Do LLMs Feel? Teaching Emotion Recognition with Prompts, Retrieval, and Curriculum Learning

Xinran Li, Yu Liu, Jiaqi Qiao, Xiujuan Xu*

School of Software Technology, Dalian University of Technology
Dalian, China
963707605@mail.dlut.edu.cn, yuliu@dlut.edu.cn, qiaobright@mail.dlut.edu.cn, xjxu@dlut.edu.cn

Abstract

Emotion Recognition in Conversation (ERC) is a crucial task for understanding human emotions and enabling natural human-computer interaction. Although Large Language Models (LLMs) have recently shown great potential in this field, their ability to capture the intrinsic connections between explicit and implicit emotions remains limited. We propose a novel ERC training framework, **PRC-Emo**, which integrates **P**rompt engineering, demonstration **R**etrieval, and **C**urriculum learning, with the goal of exploring whether LLMs can effectively perceive emotions in conversational contexts. Specifically, we design emotion-sensitive prompt templates based on both explicit and implicit emotional cues to better guide the model in understanding the speaker's psychological states. We construct the first dedicated demonstration retrieval repository for ERC, which includes training samples from widely used datasets, as well as high-quality dialogue examples generated by LLMs and manually verified. Moreover, we introduce a curriculum learning strategy into the LoRA fine-tuning process, incorporating weighted emotional shifts between same-speaker and different-speaker utterances to assign difficulty levels to dialogue samples, which are then organized in an easy-to-hard training sequence. Experimental results on two benchmark datasets—IEMOCAP and MELD—show that our method achieves new state-of-the-art (SOTA) performance, demonstrating the effectiveness and generalizability of our approach in improving LLM-based emotional understanding.

Code — https://github.com/LiXinran6/PRC-Emo
Extended Version — https://github.com/LiXinran6/PRC-Emo/blob/main/PRC_Emo_withAppendix.pdf

Introduction

"Emotions play a critical role in human intelligence. For truly intelligent machines, emotional intelligence is not optional," based on the ideas of Rosalind W. Picard, the pioneer of affective computing. As conversational agents (Lee et al. 2020) and large language models (LLMs) become increasingly integrated into our daily lives, it is essential that these systems not only understand syntax and semantics but also perceive and interpret human emotions. Emotion Recognition in Conversation (ERC) emerges as a crucial task toward building emotionally-aware AI, enabling natural and empathetic human-computer interaction (Hu et al. 2024).

Large Language Models, with their generative architecture, have achieved significant performance improvements across various natural language processing tasks (Laskar et al. 2024). Currently, the field of ERC has also entered the era of LLMs, with increasing research efforts attempting to leverage LLMs to model conversational context and predict speakers' emotional states. By harnessing the powerful prior knowledge of LLMs, researchers incorporate contextual utterance information, speaker personality, and other background knowledge to design high-quality prompts, combined with parameter-efficient fine-tuning methods such as LoRA (Hu et al. 2021b) for adaptive training. Although LLMs show great promise in ERC tasks, they still face the following three challenges:

(1) Most existing studies have not adequately considered both explicit and implicit emotions when designing prompts. Emotion recognition tasks can be divided into explicit emotion recognition and implicit emotion recognition (Koga, Kando, and Miyao 2024), as illustrated in Figure 1. Explicit emotion recognition focuses on predicting emotions conveyed by the speaker through direct expressions, strong intonations, or obvious facial cues, while implicit emotion recognition aims to capture emotions genuinely felt by the speaker but not necessarily expressed through language. Ignoring the balance between these two aspects limits the model's ability to fully comprehend complex emotional states, thereby constraining the accuracy of emotion recognition.

(2) Current research often employs demonstration retrieval in prompt design, extracting the most similar sentences and their corresponding labels from a demonstration repository to assist the model in understanding and predicting emotions (Lei et al. 2023). However, existing methods mostly construct the demonstration repository using only the training sets of commonly used datasets, which limits the diversity and coverage of demonstration samples. Due to the relatively homogeneous content of the repository, the model's reasoning is often confined to the context of specific datasets, lacking sufficient generalization ability. This restricts the model's adaptability to more complex and di-

*Corresponding author.

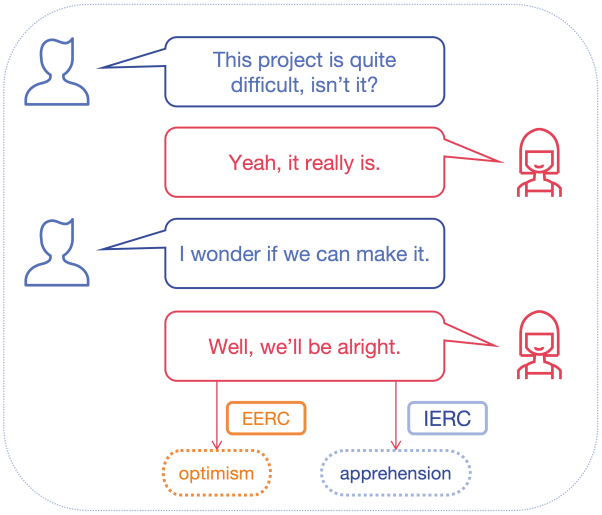

Figure 1: An example of Explicit Emotion Recognition in Conversation (EERC) and Implicit Emotion Recognition in Conversation (IERC). In this conversation, although the girl feels apprehension due to the boy's anxious utterance, "I wonder if we can make it," she expresses optimism in her own response to encourage both of them.

verse conversational scenarios, thereby limiting its practical effectiveness.

