zhang, Frederic Sala, and Christopher Ré. 2024. Skill-it! a data-driven skills framework for understanding and training language models. *Advances in Neural Information Processing Systems*, 36.

Yuyan Chen and Yanghua Xiao. 2024. Recent advancement of emotion cognition in large language models. *CoRR*, abs/2409.13354.

DeepSeek-AI, Daya Guo, and Dejian Yang et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

Zixiang Ding, Rui Xia, and Jianfei Yu. 2020. End-to-end emotion-cause pair extraction based on sliding window multi-label learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3574–3583, Online. Association for Computational Linguistics.

Yumeng Fu. 2024. CKERC : Joint large language models with commonsense knowledge for emotion recognition in conversation. *CoRR*, abs/2403.07260.

Qingqing Gao, Biwei Cao, Xin Guan, Tianyun Gu, Xing Bao, Junyan Wu, Bo Liu, and Jiuxin Cao. 2022. Emotion recognition in conversations with emotion shift detection based on multi-task learning. *Knowledge-Based Systems*, 248:108861.

Deepanway Ghosal, Navonil Majumder, Alexander F. Gelbukh, Rada Mihalcea, and Soujanya Poria. 2020. COSMIC: commonsense knowledge for emotion identification in conversations. In *Findings of the Association for Computational Linguistics: EMNLP 2020, Online Event, 16-20 November 2020*, volume EMNLP 2020 of *Findings of ACL*, pages 2470–2481. Association for Computational Linguistics.

Deepanway Ghosal, Navonil Majumder, Soujanya Poria, Niyati Chhaya, and Alexander F. Gelbukh. 2019. Dialoguegcn: A graph convolutional neural network for emotion recognition in conversation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 154–164. Association for Computational Linguistics.

Jia-Chen Gu, Chao-Hong Tan, Chongyang Tao, Zhen-Hua Ling, Huang Hu, Xiubo Geng, and Daxin Jiang. 2022. Hetermpc: A heterogeneous graph neural network for response generation in multi-party conversations. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 5086–5097. Association for Computational Linguistics.

Dou Hu, Yinan Bao, Lingwei Wei, Wei Zhou, and Songlin Hu. 2023. Supervised adversarial contrastive learning for emotion recognition in conversations. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, pages 10835–10852, Toronto, Canada. Association for Computational Linguistics.

Guimin Hu, Ting-En Lin, Yi Zhao, Guangming Lu, Yuchuan Wu, and Yongbin Li. 2022. Unimse: Towards unified multimodal sentiment analysis and emotion recognition. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 7837–7851.

Aaron Hurst, Adam Lerer, and Adam P. Goucher et al. 2024. Gpt-4o system card. *CoRR*, abs/2410.21276.

Liangyi Kang, Jie Liu, Lingqiao Liu, Zhiyang Zhou, and Dan Ye. 2021. Semi-supervised emotion recognition in textual conversation via a context-augmented auxiliary training task. *Information Processing and Management*, 58(6):102717.

Po-Nien Kung, Sheng-Siang Yin, Yi-Cheng Chen, Tse-Hsuan Yang, and Yun-Nung Chen. 2021. Efficient multi-task auxiliary learning: selecting auxiliary data by feature similarity. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 416–428.

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*.

Chenbing Li, Jie Hu, Tianrui Li, Shengdong Du, and Fei Teng. 2023. An effective multi-task learning model for end-to-end emotion-cause pair extraction. *Applied Intelligence*, 53(3):3519–3529.

Jingye Li, Meishan Zhang, Donghong Ji, and Yijiang Liu. 2020. Multi-task learning with auxiliary speaker identification for conversational emotion recognition. *CoRR*, abs/2003.01478.

Yingjian Liu, Jiang Li, Xiaoping Wang, and Zhigang Zeng. 2024. Emotionic: emotional inertia and contagion-driven dependency modeling for emotion recognition in conversation. *Science China Information Sciences*, 67(8):1–17.

Andrea Madotto, Zhaojiang Lin, Chien-Sheng Wu, and Pascale Fung. 2019. Personalizing dialogue agents via meta-learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5459, Florence, Italy. Association for Computational Linguistics.

