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

Shanglin Lei[12†], Guanting Dong[3†], Xiaoping Wang[12*], Keheng Wang[4], Sirui Wang[5]
[1]School of Artificial Intelligence and Automation, Huazhong University of Science and Technology , Wuhan , China
[2]Key Laboratory of Image Processing and Intelligent Control of Education Ministry of China, Wuhan, China
[3]School of Artificial Intelligence, Beijing University of Posts and Telecommunications, Beijing, China
[4]Sino-French Engineer School, Beihang University, Beijing, China
[5]Meituan Inc., Beijing, China
{lawson, wangxiaoping}@hust.edu.cn, dongguanting@bupt.edu.cn, wkh9575638@buaa.edu.cn, wangsirui@meituan.com

Abstract

The development of emotion recognition in dialogue (ERC) has been consistently hindered by the complexity of pipeline designs, leading to ERC models that often overfit to specific datasets and dialogue patterns. In this study, we propose a novel approach, namely InstructERC, to reformulates the ERC task from a discriminative framework to a generative framework based on Large Language Models (LLMs) . InstructERC has two significant contributions: Firstly, InstructERC introduces a simple yet effective retrieval template module, which helps the model explicitly integrate multi-granularity dialogue supervision information by concatenating the historical dialog content, label statement, and emotional domain demonstrations with high semantic similarity. Furthermore, 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. Our LLM-based plug-and-play 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 provide empirical guidance for applying InstructERC in practical scenarios. Our code will be released after blind review.

Introduction

"The question is not whether intelligent machines can have emotions, but whether machines without emotions can achieve intelligence", as pointed out by the pioneer of artificial intelligence, Minsky, in his book "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 recent years, the task of dialogue emotion recognition has become a hot research topic in the field of natural language processing (NLP) due to its enormous potential application scenarios in human-computer interaction (MacKenzie 2012) and dialogue systems (Ma et al. 2020). [1]

In contrast to conventional binary sentiment analysis tasks (Pontiki et al. 2016) , which only rely on text with explicit attitude tendencies, the dialogue emotion recognition (ERC) task aims to identify more fine-grained emotional tendencies in each sentence of a conversation. Currently, the ERC task faces challenges primarily in two aspects: context modeling and speaker modeling. Specifically, for the ERC task, the recognition of emotional tendencies in the target sentence is highly dependent on its historical utterances, and different speakers exhibit significant variations in perceiving and conveying emotions. Therefore, it is imperative to meticulously model the speakers and dialogue context.

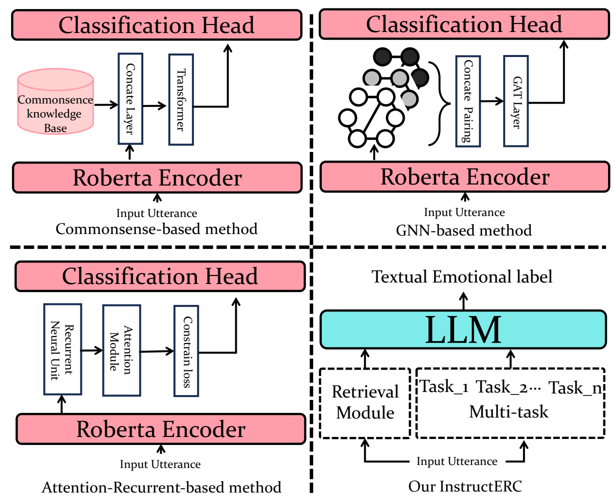

Figure 1: The illustration of different paradigms for ERC

[2] Figure 1 illustrates that previous work based on Roberta (Liu et al. 2019) in ERC can be roughly divided into three categories: attention-based, recurrent-based, and commonsense-based. (1) **Attention-Recurrent-based methods** (Yingjian et al. 2023; Hu et al. 2023; Hu, Wei, and Huai 2021), often design different levels of encoders and recurrent neural units to extract emotional features. However, bidirectional encoding allows the current utterance to consider information from the future, which is clearly illogical. (2) **Commonsense-based methods** (Ghosal et al. 2020;

[1]† The first two authors contribute equally.
Work done during internship at Meituan Inc.
* Xiaoping Wang is the corresponding author.

[2]You can find the code on [GitHub](https://github.com/LIN-SHANG/InstructERC).