(3) At present, most studies mainly focus on optimizing prompt design and fine-tuning strategies, with relatively little attention paid to improving the overall training process (Yang et al. 2022). These methods often involve complex prompt designs and multi-stage training strategies, resulting in high computational resource consumption but limited performance gains. Moreover, ERC datasets generally suffer from severe class imbalance, making it difficult for models to adequately learn the features of low-frequency categories, which in turn affects overall performance and generalization ability. Therefore, optimizing training strategies to enhance model performance on imbalanced datasets remains an urgent and important challenge.

To address the above challenges, we propose PRC-Emo: a novel Prompt-Retrieval-Curriculum framework for ERC. Leveraging LLMs, our method takes both the current utterance and its dialogue history as input to generate two types of emotion interpretations: explicit (directly expressed emotions) and implicit (underlying, unspoken emotions). These complementary signals are integrated into structured prompts, enabling the model to better capture the nuanced emotional states of speakers. We further introduce the first ERC-specific demonstration retrieval repository, which combines high-quality samples from multiple datasets (IEMOCAP, MELD, EmoryNLP) and over 10,000 additional utterances generated by LLMs and refined by human annotators. These samples span six real-world domains—healthcare, workplace, education, family, social, and entertainment—offering greater diversity and contextual richness for few-shot prompting. We improve the Curriculum Learning strategy by defining a dialogue difficulty metric that combines weighted emotional shifts both within the same speaker and between different speakers. Based on this metric, we divide the training data into multiple levels (i.e., "training buckets") from easy to hard, enabling the model to learn progressively. The main contributions are:

- We propose a method that generates explicit and implicit emotion interpretations as external knowledge in prompts, enabling a better capture of the speaker's true emotional state.
- We construct the first demonstration retrieval repository for ERC that combines multi-source data and LLM-generated, human-refined samples, significantly enhancing the model's generalization ability.
- We enhance a curriculum learning strategy based on dialogue difficulty, using dynamic training buckets for progressive learning and better robustness.
- Experiments on multiple benchmarks show our method achieves consistent SOTA results. The code and data have already been released on GitHub.

Related Work

With the rise of LLMs, the field of ERC can now be broadly divided into two categories: traditional methods and LLM-based methods. This section provides an overview of the methods previously used, as well as the application of curriculum learning in ERC.

Traditional Methods

Traditional research in the field of ERC has mainly focused on three directions: Recurrent Neural Networks (RNNs), Graph Neural Networks (GNNs), and Pre-trained Language Models (PLMs).

In RNN-based studies, models typically leverage RNNs to perform sequential modeling of utterances or entire conversations. DialogueRNN (Majumder et al. 2019) utilizes an RNN structure to dynamically track the individual state of each speaker throughout the conversation. DialogueCRN (Hu, Wei, and Huai 2021) builds upon this by incorporating cognitive factors to enhance the understanding of both contextual and speaker-level information. In GNN-based studies, models construct graph structures to capture dependencies between utterances and speakers. DialogueGCN (Ghosal et al. 2019) treats utterances as nodes in a graph and builds edges based on contextual information to represent semantic relationships. DAG-ERC (Shen et al. 2021) further considers speaker identity and utterance position when constructing a directed acyclic graph to model dialogue structures more precisely. In the area of pre-trained language models, models such as BERT (Devlin et al. 2019) and RoBERTa (Liu et al. 2019) have been widely adopted due to their powerful language representation capabilities. BERT-ERC (Qin et al. 2023) enhances model performance by introducing guided texts, fine-grained emotion classification modules, and a two-stage training strategy.