Navonil Majumder, Soujanya Poria, Devamanyu Hazarika, Rada Mihalcea, Alexander F. Gelbukh, and Erik Cambria. 2019. Dialoguernn: An attentive RNN for emotion detection in conversations. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 6818–6825. AAAI Press.

Scott Monteith, Tasha Glenn, John Geddes, Peter C Whybrow, and Michael Bauer. 2022. Commercial use of emotion artificial intelligence (ai): implications for psychiatry. *Current Psychiatry Reports*, 24(3):203–211.

OpenAI, Ahmed El-Kishky, Alexander Wei, Andre Saraiva, Borys Minaev, Daniel Selsam, David Dohan, Francis Song, Hunter Lightman, Ignasi Clavera, Jakub Pachocki, Jerry Tworek, Lorenz Kuhn, Lukasz Kaiser, Mark Chen, Max Schwarzer, Mostafa Rohaninejad, Nat McAleese, o3 contributors, Oleg Mürk, Rhythm Garg, Rui Shu, Szymon Sidor, Vineet Kosaraju, and Wenda Zhou. 2025. Competitive programming with large reasoning models. *Preprint*, arXiv:2502.06807.

Sudarshan Pant, Eunchae Lim, Hyung-Jeong Yang, Guee-Sang Lee, Soo-Hyung Kim, Young-Shin Kang, and Hyerim Jang. 2022. Korean drama scene transcript dataset for emotion recognition in conversations. *IEEE Access*, 10:119221–119231.

Avinash Patil. 2025. Advancing reasoning in large language models: Promising methods and approaches. *Preprint*, arXiv:2502.03671.

Soujanya Poria, Erik Cambria, Devamanyu Hazarika, Navonil Majumder, Amir Zadeh, and Louis-Philippe Morency. 2017. Context-dependent sentiment analysis in user-generated videos. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pages 873–883.

Soujanya Poria, Navonil Majumder, Devamanyu Hazarika, and Rada Mihalcea. 2019. Meld: A multimodal multi-party dataset for emotion recognition in conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 527–536.

Xiangyu Qin, Zhiyu Wu, Tingting Zhang, Yanran Li, Jian Luan, Bin Wang, Li Wang, and Jinshi Cui. 2023. Bert-erc: Fine-tuning bert is enough for emotion recognition in conversation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 13492–13500.

Sahand Sabour, Siyang Liu, Zheyuan Zhang, June M. Liu, Jinfeng Zhou, Alvionna S. Sunaryo, Tatia M. C. Lee, Rada Mihalcea, and Minlie Huang. 2024. Emobench: Evaluating the emotional intelligence of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 5986–6004. Association for Computational Linguistics.

Yunfan Shao, Linyang Li, Junqi Dai, and Xipeng Qiu. 2023. Character-LLM: A trainable agent for role-playing. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13153–13187, Singapore. Association for Computational Linguistics.

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, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 5197–5206. Association for Computational Linguistics.

Petru Soviany, Radu Tudor Ionescu, Paolo Rota, and Nicu Sebe. 2022. Curriculum learning: A survey. *International Journal of Computer Vision*, 130(6):1526–1565.

Selvarajah Thuseethan, Sutharshan Rajasegarar, and John Yearwood. 2022. Emosec: Emotion recognition from scene context. *Neurocomputing*, 492:174–187.

Geng Tu, Feng Xiong, Bin Liang, and Ruifeng Xu. 2024. A persona-infused cross-task graph network for multimodal emotion recognition with emotion shift detection in conversations. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2024, Washington DC, USA, July 14-18, 2024*, pages 2266–2270. ACM.

Noah Wang, Z. y. Peng, Haoran Que, Jiaheng Liu, Wangchunshu Zhou, Yuhan Wu, Hongcheng Guo, Ruitong Gan, Zehao Ni, Jian Yang, Man Zhang, Zhaoxiang Zhang, Wanli Ouyang, Ke Xu, Wenhao Huang, Jie Fu, and Junran Peng. 2024. Rolellm: Benchmarking, eliciting, and enhancing role-playing abilities of large language models. In *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, pages 14743–14777. Association for Computational Linguistics.

Xin Wang, Yudong Chen, and Wenwu Zhu. 2021. A survey on curriculum learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(9):4555–4576.