Li et al. 2021) usually use external knowledge bases to enhance semantics, but this greatly increases the consumption of computational resources and may introduce additional noise. (3) **GNN-based methods** (Liang et al. 2021; Shen et al. 2021) typically use nodes and edges to model characters and dialogue relationships in conversations. However, the numerous types of nodes and edges can lead to redundancy and overfitting to a single dataset. In summary, these Highly structured models tend to overfit to specific data features (long dialogues or complex dialogue relationships), leading them difficult to generalize to real-world scenarios where dialog content distributions are unknown.

Fortunately, the recent successful application and emergence capabilities of pre-trained large language models (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 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 real dialogue system.

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 emotional domain 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.

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 unified Seq2Seq paradigm and present a 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. Further analysis provides empirical guidance for application in practical scenarios.

Related Work

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. 2023). 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).

Emotion Recognition in Conversation

As a task that has been increasingly receiving attention in the NLP field, the ERC domain has recently seen many works based on encoder. Methods based on commonsense usually utilize external knowledge bases to semantically enhance the original datasets, in order to obtain more distinctive utterance features for emotional recognition. However, this often leads to additional noise. (Dong et al. 2022; Guo et al. 2023) Utilizing a hybrid attention-recurrent network structure, techniques have the potential to provide a more accurate and faithful depiction of the emotional dynamics in a conversation. On the other hand, GNN-based methods (Liang et al. 2021; Shen et al. 2021; Zhao et al. 2022) employ Graph Neural Network to facilitate coarse-grained and fine-grained extraction of emotional features from dialog. Recently, prompt-based methods have shown great capabilities in NLP tasks. They transformed the traditional task paradigm into a unified Seq2Seq framework (Lu et al. 2022; Wang et al. 2022; Li et al. 2023b; Qixiang et al. 2022; Wang et al. 2023). However, ERC tasks are still an unexplored field, which motivates us to reformulate ERC tasks as a unified generative paradigm.

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.

Problem Definition

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.

In the generative framework based on LLMs, for a given utterance, we process it into formatted text according to the

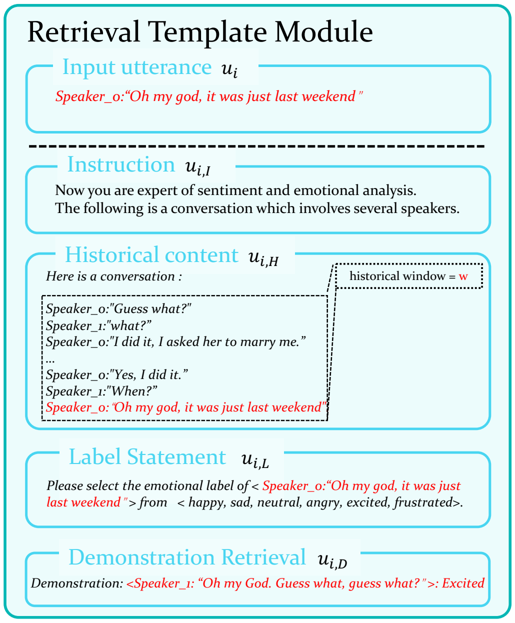

Figure 2: The Schematic of Retrieval Template Module.

pre-designed template and input it into LLMs. The objective of ERC in this case is to enable LLMs generate the most reasonable text emotional label, which must belong to the predefined text emotional label set $\mathcal{E} = \{e_1, e_2, ..., e_o\}$. o is the number of emotoinal categories.

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-retrieve.

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. The ERC task is heavily reliant on contextual information, yet in daily conversations, the affective state of a speaker in the present moment is impervious to the emotional influence of future utterances. Therefore, the historical content that is included in the model's input is limited to those utterances that precede the current recognized utterance. We employ a hyperparameter, the historical window (denoted as w), to indicate the specific rounds 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, facilitate statistical analysis of the model's output, and enable the model to focus on the current utterance being recognized, our label statement $u_{i,L}$ is shown in Figure 2.

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:

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

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:

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

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.

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, which is 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, 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* $<p_1,...,p_M>$*"*