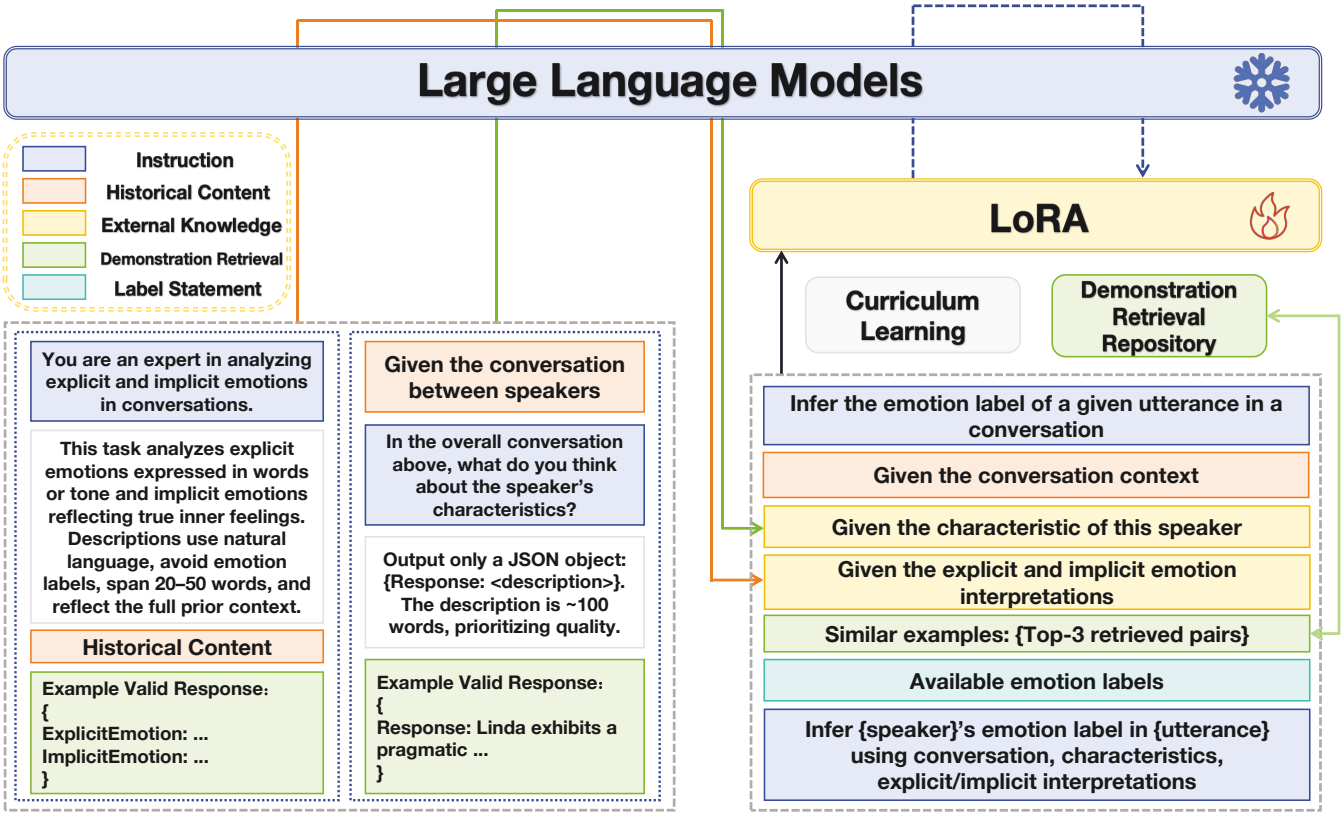

Figure 2: PRC-Emo's architecture has two main stages: extracting external supplementary knowledge and predicting emotion labels, with curriculum learning applied during training. The two prompts at the bottom-left extract explicit and implicit emotion interpretations and speaker characteristics as external knowledge. This information is passed to the bottom-right prompt, which performs the final emotion recognition by retrieving similar pairs from a retrieval repository to aid the process.

LLM-based Methods

Traditional ERC methods predominantly adopt discriminative architectures. With the rise of LLMs, InstructERC (Lei et al. 2023) is the first to introduce a generative architecture for tackling ERC tasks, opening up a novel research direction. This approach incorporates an emotion-alignment auxiliary task and a retrieval-augmented prompting module to enhance the model's understanding of emotions. Moreover, BiosERC (Xue et al. 2024) enriches prompts with background information, such as speaker characteristics, and applies LoRA for the efficient fine-tuning of large models. These LLM-based approaches have significantly improved performance across various ERC benchmarks, demonstrating powerful generalization and adaptability, and paving the way for future advancements in the field.

Curriculum Learning

Curriculum Learning (CL) (Bengio et al. 2009), a training strategy that simulates human learning, has gained widespread attention in ERC in recent years. Yang et al. (2022) proposed a Hybrid Curriculum Learning framework that combines difficulty designs at both the conversation and utterance levels, guiding the model to gradually learn complex emotions based on the frequency of emotion shifts and emotional similarity. Nguyen et al. (2024) introduced the MultiDAG+CL method, which integrates textual, acoustic, and visual features and leverages curriculum learning to address challenges related to emotional variations and data imbalance. Li, Xu, and Qiao (2025) introduce the concept of weighted emotion shifts, with a difficulty design that focuses on modeling transitions between similar emotions. However, these methods only consider emotion transitions within the same speaker and do not take into account transitions between different speakers.

Methodology

This section presents the PRC-Emo architecture. We first define the problem, then describe the extraction of explicit and implicit emotion interpretations as external knowledge. Next, we detail the prompt retrieval template module and its ERC-specific demonstration retrieval repository. Finally, we explain the curriculum learning strategy used in training. The overall architecture of the PRC-Emo is illustrated in Figure 2.

Problem Definition

In Emotion Recognition in Conversation, a conversation is represented as a sequence of utterances $D =$

$\{u_1, u_2, u_3, ..., u_N\}$, where N denotes the total number of utterances. Each utterance u_{i,s_j} is spoken by a specific speaker s_j, indicating that the i^{th} utterance is from speaker s_j. The objective of ERC is to assign an emotion label $y_k \in Y$, such as joy or sadness, to each utterance u_i, where Y is the set of possible emotion categories.

Extraction of External Supplementary Knowledge

Given the extensive knowledge and powerful language understanding abilities of LLMs, we adopt a Prompt Engineering approach by designing query templates to extract supplementary information valuable for ERC tasks. Inspired by Koga, Kando, and Miyao (2024), we observe that the current ERC field generally lacks clear definitions and systematic interpretations of explicit emotion and implicit emotion. Traditional models typically infer the emotion of an utterance based solely on the textual content of the current and historical utterances, without truly understanding the speaker's internal emotional state.