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.

Jieying Xue, Minh-Phuong Nguyen, Blake Matheny, and Le-Minh Nguyen. 2024. Bioserc: Integrating biography speakers supported by llms for ERC tasks. In *Artificial Neural Networks and Machine Learning - ICANN 2024 - 33rd International Conference on Artificial Neural Networks, Lugano, Switzerland, September 17-20, 2024, Proceedings, Part V*, volume 15020 of *Lecture Notes in Computer Science*, pages 277–292. Springer.

Kailai Yang, Shaoxiong Ji, Tianlin Zhang, Qianqian Xie, Ziyan Kuang, and Sophia Ananiadou. 2023. Towards interpretable mental health analysis with large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 6056–6077. Association for Computational Linguistics.

Fangxu Yu, Junjie Guo, Zhen Wu, and Xinyu Dai. 2024. Emotion-anchored contrastive learning framework for emotion recognition in conversation. In *Findings of the Association for Computational Linguistics: NAACL 2024, Mexico City, Mexico, June 16-21, 2024*, pages 4521–4534. Association for Computational Linguistics.

Taeyang Yun, Hyunkuk Lim, Jeonghwan Lee, and Min Song. 2024. Telme: Teacher-leading multimodal fusion network for emotion recognition in conversation. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies NAACL 2024, Mexico City, Mexico, June 16-21, 2024*, pages 82–95. Association for Computational Linguistics.

Samira M Zahiri and Jinho D Choi. 2018. Emotion recognition in conversations with transfer learning from generative conversation modeling. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 27–36.

Eric Zelikman, Yuhuai Wu, and Noah D Goodman. 2022. Star: Self-taught reasoner. *arXiv preprint arXiv:2203.14465*.

Duzhen Zhang, Feilong Chen, and Xiuyi Chen. 2023a. Dualgats: Dual graph attention networks for emotion recognition in conversations. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, pages 7395–7408.

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

Yazhou Zhang, Mengyao Wang, Prayag Tiwari, Qiuchi Li, Benyou Wang, and Jing Qin. 2023b. Dialoguellm: Context and emotion knowledge-tuned llama models for emotion recognition in conversations. *CoRR*, abs/2310.11374.

A Details of CoE

In this study, we develop a CoE framework to fine-tune pre-trained language models for emotion-related tasks through supervised fine-tuning. Each item in the square bracket represents a key element, with the following descriptions:

- **[System]:** This part outlines the type of task the participant is required to perform (i.e., ERC, role-playing, or speaker identification), as well as the role the large language model needs to assume during the task.
- **[Speaker_Persona]:** This part provides detailed descriptions of the character's personality traits, professional backgrounds, and roles within the group. These descriptions help in establishing a clearer understanding of each character's persona.
- **[Scene]:** This section describes the specific location and environment where the dialogue takes place, such as "Monica and Rachel's apartment." Setting the scene is crucial for building the conversation context.
- **[History_Dialogue]:** This part records prior interactions between the characters, allowing the model to better understand the current conversational context and the dynamics between participants.
- **[Predict]:** This section provides instructions to guide the model in making predictions. For the ERC task, this might involve predicting the emotion of a specific dialogue segment; for role-playing, it entails predicting what the character will say next; and for speaker identification, it involves predicting the speaker of a particular utterance.
- **[Target]:** This represents the model's expected output, such as the prediction results that the model needs to learn.

B Dialogue Processing Steps

To process dialogue segments from various scenarios, we follow these steps. In our notation, C_i represents the i-th dialogue segment in a conversation C, and $u_{i,j}$ denotes the j-th utterance in segment C_i. We define P as the predefined subset of speakers we wish to retain from the complete speaker set S.

The processing consists of two main steps:

1. For each dialogue segment C_i, we remove the first utterance $u_{i,1}$:

$$F_1(C_i) = C_i \setminus \{u_{i,1}\} \tag{4}$$

2. We retain only those dialogue segments where the speaker of the last utterance is in set P:

$$F_2(C_i) = \begin{cases} C_i & \text{if speaker of last utterance} \in P \\ \emptyset & \text{otherwise} \end{cases} \tag{5}$$

Combining the above steps, the processed dialogue set $P(C)$ is formed by:

$$P(C) = \bigcup_{i=1}^{n} F_2(F_1(C_i)) \tag{6}$$

where n is the total number of dialogue segments in conversation C.

