# InstructERC: Reforming Emotion Recognition in Conversation with a Multi-task Retrieval-based LLMs Framework

Anonymous ACL submission

## Abstract

The field of emotion recognition of conversation (ERC) has been focusing on separating sentence feature encoding and context modeling, lacking exploration in generative paradigms based on unified designs. In this study, we propose a novel approach, **InstructERC**, to reformulate the ERC task from a discriminative framework to a generative framework based on Large Language Models (LLMs). InstructERC makes three significant contributions: (1) it introduces a simple yet effective retrieval template module, which helps the model explicitly integrate multi-granularity dialogue supervision information. (2) We introduce two additional emotion alignment tasks, namely speaker identification and emotion prediction tasks, to implicitly model the dialogue role relationships and future emotional tendencies in conversations. (3) Pioneeringly, we unify emotion labels across benchmarks through the feeling wheel to fit real application scenarios. InstructERC still perform impressively on this unified dataset. Our LLM-based plugin framework significantly outperforms all previous models and achieves comprehensive SOTA on three commonly used ERC datasets. Extensive analysis of parameter-efficient and data-scaling experiments provides empirical guidance for applying it in practical scenarios. Our code and aligned unified dataset are in the supplementary.

## 1 Introduction

“The question is not whether intelligent machines can have emotions, but whether machines without emotions can achieve intelligence”, as mentioned in “Society of Mind” (Minsky, 1988). Empowering machines with the ability to understand emotions in various scenarios has always been the unwavering direction of researchers.

In contrast to conventional binary sentiment analysis tasks (Pontiki et al., 2016) , which only rely on text with explicit attitude tendencies, the emotion recognition in conversation (ERC) task aims to identify more fine-grained emotional tendencies in each sentence of a conversation. Specifically, for a given complete dialogue sequence input and a set of emotional labels, the model is required to accurately assign an emotional label to each sentence. Intuitively, the recognition of emotional tendencies in the target sentence is heavily influenced by its historical utterances (Yingjian et al., 2023), and there is significant variation in how different speakers perceive and express emotions (Shen et al., 2021). Therefore, it is imperative to meticulously model the speakers and dialogue context.

Figure 1 illustrates that previous work based on Roberta (Liu et al., 2019) in ERC can be roughly divided into three categories: (1) **Transformer-based methods** (Li et al., 2020; Song et al., 2022; Liu et al., 2023; Chudasama et al., 2022) attempt to establish long-range emotional correlations in conversational scenarios by directly adopting or modifying the original transformer block. (2) **Recurrent-based methods** (Hu et al., 2023; Lei et al., 2023; Majumder et al., 2019; Hazarika et al., 2018; Poria et al., 2017) utilize various forms of RNNs, like LSTM and GRU, to model individual emotional states and global emotional impacts separately. (3) **GNN-based methods** (Ghosal et al., 2019; Ishiwatari et al., 2020; Shen et al., 2021; Li et al., 2023a) typically use nodes and edges to model characters and dialogue relationships in conversations. Above approaches have their strengths in modeling dialogue at the sentence level, but they still generally adhere to the paradigm of fine-tuning sentence features and separately modeling dialogue context. However, in realistic scenarios, end-to-end model designs are often more practical. [1].

Fortunately, the recent successful application (OpenAI, 2023) and emergence capabilities (Zhao et al., 2023) of pre-trained large language mod-

[1] The discussion between discriminant model and InstructERC can refer to D.2

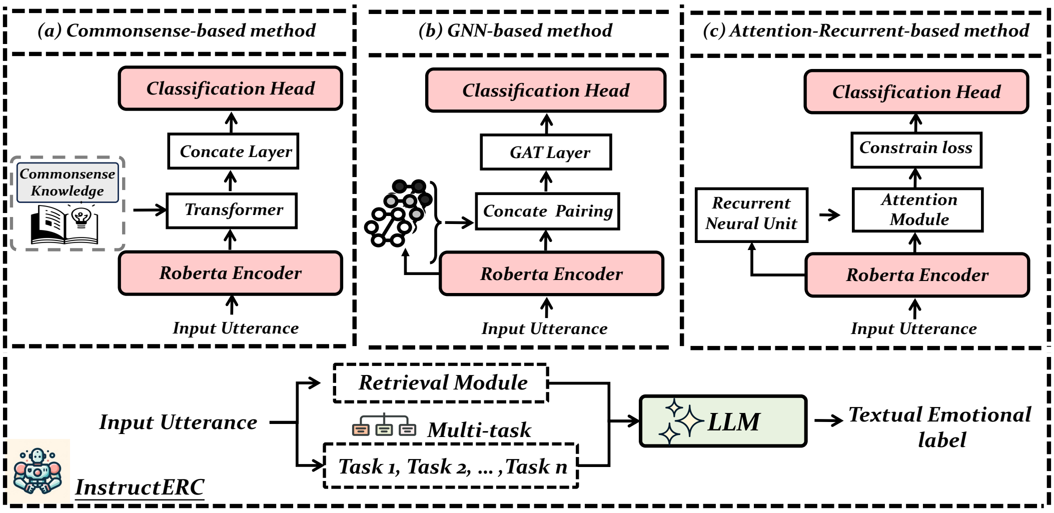


Figure 1: The illustration of different paradigms for ERC

els (LLMs) have demonstrated remarkable performance in natural language reasoning tasks. By using a generative architecture, LLMs unify the output and input of different tasks and have shown significant performance improvements in all NLP tasks. Despite their powerful capabilities, enabling these abilities for specific sub-tasks requires high-quality prompts (Wei et al., 2021; Chung et al., 2022) and designs to fill the reasoning gap. Therefore, how to use LLMs framework to reconstruct ERC while considering context modeling, speaker modeling, and capturing conversation relationships poses a significant challenge in pushing this framework towards a realistic ERC application.

In this work, we reformulate the ERC task using LLMs. Specifically, we design a simple but efficient retrieval template module, which consists of instruction, historical utterance, label statements, and demonstration retrieval to explicitly integrate multi-granularity dialogue supervision information during reasoning. In addition, we separately design two auxiliary tasks for the ERC task: speaker identification task and emotion prediction task. The speaker identification task assists LLMs in modeling dialogue role relationships by predicting the speaker of each sentence, while the emotion prediction task models future emotional tendencies in conversations. Furthermore, due to biases in data distribution and labeling across different ERC domains, it's still challenging for discriminative ERC models to achieve multi-domain ERC capabilities, both in terms of engineering and performance. To dive deeper into this topic, we pioneeringly align labels for three benchmarks and conduct a series of unified dataset experiments. Looking ahead, we contend that IERC, as the first framework transitioning from single-domain to multi-domain ERC, offers us a glimpse into the prospective landscape of open-domain emotional artificial intelligence (Emotional AGI).

In conclusion, our work can be outlined as follows:

- To the best of our knowledge, we are the first to reformulate the ERC task as a retrieval based Seq2Seq paradigm with LLMs and present an effective instruction template which can adapt to different dialog scenarios.

- We propose two novel emotional auxiliary tasks to implicitly model the dialogue role relationships and future emotional tendencies in conversations.

- Our InstructERC significantly outperforms all previous models and achieves comprehensive SOTA on three commonly used ERC datasets.

- To advance towards multi-domain ERC scenario, we pioneeringly align labels for three benchmark to form the UIME ERC dataset, a series of unified dataset experimental results provides empirical guidance for application in practical scenarios.

## 2 Methodology

In this section, we present a comprehensive overview of the proposed InstructERC framework shown as Figure 3. Firstly, we provide a brief introduction to the task definition of ERC. Next, we discuss the framework of InstructERC, which consists of two major parts: retrieval template module and emotional alignment tasks. Finally, we introduce training and inference process of our framework. [2]

## 2.1 Problem Definition[3]

Assuming a dialogue text $U = [u_1, u_2, ...u_n]$ of length $n$ is given, which includes $M$ speakers/parties $p_1, p_2, ..., p_M$ $(M \geq 2)$ in the dialogue, and each utterance $u_i$ spoken by the corresponding speaker $p_{K(u_i)}$. Function $K$ is employed to establish a mapping between each utterance and its corresponding speaker. $o$ is the number of emotional categories, which varies with the number of emotional types in different evaluation datasets.

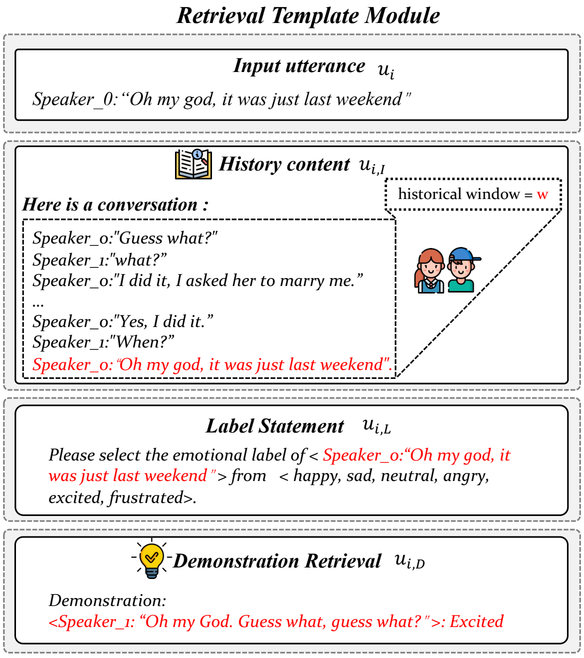
## 2.2 Retrieval Template Module

To better transfer and utilize the inference ability of pre-trained large language models, we reconstruct the ERC task to the seq2seq form and solve it through fine-tuning LLMs. Therefore, we construct a efficient retrieval template module to bridge the gap when applying LLMs to specific NLP subtasks. As shown in Figure 2, for ERC task, each input consists of four parts: instructions, historical content, label statement, and demonstration retrieval.