To address this, we design a set of query templates specifically for generating explicit and implicit emotion interpretations. These interpretations serve as high-quality auxiliary signals and are injected into the model input, guiding the model to attend to both the speaker's expressed emotions and their underlying emotional state. This enhances the model's ability to capture multi-layered emotional expressions. In addition, inspired by Xue et al. (2024), we also extract speaker characteristic as supplementary knowledge to further support emotion recognition. These external supplementary knowledge sources are incorporated into the Retrieval Template Module for the final emotion recognition. For the interpretations of explicit emotion and implicit emotion, we input the historical utterances. For the speaker characteristic, we use the entire dialogue. The simplified prompt design is shown in the bottom-left corner of Figure 2.

Retrieval Template Module

To better leverage LLMs, we construct a carefully designed Retrieval Template Module and fine-tune the model using LoRA technology. In this section, we introduce the components of the Retrieval Template Module, as well as the first demonstration retrieval repository specifically built for the ERC task. The simplified prompt design is shown in the bottom-right corner of Figure 2.

Components of Retrieval Template Module Each input consists of an Instruction, Historical Content, External Knowledge, Demonstration Retrieval, and a Label Statement.

- **Instruction**. Defines the model's role and core task objectives.
- **Historical Content**. In order to determine the emotion of a given utterance, it is necessary to provide the model with historical context. We adopt a history window w, which includes the past and current w utterances along with the corresponding speaker names.
- **External Knowledge**. It consists of two parts: speaker characteristic descriptions and expert-level interpretations of explicit/implicit emotions. The speaker characteristics and expert emotion interpretations vary across different conversations.
- **Demonstration Retrieval**. Numerous studies have demonstrated the importance of demonstration retrieval (an Luo et al. 2024). To extract utterances most similar to the current one, we select three ERC examples from our self-constructed demonstration retrieval repository D_{domain}. We utilize $SBERT$ (Reimers and Gurevych 2019) to retrieve relevant demonstrations from the repository. By calculating the cosine similarity and comparing with all vectors in D_{domain}, the top three demonstration-label pairs with the highest scores are selected.
- **Label Statement**. We constrain the model output to a limited set of labels based on the specific dataset used.

Demonstration Retrieval Repository To address the issue of imbalanced emotion label distribution in emotion recognition tasks, we construct an emotion dialogue augmentation dataset based on OpenAI's GPT-4o, covering a wide range of real-life scenarios. The dataset spans common contexts such as healthcare, workplace, education, family, social interactions, and entertainment, and includes five emotion categories: happiness, neutral, fear, sadness, and anger.

We adopt a two-stage prompt strategy for data generation. First, the model produces 30 subtopics under a given scenario to increase diversity. Then, it generates coherent two-person dialogues for these subtopics, with each utterance assigned an emotion label. This process enables targeted creation of underrepresented emotions while preserving contextual consistency, resulting in a more balanced and diverse corpus for downstream ERC models.

For annotation, we adopt a "label masking + human verification" strategy. GPT-4o first generates text samples with emotion labels, after which the labels are removed. Then, two researchers independently annotate each utterance with an emotion category. A sample is officially included in the augmented dataset only if both annotators' judgments exactly match the original label. After one round of filtering, we identify emotion categories still lacking sufficient samples and generate additional data accordingly. This process of generation and filtering repeats three times to finally obtain a high-quality emotion dataset with the desired distribution. Appendix A provides the detailed composition of our self-constructed dataset along with example sentences.

Based on our self-constructed dataset and the training sets of three widely-used ERC benchmarks—IEMOCAP, MELD, and EmoryNLP—we build the first demonstration retrieval repository in the ERC domain. This repository contains a total of 36,712 utterances, including 14,009 from our dataset, 5,163 from IEMOCAP, 9,989 from MELD, and 7,551 from EmoryNLP. Each utterance in the repository includes the following information: the original text, emotion label, source dataset, dialogue ID, sentence position within the dialogue, and its vector representation. By default, we use $SBERT$ to encode the utterances into high-quality semantic embeddings. This demonstration retrieval repository provides a unified, structured, and high-quality data foun-

dation for future research on emotion-aware retrieval and prompt engineering.

Curriculum Learning

Curriculum learning (Bengio et al. 2009) trains models from easy to hard. This section presents our difficulty measure and training scheduler.

Difficulty Measure Function Previous studies (Nguyen et al. 2024; Li, Xu, and Qiao 2025) use emotional shifts as difficulty metrics but overlook transitions between different speakers. We address this by introducing a difficulty function based on weighted emotional shift frequency that accounts for emotional similarity, enabling more accurate conversation complexity estimation.