This indicates that for each dialogue segment C_i, we first discard its first utterance (applying F_1), then check if the speaker of the last utterance is in our predefined set P (applying F_2). If so, we keep the entire segment; otherwise, we discard it. Finally, we merge all retained segments to form the final dialogue set $P(C)$.

C Initial Data Generation Phase

The initial data generation phase, as illustrated in Algorithm 2, describes how we generate the first batch of rationales using GPT-4o. In this process, we iterate through the dataset and generate rationales for each sample. When the generated emotion label matches the ground truth, we add the corresponding rationale to our training dataset. This approach ensures that we only use high-quality rationales for training the model. After each iteration, we remove the processed samples from the original dataset to avoid redundant processing.

D Analysis of Task Combinations

To investigate the contribution of different task combinations, we conducted experiments comparing three training strategies.

Our experiment shows that using only the Speaker Task followed by the STEeR Task and the ERC task yields a 43.43% W-F1 score on EmoryNLP. Similarly, using only the Role-Playing Task followed by the STEeR Task and the ERC task achieves 43.10%. However, combining both Speaker and Role-Playing Tasks before the STEeR

Algorithm 2: Pre Data Generation Process

Input: GPT Model $GPT4$, Dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{D}$

$\mathcal{D}_0 \leftarrow \mathcal{D}$, $\mathcal{D}_{\text{train}} \leftarrow \emptyset$ // Initialize a temporary dataset and the training dataset

for $n \in 1 \dots N$ **do**

  for $(x_i, y_i) \in \mathcal{D}_{n-1}$ **do**

    $(\hat{r}_i, \hat{y}_i) \leftarrow GPT4(x_i, y_i) \quad \forall (x_i, y_i)$ // Perform rationale generation

    $\mathcal{D}_{\text{train}} \leftarrow \mathcal{D}_{\text{train}} \cup \{(x_i, \hat{r}_i, y_i) \mid \hat{y}_i = y_i\}$ // Add correct rationales

  end

  $\mathcal{D}_n \leftarrow \mathcal{D}_{n-1} \setminus \mathcal{D}_{\text{train}}$ // Update dataset by removing generated samples

end

Task produces the best performance (44.29%). The results indicate that while both speaker identification and role-playing tasks show meaningful contributions when paired with the STEeR task, the combination of both tasks achieves the best performance. This suggests that integrating role-playing with speaker identification creates a synergistic effect, improving the model's ability to better understand personas, including the characters' personalities and conversational dynamics. This enhances the model's overall emotion recognition ability, as the role-playing task complements the speaker identification task by focusing on different aspects of the speaker's identity and interaction style.

The findings suggest that the full potential of the role-playing task is realized when it is integrated with other tasks. Further evaluation of the auxiliary task could clarify its contribution to overall performance and justify its role in the training strategy.

E Prompts Settings

In this section, we present the detailed prompt templates used for each task in the CoE framework. These prompts illustrate how we structured the inputs for the LLM to perform different tasks. Each prompt follows a consistent format with sections for system instructions, speaker persona information, scene description, dialogue history, prediction task, and target output. The following tables demonstrate the specific prompts used for the main ERC task and each auxiliary task (speaker identification, role-playing, and emotion reasoning). These prompts show how we systematically incorporated textual clues and task-specific instructions to guide the model's learning process.

Prompt for ERC	
[System]	As an expert in Conversational Emotion Recognition (ERC), your task is to [Predict] the emotion by analyzing the [Speaker_Persona], [Scene] and [History_dialogue] information.
[Speaker_Persona]	Chandler: Known for wit and sarcasm, works in advertising, group's humorist with logical insights. Phoebe: A masseuse and musician, eccentric with unique wisdom, offering empathy to the group. Ross: A paleontologist and professor, adds intellectual depth, connections through family ties with Monica and past relationship with Rachel. Rachel: Progresses from a waitress to fashion, influences group dynamics, evolving relationship with Ross. Monica: A chef, known for organization, group's anchor with hospitality, Ross's sister.
[Scene]	Monica and Rachel's place, the entire gang is present.
[History_Dialogue]	Monica: "So ah, Phoebe, how was your date?" Phoebe: "Oh well y'know." Monica: "Yeah, I do know." ... Chandler: "Don't worry." Phoebe: "God, I hope they kick his ass!"
[Predict]	Please select the emotional label of <Phoebe: "God, I hope they kick his ass!">from <Joyful, Mad, Peaceful, Neutral, Sad, Powerful, Scared>.
[Target]	Powerful