**Instruction.** The instructions serve to provide the model with a well-defined role, precise details of the ERC task, and a standardized format for the input dialogue text. For the primary ERC task, our instruction $u_{i,I}$ is shown in Figure 2.

**Historical Content.** To model the context in realistic ERC scenarios, We employ a hyperparameter, the historical window (denoted as $w$), to indicate the specific rounds (including current utterance) of historical dialogue along with the corresponding speaker information. For the emotion recognition of the target utterance $u_n$, its historical content $u_{i,H}$ is shown in Figure 2.

**Label Statement.** To confine the model's output within a finite range of labels and enable the model to focus on the current utterance being recognized, our label statement $u_{i,L}$ is shown in Figure 2.


Retrieval Template Module
Input utterance $u_i$
History content $u_{i,I}$
historical window = w
Label Statement $u_{i,L}$
Demonstration Retrieval $u_{i,D}$

Speaker_0:"Oh my god, it was just last weekend"

Here is a conversation :

Speaker_o:"Guess what?"
Speaker_1:"what?"
Speaker_o:"I did it, I asked her to marry me."
...
Speaker_o:"Yes, I did it."
Speaker_1:"When?"
Speaker_o:"Oh my god, it was just last weekend".

Please select the emotional label of < Speaker_o:"Oh my god, it was just last weekend"> from < happy, sad, neutral, angry, excited, frustrated>.

Demonstration:
<Speaker_1: "Oh my God. Guess what, guess what?">: Excited

Figure 2: The Schematic of Retrieval Template Module.

**Demonstration Retrieval.** In order to further integrate emotional information to assist reasoning, we have developed a domain demonstration recall module based on semantic similarity. In detail, we construct a domain base $\mathcal{D}_{domain}$ from the training dataset that removes speaker identity information and balances the number of emotion labels, which ensures that the demonstrations is not influenced by the distribution of speakers or emotion labels in the dataset. For a given utterance $u_i$ to be identified, we retrieve the most relevant ERC example from $\mathcal{D}_{domain}$ as the demonstration. To perform the retrieval, we use a bidirectional encoder SBERT (Reimers and Gurevych, 2019) to find the most semantically similar ERC example $d_{rvl}$. SBERT generates independent CLS embeddings for the target utterance $u_i$ and each element $d_j$ in $\mathcal{D}_{domain}$. After sorting all target-demonstration pairs by cosine similarity, we select the pair with the highest score as the most relevant element $d_{rvl}$. An abstract mathematical description of this process is as follows:

$$d_{rvl_i} = \underset{d_j \in \mathcal{D}_{domain}}{\text{argmax}} \ \text{SBERT}(u_i, d_j) \quad (1)$$

The textual input $u_{i,D}$ for the demonstration retrieval part is shown in Figure 2. In summary, after constructing the Retrieval template, the simplified input $x_i$ for the main task is as follows:

$$x_i = [u_{i,I}; u_{i,H}; u_{i,L}; u_{i,D}] \quad (2)$$

where [;] means the textual concatenation, $u_{i,I}$, $u_{i,H}$, $u_{i,L}$, and $u_{i,D}$ indicate Instructions, Historical content, Label statement, demonstration retrieval for a given utterance $u_i$.

[2] Due to the space limitation, we have included the related works in Appendix B.

[3] The difference of problem definition between two paradigms can be refer to Appendix D.2.

### 2.3 Emotional alignment tasks

To better capture the dialogue role relationships and future emotional tendencies in conversations, we have incorporated two auxiliary tasks, namely speaker identification and emotion impact prediction, which constitute the fine-grained subtasks of the InstructERC framework. The model is jointly trained with these auxiliary tasks to improve its overall performance, illustrated in Figure 3.

**Speaker Identification task.** Emotions are expressed differently among different speakers. Previous models have used techniques such as speaker-based masked attention modules or multiple GRUs to capture the emotional expression features of different characters. This modeling of emotional expression in the task can also be transformed into a generative task using our InstructERC. To enable the LLM to capture the speaking styles of different individuals, beyond (Li et al., 2020), the model is trained to identify the relevant speaker for a given utterance, without considering the historical context. For a given dataset, a predefined set of speaker labels is provided. Consistent with the main task, the Instruction text input $x_i^p$ for this task is constructed as follows:

> *"Now you are an expert of sentiment and emotional analysis. Please select the Speaker label of the utterance <Speaker:$u_i$> from <$p_1$,...,$p_M$>"*

The loss function for the Speaker Identification is as follows:

$$\mathcal{L}_p = \sum_i^N - \log P(\mu_i | x_i^p, \theta_p) \quad (3)$$

Here, $\mu_i$ represents the token of the corresponding speaker label for the given speaker identification task input sample $x_i^p$. Unless otherwise specified, $N$ stands for the total number of utterances in the dataset, while $\theta_*$ represents the parameters of the LLM in different periods.

**Emotion Impact Prediction task.** In the daily conversations, the intricate relationships between individuals can have a significant impact on the emotional states of subsequent dialog. Prior research has attempted to address this issue by constructing a dialogue relationship graph and utilizing a complex graph neural network to model the emotional impacts of these relationships. However, these methods are often associated with a highly intricate data preprocessing pipeline and are susceptible to overfitting on certain datasets. To address these issues, we propose a generative framework for the emotion impact prediction task, which implicitly captures the interplay between dialogues and emotional impacts.

Specifically, the input for emotion impact prediction consists of three parts: instruction, historical content, and label statement. First, the instruction part of this task is kept consistent with the instruction part of the main task. Then, since the task requires predicting the impact of previous historical utterances on the current utterance, unlike the main task, the historical content $u_{i,H}^e$ with a window of "w" will not include the current utterance. Correspondingly, to stay aligned with the original design intention of the task, the label statement of this task is modified as follows:

> *"Based on the above historical utterances, the next utterance is spoken by <$P_{K(u_i)}$ >, please predict the emotion states of <$P_{K(u_i)}$ >from <$e_1$, $e_2$, ..., $e_o$ >:"*

Hence, the overall input for emotion impact prediction is:

$$x_i^e = [u_{i,I}; u_{i,H}^e, u_{i,L}^e] \quad (4)$$

The loss calculation for the emotion impact prediction task is as follows:

$$\mathcal{L}_e = \sum_i^N - \log P(\epsilon_i | x_i^e, \theta_e) \quad (5)$$

Here, $\epsilon_i$ represents the emotional label token of the text label $e_i$ corresponding to the formatted input utterance $x_i$.

### 2.4 Overview of InstructERC

To sum up the instruction based generative framework for ERC, given an input utterance $x_i$ after concatenating the retrieval template $d_{rvl}$ and a LLM, the model returns the logits $g_i$ and the generated text $y_i$ for the entire sentence, including both input and output tokens. This is represented by the following equation:

$$y_i, \mathbf{g_i} = \mathrm{LLM}(x_i, \theta_{all}) \quad (6)$$

Here, $\theta$ is the same as mentioned. The LLM predicts the conditional probability $p(\gamma_i|x_i, \theta)$ of generating each token $\gamma_i$ of the generated text $y_i$ until the end symbol <eos>is outputted. As for

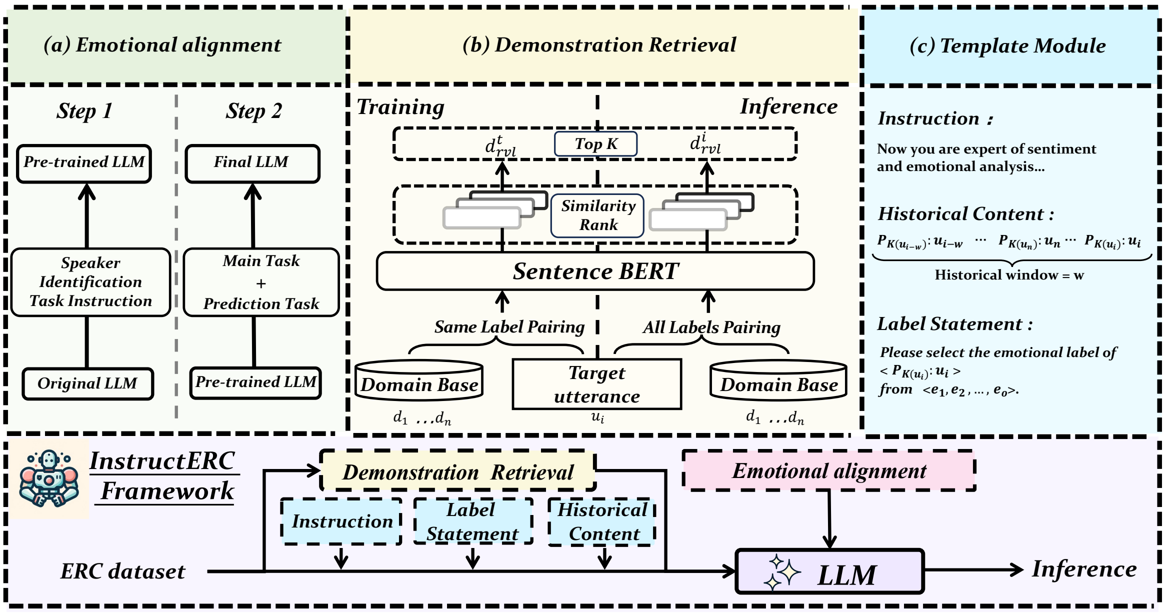


Figure 3: The overview of InstructERC framework

logits $\mathbf{g_i} \in \mathbf{R}^{L \times V}$, where $L$ and $V$ denote the length of the entire sentence and the size of the vocabulary used by the LLM, respectively.