The weighted emotional shift is defined as the emotional change between two consecutive utterances. Based on the two-dimensional arousal-valence emotion wheel, each emotion label corresponds to a point on the unit circle, covering all emotion categories in the ERC dataset. The similarity between different emotions is calculated by Equation (1):

$$s_{ij} = \begin{cases} \max(\cos(\theta_{ij}), 0) & \text{if } \mathbf{v}_i \cdot \mathbf{v}_j > 0 \\ 0 & \text{if } \mathbf{v}_i \cdot \mathbf{v}_j < 0 \\ \frac{1}{N} & \text{if } \mathbf{v}_i \cdot \mathbf{v}_j = 0 \end{cases} \tag{1}$$

Here, s_{ij} denotes the similarity between emotion labels i and j, $\mathbf{v}_i$ represents the valence vector of label i, θ_{ij} is the angle between labels i and j, and N is the total number of emotion labels in the dataset. The closer two emotions are, the higher their similarity score. The weighted emotional shift (WES) is then defined as shown in Equation (2):

$$N^{WES} = k \times s_{ij} + b \tag{2}$$

We apply a linear transformation to s_{ij}, where k is a weighting factor and b is a bias term. Thus, the difficulty of a conversation c_i is defined as shown in Equation (3), (4) and (5):

$$DIF(c_i) = \frac{WES_{same}(c_i) + WES_{diff}(c_i) + N_{sp}(c_i)}{N_u(c_i) + N_{sp}(c_i)} \tag{3}$$

$$WES_{same}(c_i) = \sum_{j=1}^{N_{shift}^{same}(c_i)} N_j^{WES} \tag{4}$$

$$WES_{diff}(c_i) = \sum_{j=1}^{N_{shift}^{diff}(c_i)} N_j^{WES} \tag{5}$$

where $N_{shift}^{same}(c_i)$ and $N_{shift}^{diff}(c_i)$ represents the total number of emotional shifts between utterances from the same speaker and between consecutive utterances from different speakers in conversation c_i. $N_u(c_i)$ is the total number of utterances. $N_{sp}(c_i)$ denotes the number of speakers appearing in conversation c_i, which serves as a smoothing factor. N_j^{WES} is the Weighted Emotional Shift at the j-th emotional shift. The proposed algorithm is presented in Algorithm 1.

Algorithm 1: Curriculum Learning Training with Difficulty Measure Function

Require:
Input: Training dataset D, Model M, Number of buckets n, Training epochs t
Parameters: Linear transformation coefficients (k, b) for WES calculation
Ensure: Trained model M^*
// Phase 1: Calculate difficulty for each conversation
for each conversation $c_i \in D$ **do**
  Initialize: $WES_{same} \leftarrow 0$, $WES_{diff} \leftarrow 0$, $N_{sp} \leftarrow 0$, $N_u \leftarrow 0$
  $S \leftarrow \emptyset$ {Speaker emotion sequences}
  // Build speaker emotion sequences
  for each utterance u_j in c_i **do**
    $speaker\_id \leftarrow$ get_speaker(u_j)
    $emotion \leftarrow$ get_emotion(u_j)
    $S[speaker\_id] \leftarrow S[speaker\_id] \cup \{emotion\}$
    $N_u \leftarrow N_u + 1$
  end for
  $N_{sp} \leftarrow |S|$ {Number of unique speakers}
  // Calculate same speaker emotional shifts
  for each speaker $p \in S$ **do**
    for $j = 1$ to $|S[p]| - 1$ **do**
      if $S[p][j] \neq S[p][j+1]$ **then**
        $s \leftarrow$ similarity$(S[p][j], S[p][j+1])$
        $WES_{same} \leftarrow WES_{same} + (k \times s + b)$
      end if
    end for
  end for
  // Calculate different speaker emotional shifts
  for $j = 1$ to $N_u - 1$ **do**
    if speaker$(u_j) \neq$ speaker(u_{j+1}) **then**
      $s \leftarrow$ similarity(emotion(u_j), emotion(u_{j+1}))
      $WES_{diff} \leftarrow WES_{diff} + (k \times s + b)$
    end if
  end for
  // Compute conversation difficulty
  $DIF(c_i) \leftarrow \frac{WES_{same} + WES_{diff} + N_{sp}}{N_u + N_{sp}}$
end for
// Phase 2: Sort and partition dataset
$D_{sorted} \leftarrow$ sort$(D,$ by DIF ascending$)$
Partition D_{sorted} into n buckets: $\{D_1, D_2, \ldots, D_n\}$
where $\max(DIF(D_i)) \leq \min(DIF(D_{i+1}))$ for $i = 1, \ldots, n-1$
// Phase 3: Curriculum training
$D_{train} \leftarrow \emptyset$
for $epoch = 1$ to t **do**
  if $epoch \leq n$ **then**
    $D_{train} \leftarrow D_{train} \cup D_{epoch}$ {Add next difficulty bucket}
  end if
  $M \leftarrow$ train(M, D_{train})
end for
return M^*

Training Scheduler The training scheduler aims to divide the entire dataset D into several subsets with sim-

Dataset	Partition	Utterance	Dialogues
IEMOCAP	train + val	5810	120
	test	1623	31
MELD	train + val	11098	1152
	test	2610	280

Table 1: Statistics of the two datasets.

ilar difficulty levels (i.e., "training buckets"), denoted as $\{D_1, D_2, ..., D_n\}$. The model training starts with the easiest subset. After completing several epochs, the next more difficult subset is gradually introduced. Once all subsets have been involved in training, additional epochs are conducted on the entire dataset to further improve model performance.

Experimental Settings

This section introduces the datasets, baselines, and implementation details.