Prompt for Speaker Recognition Task	
[System]	You are now an expert in Conversational Emotion Recognition. Your task is to predict the speaker by analyzing sentences. Here's a list of individuals and their [Speaker_Persona], which will assist you in understanding them better
[Speaker_Persona]	Chandler: Known for his wit and sarcasm, works in advertising, group's humorist with logical insights. Ross: A paleontologist and professor, adds intellectual depth through family ties with Monica and past relationship with Rachel. Joey: An aspiring actor, the group's innocent and charming member, deepening bond with Chandler through loyalty and simplicity. Phoebe: A masseuse and musician, introduces eccentricity and unique wisdom, offering empathy to the group. Monica: A chef, known for organization, group's anchor with hospitality, directly linking to Ross as his sister. Rachel: Progresses from a waitress to fashion, influencing group dynamics, especially her evolving relationship with Ross.
[Predict]	Please select the Speaker of <"God, I hope they kick his ass!", Emotion: Powerful >from <Chandler, Ross, Joey, Phoebe, Monica, Rachel>.
[Target]	Phoebe

Prompt for Role-play Task	
[System]	You are now an expert role-play actor. Analyze the character's [Speaker_Persona], [Scene], and [History_dialogue] to predict their most likely next line.
[Speaker_Persona]	Monica: Known for organization, acts as the group's anchor, fostering cohesion with hospitality, linking to Ross as his sister. Rachel: Progresses from a waitress to a career in fashion, influencing group dynamics, especially her relationship with Ross. Ross: A paleontologist and professor, adds intellectual depth through family ties with Monica and past relationship with Rachel. Chandler: Known for wit and sarcasm, works in advertising, the group's humorist with logical insights. Phoebe: A masseuse and musician, introduces eccentricity and unique wisdom, offering empathy to the group.
[Scene]	Monica and Rachel's, The entire gang is there.
[History_Dialogue]	Monica: "So ah, Phoebe, how was your date?" Phoebe: "Oh well y'know." Monica: "Yeah, I do know." ... Chandler: "Don't worry."
[Predict]	You will stay in character whenever possible, and generate responses in <Powerful>emotion as if you were <Phoebe>.
[Target]	God, I hope they kick his ass!

Prompt for Reasoning Task (STEeR)	
[System]	As an expert in Conversational Emotion Recognition (ERC), your task is to [Predict] the emotion by analyzing the [Speaker_Persona], [Scene] and [History_dialogue] information.
[Speaker_Persona]	Monica: Known for organization, acts as the group's anchor, fostering cohesion with hospitality, linking to Ross as his sister. Rachel: Progresses from a waitress to a career in fashion, influencing group dynamics, especially her relationship with Ross. Ross: A paleontologist and professor, adds intellectual depth through family ties with Monica and past relationship with Rachel. Chandler: Known for wit and sarcasm, works in advertising, the group's humorist with logical insights. Phoebe: A masseuse and musician, introduces eccentricity and unique wisdom, offering empathy to the group.
[Scene]	Monica and Rachel's, The entire gang is there.
[History_Dialogue]	Monica: "So ah, Phoebe, how was your date?" Phoebe: "Oh well y'know." Monica: "Yeah, I do know." ... Chandler: "Don't worry."
[Predict]	Please analyze the emotional state of <Phoebe: "God, I hope they kick his ass!">by first providing your rationale based on the given context, then select the emotional label from <Joyful, Mad, Peaceful, Neutral, Sad, Powerful, Scared>.
[Target]	[Rationale] In the dialogue, Phoebe starts feeling vulnerable about her date but transitions to confidence after receiving support from her friends. Her assertive statement, indicating a desire for action and control, reflects an empowered emotional state. The emotional label of <Phoebe: "God, I hope they kick his ass!">is <Powerful>.