In accordance with the original training method of LLMs, we adopt the next token prediction loss to measure the model's output error. Therefore, the loss calculation of the main task, denoted as $\mathcal{L}_{main}$, is defined as follows:

$$\mathcal{L}_{main} = \sum_{i}^{N} -\log P(\epsilon_i | x_i, \theta_{all}) \tag{7}$$

**Training and Inference.**

During training and inference, our retrieval process, emotional alignment tasks and main tasks in InstructERC can be divided into two stages:

In the first stage of joint training, the characteristics of the speaker intuitively form the basis of emotional expression. Therefore, we use the speaker identification task for LLM pre-training to fine-tune speaker characteristics, which aims to preheat parameters for subsequent ERC tasks.

In the second stage, we fine-tune LLM using both the ERC main task and the emotion influence prediction task to improve overall performance. The training loss at this stage is $\mathcal{L}_{main} + \alpha * \mathcal{L}_e$, where $\alpha$ is a hyperparameter used to adjust the weight of the emotion influence prediction task loss in the second overall joint training loss.

The difference of demonstration retrieval on training and inference stage is shown in figure 3, we limit the retrieved examples to those with the same emotion label as the current recognized speech, namely same label pairing ,in order to provide more diverse emotional understanding while avoiding excessive noise during training. During inference, there are no restrictions on the retrieved demonstrations due to the labels are unknown, namely All labels pairing. The retrieval results, simply referred as $d_{rvl}$, are specialized as $d^t_{rvl}$ and $d^i_{rvl}$ in training and inference stage, respectively.

## 3 Experiments and Results

### 3.1 Dataset

We evaluate the efficacy of InstructERC on three standard benchmark datasets: IEMOCAP, MELD, and EmoryNLP. The specifics of the datasets are outlined in Table 6. The details of dataset can be refer to Appendix C.1.

### 3.2 Baselines

Align with the related works, we select several only textual modality baselines to compare with our InstructERC. **1) Transformer-based**: SPCL+CL(Song et al., 2022) and MPLP (Zhang et al., 2023b) , **2) Recurrent-based**: EmotionIC(Yingjian et al., 2023) and SACL-LSTM(Hu et al., 2023), **3) GNN-based**: DualGATs(Zhang et al., 2023a) and Skier(Li et al., 2023b). **4) LLM backbones**: ChatGLM-6B & ChatGLM2-6B (Du et al., 2022) and LLaMA-7B & LLaMA2-7B (Touvron et al., 2023). More details of baselines and implementations can be refered to Appendix C.2 and D.1.

Table 1: The main results on three benchmarks.

| Dataset<br>Models | IEMOCAP<br>W-F1 | MELD<br>W-F1 | EmoryNLP<br>W-F1 | Average<br>W-F1 |
|---|---|---|---|---|
| Disciminant Models | | | | |
| SPCL+CL$^\dagger$ | 69.74 | 66.35 | 40.25 | ***58.78*** |
| MPLP$^*$ | 66.65 | 66.51 | - | - |
| EmotionIC$^\dagger$ | ***69.61*** | 66.40 | 40.01 | 58.63 |
| SACL$^*$ | 69.22 | 66.45 | 39.65 | 58.44 |
| DualGATs$^*$ | 67.68 | 66.90 | ***40.29*** | 58.29 |
| Skier$^\dagger$ | - | ***67.39*** | 40.07 | - |
| Zero-shot + InstructERC | | | | |
| ChatGPT3.5$^\dagger$ | ***53.38*** | ***65.07*** | ***37.00*** | ***51.81*** |
| ChatGLM$^\dagger$ | 38.6 | 38.8 | 19.6 | 32.33 |
| ChatGLM2$^\dagger$ | 21.1 | 21.8 | ***24.4*** | 22.43 |
| Llama$^\dagger$ | 0.753 | 9.12 | 5.31 | 5.06 |
| Llama2$^\dagger$ | 2.774 | 16.28 | 8.36 | 9.46 |
| LoRA + Backbone | | | | |
| ChatGLM$^\dagger$ | 17.98 | 40.54 | 25.71 | 28.07 |
| ChatGLM2$^\dagger$ | 52.88 | 64.85 | 37.69 | 51.80 |
| Llama$^\dagger$ | 55.81 | ***66.15*** | 37.98 | 53.21 |
| Llama2$^\dagger$ | ***55.96*** | 65.84 | ***38.21*** | ***53.33*** |
| LoRA + InstructERC | | | | |
| ChatGLM$^\dagger$ | 36.04 | 46.41 | 30.86 | 37.77 |
| ChatGLM2$^\dagger$ | 67.54 | 65.58 | 39.09 | 57.40 |
| Llama$^\dagger$ | 64.17 | 67.62 | 39.34 | 57.04 |
| Llama2$^\dagger$ | **71.39** | **69.15** | **41.37** | **60.64** |

NOTE: The best results of other baselines are in gold font, while SOTA results across all models are emphasized in red font. * indicate results sourced from the model's paper, and a (†) denotes results from reproductions conducted by the authors.

### 3.3 Main Results

Table 1 illustrates the results of comparing our InstructERC model with other models and backbones from different perspectives. Based on this, We make the following observations:

(1) Our methods achieves significant improvements over the SOTA of discriminative models on all benchmarks. Specifically, we outperform EmotionIC, Skier, and DuaGATs by 1.73%, 1.76%, and 1.08% on IEMOCAP, MELD and EmoryNLP respectively. Notably, we completely outperformed commonsense knowledge models (Skier) on two benchmarks without any external knowledge, demonstrating the extreme utilization of our method for textual data.

(2) To gain an insight into LLM models under different supervision scenarios for ERC task, we conduct experiments on Zero-shot + InstructERC and LoRA + InstructERC settings. It can be observed that even with carefully designed primary task instructions, LLMs still struggle in zero-shot scenarios, which further confirms the existence of a significant reasoning gap in their application to ERC sub-task. Furthermore, by utilizing the LoRA + InstructERC, the performance of the four LLMs has significantly improved, especially on the IEMOCAP dataset. This fully demonstrates the effectiveness and generalization ability of our InstructERC framework, which greatly enhances the emotion recognition capability of LLM in long texts.

(3) InstructionERC is a plug-and-play method that can be adapted to multiple generative frameworks, such as prefix decoder or causal decoder. Although ChatGPT has a relevant competitive good performance on short length conversation scenrios(e.g. Meld,EmoryNLP), as can be seen, our results are far superior to the level of ChatGPT. Our unified alignment task and demonstration construction strategy are not tailored to any specific dataset design, highlighting the strong transferability and generalization capability of our approach.

### 3.4 Ablution study

We conduct an ablation study to investigate the characteristics of the main components in InstructERC. Table 2 shows the ablation results, and “w/o" denotes the model performance without a specific module. We have following observations:

(1) The performance of InstructERC drops when removing any one component, which suggests that every part of the design is necessary.

(2) Removing any one Emotional alignment task results in great performance degradation. This is consistent with our conjecture since speaker identification and emotion impact prediction provide relatively orthogonal semantic information from two perspectives. [4]

(3) Taking away the domain retrieval module resulted in a steady decline on all three datasets, demonstrating the important role of domain information in dialogue modeling.

4) Removing joint alignment task tasks causes obvious performance degradation compared with removing one of them, which indicates that jointly pre-training objectives have a mutually reinforcing effect. [5]

(5) Replacing LoRA with full-parameter finetuning results in a significant drop in performance, which indicates that the parameter-efficient approach is effective in preventing overfitting of LLMs on the ERC task. For detailed analysis,

[4] We also explore the impact of $\alpha$ on the performance of InstructERC, refer to Appendix E.2

[5] We also explore the optimal conversational turns in modeling context in ERC, please refer to the“The historical window exploration study” section in Appendix E.1.