Datasets

We use two ERC datasets: IEMOCAP (Busso et al. 2008), which consists of dyadic conversations; and MELD (Poria et al. 2019), a multiparty conversation dataset derived from *Friends*. The composition of datasets is shown in Table 1.

Baselines

We compare our proposed method with two categories of mainstream baselines: (1) Conventional models, including DialogueRNN (Majumder et al. 2019), ICON (Hazarika et al. 2018), DialogueGCN (Ghosal et al. 2019), COSMIC (Ghosal et al. 2020), MMGCN (Hu et al. 2021a), DAG-ERC (Shen et al. 2021), LR-GCN (Ren et al. 2022), Multi-DAG+CL (Nguyen et al. 2024), CBERL (Meng et al. 2024), DER-GCN (Ai et al. 2024) and LSDGNN+ICL (Li, Xu, and Qiao 2025); and (2) Large language model (LLM)-based methods, including InstructERC (Lei et al. 2023) and BiosERC (Xue et al. 2024). Following the convention in the ERC field, the evaluation metric for most comparative experiments in this paper is the weighted F1 score by default.

Implementation Details

All experiments are conducted on a single NVIDIA 4090D GPU. For the IEMOCAP dataset, Qwen2.5-7B-Instruct is used as the base model; for the MELD dataset, Qwen3-8B serves as the base model. The external large language model is consistently Qwen3-14B, which is employed to generate explicit and implicit emotion interpretations as well as speaker characteristics. Additional details on training hyperparameters and settings are provided in Appendix B. All experiments in this paper—including comparative experiments—are repeated using five different random seeds, and the average results are reported.

Results and Analysis

This section presents the comparison between our model and state-of-the-art methods, the results of ablation studies, and a series of comparative experiments.

Comparison with the State of the Art

Table 2 presents the performance of our model on the IEMOCAP and MELD, with bold values indicating the best results among all models. The results of other models are taken from their original papers. Our method, PRC-Emo, achieves SOTA performance on both the IEMOCAP and MELD datasets, with weighted F1 score improvements of 0.76% on IEMOCAP and 0.61% on MELD. As a supplement, we also compare accuracy scores and achieve the best results.

Ablation Experiments

To investigate the importance of the **P**rompt, demonstration **R**etrieval, and **C**urriculum Learning modules in the PRC-Emo model, we conduct ablation experiments on two datasets. The results are shown in Table 3.

From the results, it can be seen that each module plays an important role. Among them, the design of the Prompt contributes the most significant improvement, which demonstrates that the interpretations of explicit and implicit emotions have a remarkable effect. This enables the model to deeply understand the speaker's psychological state and make better predictions. Demonstration retrieval and curriculum learning further enhance the performance. When the P, R, and C modules are removed, the model essentially becomes equivalent to directly fine-tuning LLM using LoRA.

Comparative Experiments on Prompt Design

To verify the effectiveness of our prompt design, we conduct comparative ablation experiments on different components of the prompt. Our prompt augmentation consists of the following components: **S**peaker characteristics, explicit emotion and implicit emotion **I**nterpretations, and demonstration **R**etrieval. The training process adopts curriculum learning. The results are shown in Table 4.

Table 4 shows that the full prompt design (PRC-Emo) achieves the best performance. Gradually removing demonstration retrieval (R), explicit emotion and implicit emotion interpretations (I), and speaker characteristics (S) leads to a consistent performance drop, indicating that each component contributes positively to the model. Among them, explicit emotion and implicit emotion interpretations (I) have the most significant impact.

Comparative Experiments on Curriculum Learning Methods for ERC

Nguyen et al. (2024) proposed using emotional shifts (ES) as the key metric for measuring sentence difficulty. Li, Xu, and Qiao (2025) proposed using weighted emotional shifts (WES) as the key metric. Both methods only consider emotional changes within the same speaker. Our curriculum learning approach introduces emotional changes across different speakers, providing a better assessment of sentence difficulty. Table 5 presents the performance comparison of the three curriculum learning methods. All comparisons are performed using the same base model and prompt, with ES and WES representing their respective curriculum learning strategies.

Model	IEMOCAP		MELD		Average	
	Acc.	W-F1	Acc.	W-F1	Acc.	W-F1
DialogueRNN (Majumder et al. 2019)	63.40	62.75	60.27	57.03	61.84	59.89
ICON (Hazarika et al. 2018)	64.00	63.50	—	56.30	—	59.90
DialogueGCN (Ghosal et al. 2019)	65.25	64.18	—	58.10	—	61.14
COSMIC (Ghosal et al. 2020)	—	65.28	—	65.21	—	65.25
MMGCN (Hu et al. 2021a)	—	66.22	61.34	58.65	—	62.44
DAG-ERC (Shen et al. 2021)	67.53	68.03	63.98	63.63	65.76	65.83
LR-GCN (Ren et al. 2022)	68.50	68.30	—	65.60	—	66.95
MultiDAG+CL (Nguyen et al. 2024)	69.11	69.08	—	64.00	—	66.54
CBERL (Meng et al. 2024)	69.36	69.27	67.78	66.89	68.57	68.08
DER-GCN (Ai et al. 2024)	69.70	69.40	66.8	66.10	68.25	67.75
LSDGNN+ICL (Li, Xu, and Qiao 2025)	70.35	70.24	64.67	64.07	67.51	67.16
InstructERC (Lei et al. 2023)	—	71.39	—	69.15	—	70.27
BiosERC (Xue et al. 2024)	—	71.19	—	69.83	—	70.51
PRC-Emo (Ours)	**71.03**	**71.95**	**71.50**	**70.44**	**71.27**	**71.20**

Table 2: Performance comparison of different ERC methods on IEMOCAP and MELD datasets.

Model	IEMOCAP	MELD
PRC-Emo	**71.95**	**70.44**
w/o C	71.52 (↓ 0.43)	70.07 (↓ 0.37)
w/o R + C	70.74 (↓ 1.21)	69.62 (↓ 0.82)
w/o P + R + C	68.54 (↓ 3.41)	68.72 (↓ 1.72)

Table 3: Ablation experiments.

Model	IEMOCAP	MELD
PRC-Emo	**71.95**	**70.44**
w/o R	71.49 (↓ 0.46)	70.23 (↓ 0.21)
w/o I + R	70.27 (↓ 1.68)	69.66 (↓ 0.78)
w/o S + I + R	69.90 (↓ 2.05)	69.34 (↓ 1.10)

Table 4: Prompt design comparative experiments.

Methods	IEMOCAP	MELD
base + Ours	**71.95**	**70.44**
base + ES	70.43 (↓ 1.52)	69.87(↓ 0.57)
base + WES	71.48 (↓ 0.47)	70.14(↓ 0.30)

Table 5: Comparison of curriculum learning methods.

Model	IEMOCAP	MELD
Qwen2.5-7B	**71.95**	69.73
Qwen3-8B	70.86	**70.44**

Table 6: Comparison of different LLMs.

Comparative Experiments of Different LLMs

Due to resource constraints, we only compare two of the latest high-performing open-source large language models: Qwen2.5-7B and Qwen3-8B. The experimental results are shown in Table 6. We observe an interesting phenomenon: the performance of different LLMs varies significantly across datasets. Specifically, Qwen2.5-7B significantly outperforms Qwen3-8B on the IEMOCAP dataset, while Qwen3-8B performs significantly better than Qwen2.5-7B on the MELD dataset. This may be due to Qwen3's optimization for complex real-world dialogues and possible data leakage, as MELD is based on the widely available script of the TV series *Friends*.

Conclusion

This paper proposes PRC-Emo, a novel training framework for Emotion Recognition in Conversation (ERC) that integrates **P**rompt engineering, demonstration **R**etrieval, and **C**urriculum learning using large language models (LLMs). PRC-Emo builds the first dedicated demonstration retrieval repository for ERC and designs emotion-sensitive prompt templates based on explicit and implicit emotional cues to better understand psychological states. The curriculum learning strategy organizes training from easy to hard based on weighted emotional shifts between same-speaker and different-speaker utterances. Experimental results on the IEMOCAP and MELD benchmark datasets show that PRC-Emo achieves new state-of-the-art performance. This work highlights the potential of combining prompt-based learning with curriculum learning strategies to advance ERC tasks. In future work, we plan to further explore stronger LLMs, more efficient prompting paradigms, and advanced curriculum designs to enhance emotional reasoning and improve the robustness and generalization of ERC systems.

Acknowledgments

This work is funded in part by the National Natural Science Foundation of China Project (No. 62372078).

References

Ai, W.; Shou, Y.; Meng, T.; and Li, K. 2024. DER-GCN: Dialog and Event Relation-Aware Graph Convolutional Neural Network for Multimodal Dialog Emotion Recognition. *IEEE Transactions on Neural Networks and Learning Systems*, 1–14.

an Luo; Xu, X.; Liu, Y.; Pasupat, P.; and Kazemi, M. 2024. In-context Learning with Retrieved Demonstrations for Language Models: A Survey. *ArXiv*, abs/2401.11624.

Bengio, Y.; Louradour, J.; Collobert, R.; and Weston, J. 2009. Curriculum learning. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, 41–48. New York, NY, USA: Association for Computing Machinery. ISBN 9781605585161.

Busso, C.; Bulut, M.; Lee, C.-C.; Kazemzadeh, E. A.; Provost, E. M.; Kim, S.; Chang, J. N.; Lee, S.; and Narayanan, S. S. 2008. IEMOCAP: interactive emotional dyadic motion capture database. *Language Resources and Evaluation*, 42: 335–359.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *North American Chapter of the Association for Computational Linguistics*.

Ghosal, D.; Majumder, N.; Gelbukh, A.; Mihalcea, R.; and Poria, S. 2020. COSMIC: COmmonSense knowledge for eMotion Identification in Conversations. In Cohn, T.; He, Y.; and Liu, Y., eds., *Findings of the Association for Computational Linguistics: EMNLP 2020*, 2470–2481. Online: Association for Computational Linguistics.

Ghosal, D.; Majumder, N.; Poria, S.; Chhaya, N.; and Gelbukh, A. 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)*, 154–164. Hong Kong, China: Association for Computational Linguistics.

Hazarika, D.; Poria, S.; Mihalcea, R.; Cambria, E.; and Zimmermann, R. 2018. ICON: Interactive Conversational Memory Network for Multimodal Emotion Detection. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2594–2604. Brussels, Belgium: Association for Computational Linguistics.