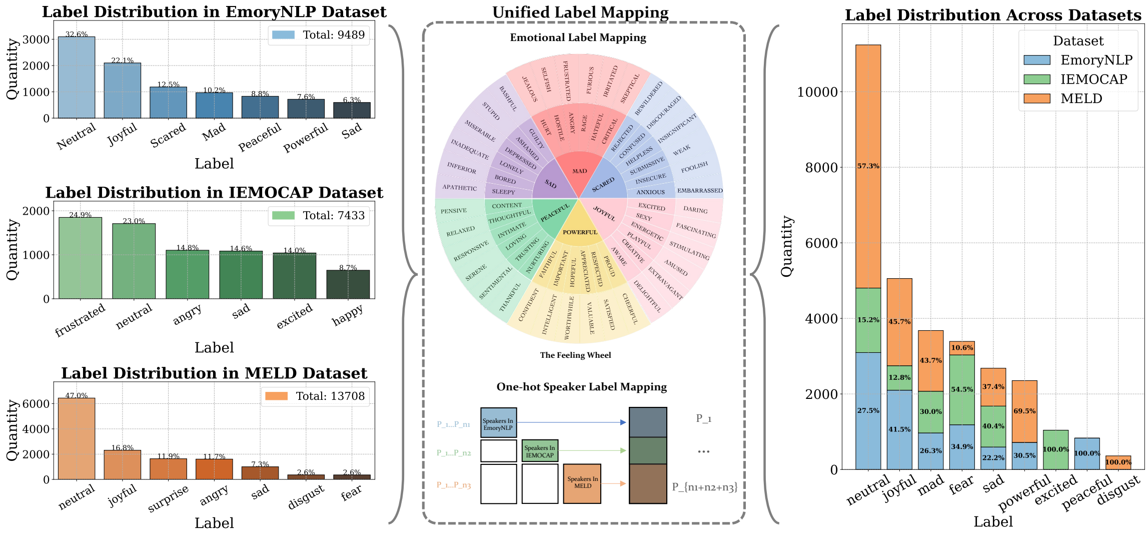


Figure 4: Unified Label Mapping Across three Open-source Benchmarks. The Feeling Wheel is proposed by (Willcox, 1982)

Table 2: The ablation results of Llama2 on three benchmarks.

| Dataset<br>Models | IEMOCAP<br>W-F1 | MELD<br>W-F1 | EmoryNLP<br>W-F1 |
|---|---|---|---|
| LoRA + InstructERC | | | |
| Llama2 | **71.39**$_{\pm 0.10}$ | **69.15**$_{\pm 0.08}$ | **41.37**$_{\pm 0.11}$ |
| w/o $\mathcal{L}_e$ | $70.50^{**}_{\pm 0.12}$ | $68.97^{*}_{\pm 0.10}$ | $40.78^{**}_{\pm 0.10}$ |
| w/o $\mathcal{L}_p$ | $70.70^{*}_{\pm 0.15}$ | $68.76^{*}_{\pm 0.14}$ | $40.59^{**}_{\pm 0.13}$ |
| w/o $\mathcal{L}_e + \mathcal{L}_p$ | $69.71^{**}_{\pm 0.17}$ | $68.39^{**}_{\pm 0.11}$ | $39.56^{**}_{\pm 0.15}$ |
| w/o $\mathcal{D}_{domain}$ | $70.91^{*}_{\pm 0.13}$ | $68.62^{*}_{\pm 0.19}$ | $40.54^{*}_{\pm 0.19}$ |
| w/o $_{LoRA}$ | $70.30^{**}_{\pm 0.11}$ | $64.80^{**}_{\pm 0.12}$ | $40.05^{**}_{\pm 0.21}$ |

Results with standard deviation and significance testing between w/o* and LLama2 (*p<0.05, **p<0.01.)

please refer to the "All Parameters vs Parameter Efficiency" section in Appendix E.4 . The further data scaling analysis of single dataset can be refer to Appendix E.5.

## 4 Unified dataset Experiments

In real-world scenarios, the ideal ERC model should be able to address ERC challenges across multiple domains, and even carry out open-domain ERC tasks. However, biases in data distribution and labeling make it challenging for small ERC models to achieve multi-domain capabilities, To better simulate real-world scenarios, we first reconstruct three ERC datasets into a single ERC dataset (UIME) with unified labels based on the Emotion Wheel (Figure 4), to better suit more industrial scenarios.

### 4.1 Unified Dataset Experiment Setup

Within the settings of this experiment, all emotional labels across the datasets are standardized, and all speaker labels are also consolidated. The unification details of speaker labels and emotional labels can be refered to Appendix A. Subsequently, we conduct data scaling experiments on the UIME. To explore the impact of different sampling methods on the final performance, two data scaling approaches are experimented with: Total Mixing and Ratio Mixing.

In the "Total Mixing" approach, all subdatasets in UIME are first merged together, and then {1, 1/2, 1/4, 1/8, 1/16, 1/32, 1/64} amounts of data are randomly sampled separately from the merged data to fine-tune instructERC. Conversely, in the "Ratio Mixing" approach, {1, 1/2, 1/4, 1/8, 1/16, 1/32, 1/64} amounts of subdatasets are first randomly sampled separately, and then they are merged in accordance with their respective ratios to form the training data. Both approaches maintain the same quantity of the final training data.

The details of results are shown in Table 5 in Appendix A, and a more intuitive presentation is shown in Figure 5.

### 4.2 The Robustness of InstructERC

As depicted in the Figure 5, Compared to the single dataset training setup, the performance of InstructERC, when fine-tuned on the UIME, has experienced a minor drop across three benchmarks. Specifically, there's a decrease of 2.4% in IEMO-

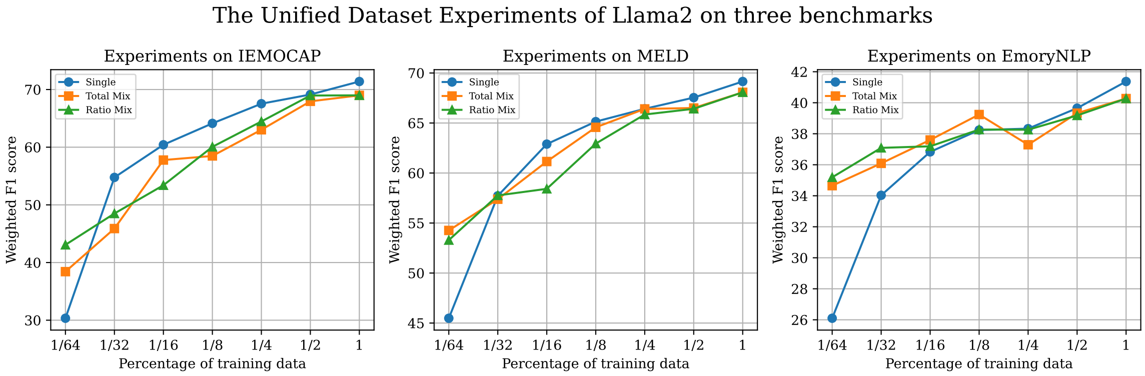


Figure 5: The data scaling analysis demonstrated on three benchmarks using different data mixing strategies

CAP, 1.08% in MELD, and 1.1% in EmoryNLP. However, a relatively high Weighted F1 score (W-F1) can still be maintained simultaneously on these three benchmarks, particularly the performance of MELD(68.07%), which continues to surpass the SOTA level of all small models. The results exhibits InstructERC 's exceptional robustness, which is capable of concurrently acquiring emotional paradigms from a multitude of distinct distributions [6].

### 4.3 The Data Scaling Exploration

The data scaling experiments are conducted on the unified dataset from 1 to 1/64. As the scale of trainig data exponentially decreases from 1 to 1/32 within the range, the performance of the model on the three benchmarks exhibits a slight fluctuation in linear decline.

We are also surprised to discover that during the final stage of training data reduction from 1/32 to 1/64, the Total Mixing and Ratio Mixing strategies continue to exhibit a linear performance decline. However, the performance of the model trained under the single method experiences a drastic drop, as depicted in Figure 5. We posit that data from different scenarios endows the model with the capability to comprehend emotions from diverse perspectives. This, in turn, allows the model to achieve robust enhancements under various data conditions. Such mutual gain is particularly pronounced in low resource scenarios (1/64). This is consistent with the findings of some existing explorations in large models (Dong et al., 2023).

### 4.4 The Discussion of Mixing Strategies

We have further investigated the impact of different mixing strategies on data scaling. The results displayed by different datasets on various mixing strategies can be interpreted from the following two perspectives:

**Data Representativeness:** In Total Mixing sampling, where each dataset's samples are equally likely to be selected, the unique traits of smaller datasets like IEMOCAP may be obscured by larger ones like MELD. In contrast, Ratio Mixinging sampling, which represents each dataset proportionally to its original sample size, may better highlight the characteristics and influence of smaller datasets.

**Effect of Class Imbalance:** In smaller datasets with internal class imbalances, Total Mixing sampling could exacerbate these imbalances. For instance, if IEMOCAP has a relatively smaller number of samples in a certain category, Total Mixing sampling might further intensify this imbalance during model training. Ratio Mixing sampling, however, better preserves the original class proportions of the datasets, potentially mitigating class imbalance impacts to a degree.

## 5 Conclusion

We introduce InstructERC, a novel approach that transforms the ERC task from a discriminative framework to a generative framework using LLMs. InstructERC presents a simple and effective retrieval template adapting to different conversation lengths. Futhermore, we introduce two emotional alignment tasks to model speaker and complex conversation relationships. InstructERC outperforms all previous models and achieve comprehensive SOTA results on three benchmarks. We also pioneer in unifying label mapping and modeling across these datasets, demonstrating the InstructERC's robust generalization capabilities. Our extensive analysis provides practical insights for implementing InstructERC in real-world ERC scenarios.

[6] The statistics of scaling analysis can be found in Table 5

## Limitation

In this work, we focus solely on the textual aspects of these datasets. The exploration of multimodal aspects is reserved for future research. We have conducted our explorations specifically on two representative large model frameworks, ChatGLM and LLaMA. Due to limitations in our graphics card capacity, the maximum parameter size of the large models we used does not exceed 7 billion.

## Ethics Statement

All the data sets we used for the experiment were published publicly. These data sets passed the ethical review at the time of publication. All the non-original methods and modules mentioned in this article have quoted other people's literature. All our science artifacts observe MIT licese.

## References

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners.

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

Vishal Chudasama, Purbayan Kar, Ashish Gudmalwar, Nirmesh Shah, Pankaj Wasnik, and Naoyuki Onoe. 2022. M2fnet: Multi-modal fusion network for emotion recognition in conversation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4652–4661.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

Guanting Dong, Hongyi Yuan, Keming Lu, Chengpeng Li, Mingfeng Xue, Dayiheng Liu, Wei Wang, Zheng Yuan, Chang Zhou, and Jingren Zhou. 2023. How abilities in large language models are affected by supervised fine-tuning data composition. *arXiv preprint arXiv:2310.05492*.

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

Deepanway Ghosal, Navonil Majumder, Soujanya Poria, Niyati Chhaya, and Alexander 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)*, pages 154–164.

Devamanyu Hazarika, Soujanya Poria, Rada Mihalcea, Erik Cambria, and Roger Zimmermann. 2018. Icon: Interactive conversational memory network for multimodal emotion detection. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pages 2594–2604.

Dou Hu, Yinan Bao, Lingwei Wei, Wei Zhou, and Songlin Hu. 2023. Supervised adversarial contrastive learning for emotion recognition in conversations. *arXiv preprint arXiv:2306.01505*.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models.

Taichi Ishiwatari, Yuki Yasuda, Taro Miyazaki, and Jun Goto. 2020. Relation-aware graph attention networks with relational position encodings for emotion recognition in conversations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7360–7370.

Shanglin Lei, Xiaoping Wang, Guanting Dong, Jiang Li, and Yingjian Liu. 2023. Watch the speakers: A hybrid continuous attribution network for emotion recognition in conversation with emotion disentanglement. *arXiv preprint arXiv:2309.09799*.

Jiang Li, Xiaoping Wang, Guoqing Lv, and Zhigang Zeng. 2023a. Graphcfc: A directed graph based cross-modal feature complementation approach for multimodal conversational emotion recognition. *IEEE Transactions on Multimedia*.

Jiangnan Li, Zheng Lin, Peng Fu, and Weiping Wang. 2021. Past, present, and future: Conversational emotion recognition through structural modeling of psychological knowledge. In *Findings of the association for computational linguistics: EMNLP 2021*, pages 1204–1214.

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

Wei Li, Luyao Zhu, Rui Mao, and Erik Cambria. 2023b. Skier: A symbolic knowledge integrated model for conversational emotion recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Junyang Lin, Rui Men, An Yang, Chang Zhou, Yichang Zhang, Peng Wang, Jingren Zhou, Jie Tang, and Hongxia Yang. 2021. M6: Multi-modality-to-multi-modality multitask mega-transformer for unified pre-training. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery Data Mining*, KDD '21, page 3251–3261, New York, NY, USA. Association for Computing Machinery.

Xiao Liu, Jian Zhang, Heng Zhang, Fuzhao Xue, and Yang You. 2023. Hierarchical dialogue understanding with special tokens and turn-level attention. *arXiv preprint arXiv:2305.00262*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Yaojie Lu, Qing Liu, Dai Dai, Xinyan Xiao, Hongyu Lin, Xianpei Han, Le Sun, and Hua Wu. 2022. Unified structure generation for universal information extraction. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5755–5772, Dublin, Ireland. Association for Computational Linguistics.

Navonil Majumder, Soujanya Poria, Devamanyu Hazarika, Rada Mihalcea, Alexander Gelbukh, and Erik Cambria. 2019. Dialoguernn: An attentive rnn for emotion detection in conversations. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 6818–6825.

Marvin Minsky. 1988. *Society of mind*. Simon and Schuster.

OpenAI. 2023. Gpt-4 technical report.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback.

Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Ion Androutsopoulos, Suresh Manandhar, Mohammed AL-Smadi, Mahmoud Al-Ayyoub, Yanyan Zhao, Bing Qin, Orphée De Clercq, et al. 2016. Semeval-2016 task 5: Aspect based sentiment analysis. In *ProWorkshop on Semantic Evaluation (SemEval-2016)*, pages 19–30. Association for Computational Linguistics.

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 (volume 1: Long papers)*, pages 873–883.

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

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Weizhou Shen, Siyue Wu, Yunyi Yang, and Xiaojun Quan. 2021. Directed acyclic graph network for conversational emotion recognition. *arXiv preprint arXiv:2105.12907*.

Yongliang Shen, Kaitao Song, Xu Tan, Dongsheng Li, Weiming Lu, and Yueting Zhuang. 2023. Hugginggpt: Solving ai tasks with chatgpt and its friends in huggingface.

Xiaohui Song, Longtao Huang, Hui Xue, and Songlin Hu. 2022. Supervised prototypical contrastive learning for emotion recognition in conversation. *arXiv preprint arXiv:2210.08713*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models.

Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2021. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models.

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.

Liu Yingjian, Li Jiang, Wang Xiaoping, and Zeng Zhigang. 2023. Emotionic: Emotional inertia and contagion-driven dependency modelling for emotion recognition in conversation. *arXiv preprint arXiv:2303.11117*.

Zheng Yuan, Hongyi Yuan, Chengpeng Li, Guanting Dong, Chuanqi Tan, and Chang Zhou. 2023. Scaling relationship on learning mathematical reasoning with large language models.

Sayyed M Zahiri and Jinho D Choi. 2017. Emotion detection on tv show transcripts with sequence-based convolutional neural networks. *arXiv preprint arXiv:1708.04299*.

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 (Volume 1: Long Papers)*, pages 7395–7408, Toronto, Canada. Association for Computational Linguistics.

Ting Zhang, Zhuang Chen, Ming Zhong, and Tieyun Qian. 2023b. Mimicking the thinking process for emotion recognition in conversation with prompts and paraphrasing. *arXiv preprint arXiv:2306.06601*.

Yupeng Zhang, Shensi Wang, Peiguang Li, Guanting Dong, Sirui Wang, Yunsen Xian, Zhoujun Li, and Hongzhi Zhang. 2023c. Pay attention to implicit attribute values: A multi-modal generative framework for AVE task. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13139–13151, Toronto, Canada. Association for Computational Linguistics.

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

Table 3: Unified Label Mapping

| Number | IEMOCAP | MELD | EmoryNLP | Final Emotion |
|---|---|---|---|---|
| 1 | happy | joyful | joyful | joyful |
| 2 | sad | sad | sad | sad |
| 3 | neutral | neutral | neutral | neutral |
| 4 | angry | angry | mad | mad |
| 5 | excited | N\A | N\A | excited |
| 6 | N\A | surprise | powerful | powerful |
| 7 | scared | fear | frustrated | fear |
| 8 | N\A | N\A | peaceful | peaceful |
| 9 | N\A | disgust | N\A | disgust |

Table 4: One-hot Speaker Label Mapping

| Speaker label | IEMOCAP | MELD | EmoryNLP |
|---|---|---|---|
| 1 | 1 | N\A | N\A |
| ... | ... | N\A | N\A |
| $n_1$ | $n_1$ | N\A | N\A |
| $n_1 + 1$ | N\A | 1 | N\A |
| ... | N\A | ... | N\A |
| $n_1 + n_2$ | N\A | $n_2$ | N\A |
| $n_1 + n_2 + 1$ | N\A | N\A | 1 |
| ... | N\A | N\A | ... |
| $n_1 + n_2 + n_3$ | N\A | N\A | $n_3$ |

## A The Details of Unified Dataset Experiment Setup

To further substantiate the efficacy and robustness of our framework, we conduct a compelling experiment involving a unified dataset. Within the settings of this experiment, all emotional labels across the datasets are standardized, and all speaker labels are also consolidated. Subsequently, we conduct data scaling experiments on the processed unified dataset. The evaluation method employed in the experimental results, utilizing the weighted F1 score, aligned with the evalution method delineated in Section Experiments.

We continue to use the previous datasets IEMOCAP, MELD, and EmoryNLP. According to The Feeling Wheel (Willcox, 1982) proposed in 1982, as shown in subfigure of Figure 4, we align all emotional labels from three datasets with this standard, the details of which are shown in Tabel 3. After completion of label mapping, there are a total of 9 types of emotional labels, which are *joyful, sad, neutral, mad, excited, powerful, fear, peaceful and disgust*. Furthermore, due to the uniqueness of character labels in each dataset, we have renumbered them using a One-hot encoding approach, as demonstrated in the "One-hot Speaker Label Mapping" Table 4, which also is shown in subfigure of Figure 4.

We still utilize the LoRA method in PEFT to train InstructERC on the unified dataset, and the training results are evaluated on the three datasets respectively. As mentioned above, these datasets have significant variations in sample size and class imbalance within each dataset. To explore the impact of different sampling methods on the final performance, two data scaling approaches were experimented with: Total Mixing and Ratio Mixing.

In the Total Mixing approach, all datasets are combined for uniform sampling. Conversely, in the Ratio Mixing approach, datasets are sampled separately and then combined. Both approaches maintain the same quantity of training data, but due to the larger absolute number of training samples in MELD and EmoryNLP, the Total Mixing approach results in a higher proportion of samples from these two datasets when varying data scaling is applied.