Hu, D.; Wei, L.; and Huai, X. 2021. DialogueCRN: Contextual Reasoning Networks for Emotion Recognition in Conversations. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 7042–7052. Online: Association for Computational Linguistics.

Hu, G.; Xin, Y.; Lyu, W.; Huang, H.; Sun, C.; Zhu, Z.; Gui, L.; and Cai, R. 2024. Recent Trends of Multimodal Affective Computing: A Survey from NLP Perspective. *arXiv preprint arXiv:2409.07388*.

Hu, J.; Liu, Y.; Zhao, J.; and Jin, Q. 2021a. MMGCN: Multimodal Fusion via Deep Graph Convolution Network for Emotion Recognition in Conversation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 5666–5675. Online: Association for Computational Linguistics.

Hu, J. E.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; and Chen, W. 2021b. LoRA: Low-Rank Adaptation of Large Language Models. *ArXiv*, abs/2106.09685.

Koga, Y.; Kando, S.; and Miyao, Y. 2024. Forecasting Implicit Emotions Elicited in Conversations. In Mahamood, S.; Minh, N. L.; and Ippolito, D., eds., *Proceedings of the 17th International Natural Language Generation Conference*, 145–152. Tokyo, Japan: Association for Computational Linguistics.

Laskar, M. T. R.; Alqahtani, S.; Bari, M. S.; Rahman, M.; Khan, M. A. M.; Khan, H.; Jahan, I.; Bhuiyan, A.; Tan, C. W.; Parvez, M. R.; Hoque, E.; Joty, S. R.; and Huang, J. X. 2024. A Systematic Survey and Critical Review on Evaluating Large Language Models: Challenges, Limitations, and Recommendations. In *Conference on Empirical Methods in Natural Language Processing*.

Lee, M.-C.; Chiang, S.-Y.; Yeh, S.-C.; and Wen, T.-F. 2020. Study on emotion recognition and companion Chatbot using deep neural network. *Multimedia Tools and Applications*, 79: 19629 – 19657.

Lei, S.; Dong, G.; Wang, X.; Wang, K.; and Wang, S. 2023. InstructERC: Reforming Emotion Recognition in Conversation with a Retrieval Multi-Task LLMs Framework. *arXiv preprint arXiv:2309.11911*.

Li, X.; Xu, X.; and Qiao, J. 2025. Long-Short Distance Graph Neural Networks and Improved Curriculum Learning for Emotion Recognition in Conversation. In *Proceedings of the 28th European Conference on Artificial Intelligence (ECAI 2025)*, volume 413 of *Frontiers in Artificial Intelligence and Applications*, 4033–4040. IOS Press.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *ArXiv*, abs/1907.11692.

Majumder, N.; Poria, S.; Hazarika, D.; Mihalcea, R.; Gelbukh, A.; and Cambria, E. 2019. DialogueRNN: An Attentive RNN for Emotion Detection in Conversations. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01): 6818–6825.

Meng, T.; Shou, Y.; Ai, W.; Yin, N.; and Li, K. 2024. Deep Imbalanced Learning for Multimodal Emotion Recognition in Conversations. *IEEE Transactions on Artificial Intelligence*, 1–15.

Nguyen, C.-V. T.; Nguyen, C.-B.; Le, D.-T.; and Ha, Q.-T. 2024. Curriculum Learning Meets Directed Acyclic Graph for Multimodal Emotion Recognition. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, 4259–4265. Torino, Italia: ELRA and ICCL.

Poria, S.; Hazarika, D.; Majumder, N.; Naik, G.; Cambria, E.; and Mihalcea, R. 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*, 527–536. Florence, Italy: Association for Computational Linguistics.

Qin, X.; Wu, Z.; Zhang, T.; Li, Y.; Luan, J.; Wang, B.; Wang, L.; and Cui, J. 2023. BERT-ERC: fine-tuning BERT is enough for emotion recognition in conversation. In *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence and Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence and Thirteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'23/IAAI'23/EAAI'23. AAAI Press. ISBN 978-1-57735-880-0.

Reimers, N.; and Gurevych, I. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Conference on Empirical Methods in Natural Language Processing*.

Ren, M.; Huang, X.; Li, W.; Song, D.; and Nie, W. 2022. LR-GCN: Latent Relation-Aware Graph Convolutional Network for Conversational Emotion Recognition. *IEEE Transactions on Multimedia*, 24: 4422–4432.

Shen, W.; Wu, S.; Yang, Y.; and Quan, X. 2021. Directed Acyclic Graph Network for Conversational Emotion Recognition. In Zong, C.; Xia, F.; Li, W.; and Navigli, R., eds., *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 1551–1560. Online: Association for Computational Linguistics.

Xue, J.; Nguyen, P. M.; Matheny, B.; and Nguyen, L. M. 2024. BiosERC: Integrating Biography Speakers Supported by LLMs for ERC Tasks. In *International Conference on Artificial Neural Networks*.

Yang, L.; Shen, Y.; Mao, Y.; and Cai, L. 2022. Hybrid Curriculum Learning for Emotion Recognition in Conversation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(10): 11595–11603.