Total Mixing and ratio Mixing modes are applied proportionally across the entire training set, while still segregating a validation set and a test set. The reported results are obtained after training on a unified training set and then testing on separate test sets. The Single mode, on the other hand, involves training on individual training sets and then testing on their respective test sets.

Meanwhile, we design Total Mixing and Ratio Mixing experiments to explore the impact of different data mixing strategies and data quantities on the model. On the basis of the following, we further explore the impact of data sampling ratio on the model's performance.The details of results are shown in Table 5, and a more intuitive presentation is shown in Figure 5.

## B Related Works

### B.1 Emotion Recoginition in Conversation

After more than a decade of development, the field of Emotion Recognition in Conversation (ERC) has seen many outstanding works. These can be broadly classified into three categories: Transformer-based, GNN-based, Recurrent-based.

Specifically, **Transformer-based** works (Li et al., 2020; Song et al., 2022; Liu et al., 2023; Yingjian et al., 2023; Chudasama et al., 2022) attempt to establish long-range emotional correlations in conversational scenarios by directly adopting or modifying the original transformer block. These efforts have made significant contributions in this direction.

**GNN-based** works (Ghosal et al., 2019; Ishiwatari et al., 2020; Shen et al., 2021; Li et al.,

Table 5: The Unified Dataset Experiments of Llama2 on three benchmarks

| Data Precent | IEMOCAP W-F1 | | | MELD W-F1 | | | EmoryNLP W-F1 | | |
|---|---|---|---|---|---|---|---|---|---|
| | Total Mixing | Ratio Mixing | Single | Total Mixing | Ratio Mixing | Single | Total Mixing | Ratio Mixing | Single |
| 1 | 68.99 | 68.99 | **71.39** | 68.07 | 68.07 | **69.15** | 40.27 | 40.27 | **41.37** |
| 1/2 | 67.95 | 68.96 | **69.13** | 66.50 | 66.42 | **67.54** | 39.18 | 39.33 | **39.65** |
| 1/4 | 63.02 | 64.46 | **67.54** | 66.41 | 65.85 | **66.42** | 38.26 | 37.29 | **38.33** |
| 1/8 | 58.48 | 60.06 | **64.13** | 64.57 | 62.94 | **65.14** | 38.27 | **39.24** | 38.24 |
| 1/16 | 57.77 | 53.40 | **60.42** | 61.15 | 58.42 | **62.89** | 37.19 | **37.60** | 36.83 |
| 1/32 | 45.89 | 48.50 | **54.76** | 57.38 | **57.76** | 57.72 | **37.09** | 36.09 | 34.03 |
| 1/64 | 38.42 | **43.07** | 30.34 | **54.26** | 53.29 | 45.48 | **35.19** | 34.65 | 26.10 |

2023a) extensively use graphs and edges to model interactions between people in conversational scenarios and the influences between different modalities. They employ various forms of multi-layer graph neural networks to fit potential conversational relations, effectively exploring this direction.

**Recurrent-based** works (Hu et al., 2023; Lei et al., 2023; Majumder et al., 2019; Hazarika et al., 2018; Poria et al., 2017) utilize various forms of RNNs, like LSTM and GRU, to model individual emotional states and global emotional impacts separately. They incorporate attention mechanisms or direct vector concatenation to represent personal and global emotional states collectively, marking effective exploration in this area.

### B.2 Large Language Models

The emergence of large-scale language models (LLMs) have brought revolutionary transformation to the field of natural language processing (NLP) (Shen et al., 2023). LLMs, such as GPT-3 (Brown et al., 2020), LLaMA (Touvron et al., 2023) and GPT-4 (OpenAI, 2023), have demonstrated impressive abilities on various tasks, as well as the use of external techniques such as reinforcement learning from human feedback (RLHF) (Ouyang et al., 2022). LLMs based on generative framework even reformulate the multi modal perspective (Lin et al., 2021; Zhang et al., 2023c). More recently, the NLP community has been exploring various application directions for LLMs. For instance, chain-of-thought prompting and RFT (Wei et al., 2023; Yuan et al., 2023) enables LLMs to generate problem-solving processes step-by-step, significantly enhancing the model's reasoning ability. Researchers have utilized the interactive capabilities of LLMs to generate commands that invoke external tools for handling of downstream tasks(Shen et al., 2023). Other researchers have proposed parameter-efficient fine-tuining (PEFT) to address the issue of excessive computational resource without sacrificing performance (Hu et al., 2021).

## C Datasets & Baselines

### C.1 Datasets

**IEMOCAP** (Busso et al., 2008) is a dataset recorded as dyadic conversational video clips with eight speaker participating in the training set while two speaker in testing set.

**MELD** dataset (Poria et al., 2018) is a multimodal dataset that has been expanded from the EmotionLines dataset. MELD is obtained from the popular TV show *Friends* and comprises over 1400 dialogues and 13000 utterances, each of which is labeled with emotion and sentiment classes.

**EmoryNLP** (Zahiri and Choi, 2017) is a dataset also collected from the TV series *Friends*. The dataset comprises utterances that are categorized into seven distinct emotional classes.

This study exclusively focuses on the emotional classes and the text modality in these datasets. Moreover, we ensure consistency with COSMIC regarding the train/val/test splits.

### C.2 Baselines

For discriminative ERC models, we selected several **SOTA** baseline for each method. For our reconstructed generative model, we chose four popular LLMs as backbones.

**Recurrent-based**: **1) EmotionIC** (Yingjian et al., 2023) uses IM-MHA and DialogGRU to capture contextual information in the dialogue, and SkipCRF to capture high-order dependencies between speakers for emotional flow simulation. **2) SACL-LSTM** (Hu et al., 2023) extracts structured representations using contrast-aware adversarial training and joint class-spread contrastive learning, an additional contextual adversarial training strategy to enhance context robustness.

**Transformer-based**: **1) MPLP** (Lu et al., 2022) is a framework that unifies multimodal sentiment analysis and emotion recognition in conversation

Table 6: The statistics of datasets. avg_utt denotes the average number of utterances in a conversation.

| Datasets | Conversations<br>Train | <br>Val | <br>Test | Utterances<br>Train | <br>Val | <br>Test | classes | type | avg_utt | Evaluation |
|---|---|---|---|---|---|---|---|---|---|---|
| IEMOCAP | 108 | 12 | 31 | 5163 | 647 | 1623 | 6 | two-person | 47 | W-F1 |
| MELD | 1038 | 114 | 280 | 9989 | 1109 | 2610 | 7 | multi-party | 9 | W-F1 |
| EmoryNLP | 713 | 99 | 85 | 9934 | 1344 | 1328 | 7 | multi-party | 11 | W-F1 |

tasks. This framework achieves this by performing modality fusion at both the syntactic and semantic levels, and by introducing contrastive learning between modalities and samples. **2) SPCL** (Song et al., 2022) is a method that addresses imbalanced classification issues using Prototypical Network and contrastive learning, without the need for large batch sizes, and incorporates a difficulty measure function and curriculum learning to mitigate the effects of extreme samples.

**GNN-based**: **1) DualGATs** (Li et al., 2021) uses a connected graph to enhance the targeted utterance with information from the past and future context, and utilizes CommonSense Knowledge (CSK) to enrich edges with knowledge representations. **2) Skier** (Li et al., 2023a) is a module that efficiently models contextual and interactive information for ERC task. It uses multiple extractors and PairCC strategy to address the heterogeneity gap in multimodal fusion.

**LLM backbones: 1) ChatGLM-6B & ChatGLM2-6B**: ChatGLM-6B is an open-source conversational language model (Du et al., 2022) for Chinese and English. It has 6.2 billion parameters and is optimized for Chinese QA. It has been trained on 1 trillion Chinese and English identifiers and further improved through various techniques. ChatGLM2-6B is the second generation of the model, pre-trained on 1.4 trillion Chinese and English identifiers with human preference alignment training. It extends the context window to 32K and speeds up inference with Multi-Query Attention. **2) Llama-7B & Llama2-7B**: Llama-7B is the 7B parameters' version of the a collection of foundation language models (Touvron et al., 2023) ranging from 7B to 65B parameters, which is trained on trillions of tokens. Llama2-7B pre-trained models are trained on 2 trillion tokens, and have double the context length than Llama 1. Its fine-tuned models have been trained on over 1 million human annotations.

# D Implementation & Discussion

## D.1 Implementation Details

We use ChatGLM and Llama as our backbone model. Considering the efficiency and effectiveness of Parameter-Efficient-Fine-Tuning (PEFT), we adopt LoRA (Hu et al., 2021) and insert low-rank adapters after self-attention layers. We set the dimension of adapters to 16 a nd the learning rate to 2e-4. The learning rate is set to 2e-5 for all parameters' finetune. The histoical window is set to 1, 5, 12, 20 for iemocap, meld and EmoryNLP respectively for all experiments. The retrieval parameter "TopK" is set to Top1 emprically. The hypermeter $\alpha$ is set to 0.1 during training. Greedy search is used during inference if not specified. Moreover, our experiments are conducted by taking the average of three runs with no hyperparameter searching. We train with FP16 precision on 4 × 80G Nvidia A100 GPUs.

## D.2 Discussion with Discriminative ERC Models

**Problem definition.**

In the discriminative framework, researchers first fine-tune an RoBERTA-style model with the context-free utterance, extract the feature vector at the CLS position as the input for the downstream ERC model. The aim is to map the feature vector of the given utterance to a scalar between 1 and $o$.

In the generative framework based on LLMs, for a given utterance, we process it into formatted text according to the pre-designed template and input it into LLMs. The aim is to enable LLMs generate the most reasonable text emotional label, which must belong to the predefined text emotional label set $\mathcal{E} = \{e_1, e_2, ..., e_o\}$.

**Parameter Scales.** As shown in Table 7, we present the publicly available statistics for all trainable parameters across the models. Although the base architecture of our model is in the 6-7B parameter range, only 12.5M LoRA parameters are actively trained, which is feasible on a single GPU. For example, on the IEMOCAP dataset, our model

Table 7: The more detailed results and Statistics on three benchmarks.

| Dataset | Parameters | IEMOCAP | MELD | EmoryNLP | Average | Extra | Model type |
|---|---|---|---|---|---|---|---|
| Models | | W-avg F1 | W-avg F1 | W-avg F1 | W-avg F1 | Knowledge | |
| Small-scale Discriminant ERC-specific Model | | | | | | | |
| KET* | 2.6M | 59.56 | 58.18 | 34.39 | 50.17 | ConceptNet | transformer |
| TODKAT† | 330M | 61.33 | 65.47 | 38.69 | 55.16 | COMET | transformer |
| MTL* | 1.2M | — | 61.90 | 35.92 | — | ✗ | transformer |
| CoG-BART* | 415.1M | 64.87 | 63.82 | 37.33 | 55.34 | ✗ | transformer |
| M2FNet* | — | 69.86 | 66.71 | — | — | ✗ | transformer |
| SPCL† | 356.7M | 68.42 | 66.13 | *40.25* | 58.26 | ✗ | transformer |
| Hidialog* | — | — | 66.96 | — | — | ✗ | transformer |
| SACL-LSTM* | 2.6M | 69.22 | 66.45 | 39.65 | 58.44 | ✗ | recurrent |
| HCAN† | 3.5M | 69.21 | 66.24 | 39.67 | 58.37 | ✗ | recurrent |
| ICON* | 0.5M | 63.50 | — | — | — | ✗ | recurrent |
| DialogueRNN† | 9.9M | 64.65 | 65.30 | 37.54 | 55.83 | ✗ | recurrent |
| DialogueCRN† | 3.3M | 67.53 | 65.77 | 38.79 | 57.36 | ✗ | recurrent |
| EmotionIC* | — | 69.50 | 66.40 | 40.01 | *58.63* | ✗ | recurrent |
| CauAIN* | 6.1M | 65.01 | 64.89 | 37.87 | 55.92 | ATOMIC | recurrent |
| COIN* | 0.5M | 65.37 | — | — | — | ✗ | recurrent |
| COSMIC† | 11.9M | 65.03 | 63.43 | 38.49 | 55.65 | COMET | recurrent |
| DialogueGCN† | 2.1M | 62.11 | 62.68 | 36.43 | 53.14 | ✗ | GNN |
| RGAT* | 13M | 65.22 | 60.91 | 34.42 | 53.52 | ✗ | GNN |
| SKAIG* | — | 66.96 | 65.18 | 38.88 | 57.01 | COMET | GNN |
| DAG-ERC† | 9.5M | 66.54 | 63.36 | 38.29 | 56.06 | ✗ | GNN |
| GraphCFC* | — | 68.91 | 58.86 | — | — | ✗ | GNN |
| Small-scale Pretrained Language Model | | | | | | | |
| KI-NET* | 500M | 67.00 | 63.24 | — | — | ConceptNet | transformer |
| DialogueXL* | 510M | 65.94 | 62.41 | 34.73 | 54.36 | ✗ | transformer |
| EmoBERTa* | 355M | 68.57 | 65.61 | — | — | ✗ | transformer |
| UniMSE* | 220M | *70.66* | 65.51 | — | — | ✗ | transformer |
| Zero-shot + InstructERC | | | | | | | |
| ChatGLM † | 12.5M(6B) | ***38.6*** | ***38.8*** | 19.6 | ***32.33*** | ✗ | LLM-based |
| ChatGLM2 † | 12.5M(6B) | 21.1 | 21.8 | ***24.4*** | 22.43 | ✗ | LLM-based |
| Llama † | 12.5M(7B) | 0.753 | 9.12 | 5.31 | 5.06 | ✗ | LLM-based |
| Llama2 † | 12.5M(7B) | 2.774 | 16.28 | 8.36 | 9.46 | ✗ | LLM-based |
| LoRA + Backbone | | | | | | | |
| ChatGLM † | 12.5M(6B) | 18.94 | 40.54 | 25.71 | 28.07 | ✗ | LLM-based |
| ChatGLM2† | 12.5M(6B) | 52.88 | 64.85 | 37.69 | 51.80 | ✗ | LLM-based |
| Llama† | 12.5M(7B) | 55.81 | ***66.15*** | 37.98 | 53.21 | ✗ | LLM-based |
| Llama2† | 12.5M(7B) | ***55.96*** | 65.84 | ***38.21*** | ***53.33*** | ✗ | LLM-based |
| LoRA + InstructERC | | | | | | | |
| ChatGLM† | 12.5M(6B) | 36.04 | 46.41 | 30.86 | 37.77 | ✗ | LLM-based |
| ChatGLM2† | 12.5M(6B) | 67.54 | 65.58 | 39.09 | 57.40 | ✗ | LLM-based |
| Llama† | 12.5M(7B) | 64.17 | 67.62 | 39.34 | 57.04 | ✗ | LLM-based |
| Llama2† | 12.5M(7B) | **71.39** | **69.15** | **41.37** | **60.64** | ✗ | LLM-based |

NOTE: The best-performing results of other models are highlighted in gold font, while SOTA results across all models are emphasized in red font. Models annotated with an * indicate results sourced from the model's paper, and a (†) denotes results from reproductions conducted by the authors.

typically converges by the 6th epoch, taking approximately 2 hours. The inference process requires about 10 minutes to handle 1000 samples. While our method is marginally slower than other approaches, such as the SPCL baseline which utilizes 356.7M training parameters, this speed reduction is not a significant drawback and remains manageable for most research contexts.

**Structural Complexity.** As shown in Table 7 and Figure 1, taking the influential work such as COSMIC as an example, COSMIC fine-tuned RoBERTA on single-sentence dialogues, extracted its features, and encapsulated them into a dataset. Many works in the baseline are based on the feature dataset extracted from this work rather than the original text data for downstream model design. This means that these models (including but not limited to all the compared baselines which adopt this practice) need to use the single-sentence speech features fine-tuned with emotional labels during inference, which clearly does not conform to reality (the sentences that need to perform emotion recognition cannot access the gold emotional labels in advance). Furthermore, even if these single-sentence features do not need fine-tuning, it is still necessary to use Roberta to infer and obtain features.

In contrast, our InstructERC can directly input text and output emotional labels. Additionally, the InstructERC framework can be migrated to multiple datasets and combine datasets across multiple domains without modification, whereas discriminative models require manual changes to the architecture of the model, specifically the number of softmax classification neurons in the last layer, to perform multi-domain operations. In terms of scalability, the generative model InstructERC is clearly more practical than discriminative models.

## E The Supplementary Experiments

### E.1 The historical window exploration study

Table 8: The historical window exploration of Llama2 on three benchmarks.

| histoical window | IEMOCAP W-F1 | MELD W-F1 | EmoryNLP W-F1 |
|---|---|---|---|
| LoRA + LLaMA2 + InstructERC | | | |
| 1 | $56.12_{\pm 1.40}$ | $65.91_{\pm 0.46}$ | $38.32_{\pm 0.38}$ |
| 5 | $68.65_{\pm 0.32}$ | $66.97_{\pm 0.21}$ | $40.48_{\pm 0.23}$ |
| 12 | $\mathbf{71.39}_{\pm 0.10}$ | $\mathbf{69.15}_{\pm 0.09}$ | $\mathbf{41.37}_{\pm 0.11}$ |
| 20 | $71.01_{\pm 0.12}$ | $68.75_{\pm 0.12}$ | $40.56_{\pm 0.15}$ |

In the historical window exploration shown as Table 8, we examine how different sizes of historical windows affect emotion recognition tasks. Due to token limitations, we set the upper limit for conversational turns to 20. This is an upgrade from earlier, smaller Pretrained Language Models (PLMs, e.g. Roberta(Liu et al., 2019)), which only support up to 5 turns. We find that a window of 12 turns is optimal for capturing the necessary historical context. In general, expanding the count of historical turns aids in enhancing the accuracy of emotion detection, a trend that is readily observable in the IEMOCAP dataset featured long-term turns. However, there's a point where adding more historical turns doesn't lead to better results and might even harm performance, especially for datasets like MELD and EmoryNLP, which have an average length of 6 to 7 turns. However, these insights are beyond the reach of smaller PLMs that top out at 5 turns.

### E.2 The Exploration Experiments on $\alpha$

Table 9: The exploration experiments on $\alpha$.

| $\alpha$ | IEMOCAP W-F1 | MELD W-F1 | EmoryNLP W-F1 |
|---|---|---|---|
| LoRA + LLaMA2 + InstructERC | | | |
| 0 | 70.50 | 68.97 | 40.78 |
| 0.05 | 70.67 | 69.03 | 40.91 |
| 0.1 | 71.39 | 69.15 | 41.37 |
| 0.2 | 71.14 | 68.54 | 40.63 |

Shown as Table 9, the influence of alpha on InstructERC's performance varies across different datasets due to their unique characteristics. In general, as alpha increases, its contribution to model performance also increases, peaking at alpha=0.1. Specifically, in the IEMOCAP dataset, characterized by longer dialogues averaging 47 turns, even when alpha exceeds 0.1 significantly, there is no significant decrease in performance. However, in datasets like MELD and EmoryNLP, which have shorter dialogues averaging 7 turns, an alpha value of 0.2 can lead to a negative impact, particularly evident in MELD. Therefore, careful consideration is necessary when selecting alpha values for different datasets.

This phenomenon can be explained as follows: In the IEMOCAP dataset, with its longer dialogues, emotional changes occur relatively slowly. In contrast, datasets like MELD and EmoryNLP, sampled from the sitcom “Friends”, feature many brief and intense emotional shifts. Excessive reduction in the weight of emotion impact prediction may cause the model to overly emphasize the influence of past utterances on current emotion judgment, which may not be suitable for MELD and EmoryNLP.

### E.3 Label Ablation Experiments

To further explore the impact of using the same or unrestricted emotional labels at different stages during the demonstration retrieval process on final performance, we designed experiments as shown in the table 10, where ×represents not using the same labels, and ✓represents using the same labels. Our conclusions are as follows:

**Impact of Label Restrictions:** The performance consistently improves across all datasets when moving from unrestricted to restricted labels in both training and inference. This suggests that restricting labels helps the model learn more robust features that are better at generalizing during inference.

**Comparison Across Datasets:** IEMOCAP: Shows a steady increase in W-F1 scores as restrictions are applied first in training and then in both training and inference. The improvement from fully unrestricted to fully restricted is 1.54 points. MELD: Similar to IEMOCAP, restricted training and inference show a noticeable improvement. The gain from the least to the most restricted setup is 2.54 points, indicating a potentially more significant impact of label restriction in emotionally complex interactions, possibly due to MELD’s diverse emotional content and real-life scenarios. EmoryNLP: This dataset shows the lowest overall scores but follows the same trend. The increase is 2.14 points from no restrictions to full restrictions. Given the smaller base score, this improvement is quite significant, emphasizing how crucial precise label handling is in models trained on this data.

**Fairness and Performance Trade-offs:** The best results obtained by using restricted labels in both phases might not be fair or realistic for real-world applications, where the model shouldn’t have prior knowledge of the emotional context. This indicates a need for models that perform well under unrestricted conditions. The performance drop when moving to unrestricted labels in inference underscores the challenge in generalizing the learned emotional cues without specific hints, highlighting a potential area for further research in enhancing model robustness.

Table 10: Dataset performance with various restrictions on labels during training and inference.

| **Dataset** | **Training** | **Inference** | **W-F1** |
|---|---|---|---|
| IEMOCAP | × | × | 70.71 |
| IEMOCAP | ✓ | × | 71.39 |
| IEMOCAP | ✓ | ✓ | 72.25 |
| MELD | × | × | 68.52 |
| MELD | ✓ | × | 69.15 |
| MELD | ✓ | ✓ | 71.06 |
| EmoryNLP | × | × | 40.54 |
| EmoryNLP | ✓ | × | 41.37 |
| EmoryNLP | ✓ | ✓ | 42.68 |

Table 11: The comparison results of different parameter fine-tuning settings on three benchmarks.

| Dataset<br>Models | IEMOCAP<br>W-F1 | MELD<br>W-F1 | EmoryNLP<br>W-F1 | Average<br>W-F1 |
|---|---|---|---|---|
| All parameters + InstructERC | | | | |
| ChatGLM† | 33.94 | 37.96 | 13.25 | 28.38 |
| ChatGLM2† | 70.05 | 63.24 | ***38.77*** | 57.35 |
| Llama† | ***69.38*** | ***66.01*** | 40.21 | ***58.53*** |
| Llama2† | 70.30 | 64.80 | 40.05 | 58.38 |
| LoRA + InstructERC | | | | |
| ChatGLM† | 36.04 | 46.41 | 30.86 | 37.77 |
| ChatGLM2† | 67.54 | 65.58 | 39.09 | 57.40 |
| Llama† | 69.71 | 68.89 | 39.90 | 59.50 |
| Llama2† | **71.39** | **69.15** | **41.37** | **60.64** |

### E.4 All Parameters vs Parameter Efficiency

In order to investigate the effect of different parameter fine-tuning methods on the ERC task, we conducted comparative experiments in Table 11. We have the following observations:

(1) The all parameter fine-tuning performs weaker than LoRA’s fine-tuning on all backbones on average performance (especially ChatGLM with a 9.32 % improvement). It is worth noting that the best performance of the full parameter method is often achieved in the first 1-3 epochs in the experiment. These findings demonstrate that parameter-efficient methods are more suitable for LLMs in ERC tasks.

(2) From the perspective of model structure, the average performance of full parameter ChatGLM even decreases compared to the zero-shot results in Table 1 (from 32.33% to 28.38%), while replacing it with LoRA brings a significant improvement (from 32.33% to 37.77%). Other decoder-only backbones do not show such drastic performance fluctuations, which further indicates that the prefix-decoder paradigm is unstable in ERC tasks compared to the casual decoder, and parameter-efficient

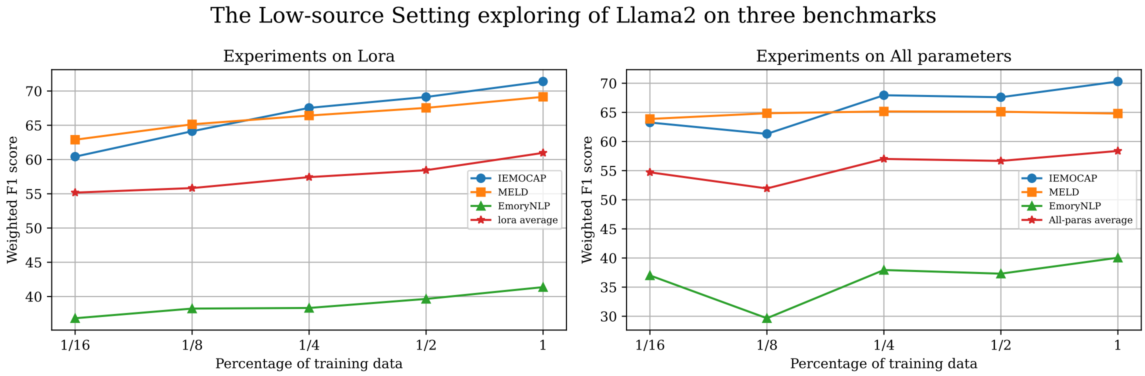


Figure 6: The scaling of data and performance for different parameter fine-tuning settings (LoRA & All Parameters)

frameworks can effectively alleviate this problem.

(3) From the perspective of datasets, compared to full parameter fine-tuning, the performance gain of the LoRA method in MELD and EmoryNLP is significantly greater than that in IEMOCAP. We believe that this is related to the characteristics of thees datasets: IEMOCAP has long dialogue texts and multiple conversation rounds, these strong supervision signals lead to good performance in both settings. However, MELD and Emory have fewer dialogue rounds, diverse speakers, and imbalanced categories. Low-parameter methods can effectively prevent LLMs from overfitting to certain semantic patterns of dialogues format and speaker's habits, thereby enhancing the generalization ability of emotion recognition in conversation.

### E.5 Scaling Analysis in Low-source Scenario

In this section, we gain an insight into the scaling analysis of data and performance for different parameter fine-tuning settings (LoRA & All Parameter), as shown in Figure 6.

**Parameter-efficient Scaling Analysis**: On the IEMOCAP dataset, our scaling curve initially increases (from 1/16 to 1/4) and then stabilizes. This may be because the dataset has long dialogue texts and multiple dialogue rounds, leading to increased diversity with the addition of early data. However, as the supervision signal strengthens, the performance gain gradually weakens. For datasets with fewer dialogue rounds and imbalanced categories, such as MELD and EmoryNLP, our method only yields a small gain in extremely low-resource scenarios (from 1/16 to 1/4) and achieves a relatively stable performance improvement with the increase of data (from 1/2 to 1). This finding supports the idea that when a unit-scaling of data only provides weak supervision signals, the data size needs to exceed a certain threshold (1/4 - 1/2) to achieve significant improvement.

**Full-Parameter Scaling Analysis**: The scaling curves of full-parameter settings on the IEMOCAP and EmoryNLP datasets showed significant fluctuations and performance degradation in two intervals (from 1/16 to 1/8, 1/4 to 1/2) compared to LoRA. Fine-tuning large models with all parameters may cause redundant parameters to overfit the patterns in the current dialogue, which hinders the model's ability to generalize new supervised signals as data volume increases. The MELD dataset also exhibited performance degradation with data augmentation (from 1/4 to 1). These findings demonstrate the stability and robustness of parameter-efficient fine-tuning in the ERC task, providing empirical guidance for large models in industrial interfaces with ERC tasks of varying data characteristics.