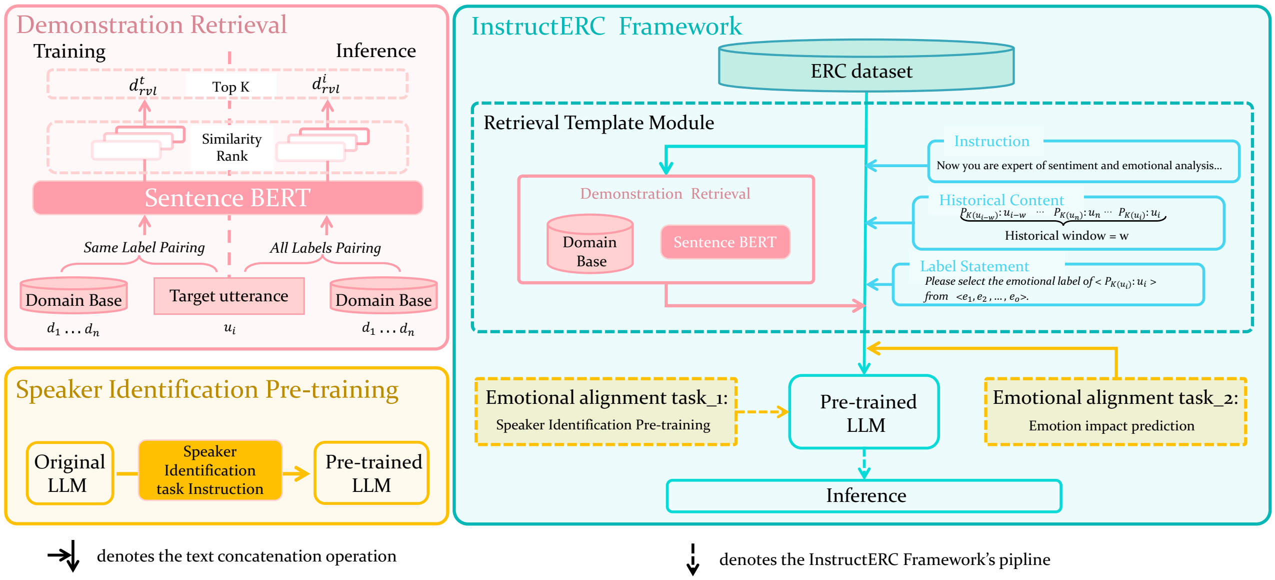

Figure 3: The overview of InstructERC framework

The loss function for the Speaker Identification is as follows:

$$\mathcal{L}_p = \sum_{i}^{N} -\log P(\mu_i | x_i^p, \theta) \tag{3}$$

Here, μ_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 θ represents the parameters of the LLM.

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.

To be specific, we maintain the instruction part $u_{i,I}$ of the input x_i of the main task and modify the historical content $u_{i,H}^e$ of the target statement u_i for the Emotion Impact Prediction task. The corresponding label statement $u_{i,L}^e$ 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:

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

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

$$\mathcal{L}_e = \sum_{i}^{N} -\log P(\epsilon_i | x_i^e, \theta) \tag{5}$$

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

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:

$$y_i, \mathbf{g_i} = \mathrm{LLM}(x_i, \theta) \tag{6}$$

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

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

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 α is a hyperparameter The parameter α is 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 2, 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.

Experiments and Results

Dataset

We evaluate the efficacy of InstructERC on three standard benchmark datasets: IEMOCAP, MELD, and EmoryNLP.

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. Emotional tags in IEMOCAP are *happy, sad, neutral, angry, excited,* and *frustrated.*

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. The emotion classes include *(i.e., happy/joy, anger, fear, disgust, sadness, surprise,* and *neutral)*, while the sentiment classes consist of *positive, negative, or neutral*.

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, namely *neutral, joyful, peaceful, powerful, scared, mad,* and *sad*, while the sentiment classes consist of *positive, negative, or neutral*.

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. The specifics of the datasets are outlined in Table 1.

Baselines

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

Attention-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.

Commonsense-based: **SKAIG** (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.

GNN-based: **GraphCFC** (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.

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

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.

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 pretrained 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.

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 12, 20 and 20 for iemocap, meld and EmoryNLP respectively for all experiments. The retrieval parameter "TopK" is set to Top1 emprically. The hypermeter α 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 $\times$ 80G Nvidia A100 GPUs.

Main Results

Table 2 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

Datasets	Conversations Train	Val	Test	Utterances Train	Val	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

Table 1: The statistics of datasets. avg_utt denotes the average number of utterances in a conversation.

Dataset Models	IEMOCAP W-F1	MELD W-F1	EmoryNLP W-F1	Average W-F1
Discriminant Models				
EmotionIC*	69.50	66.40	***40.01***	***58.63***
SACL*	69.22	***66.45***	39.65	58.44
SKAIG*	66.98	65.18	38.88	57.01
GraphCFC*	68.91	58.86	-	-
UniMSE*	***70.66***	65.51	-	-
Zero-shot + InstructERC				
ChatGLM	***38.6***	***38.8***	19.6	***32.33***
ChatGLM2	21.1	21.8	***24.4***	22.43
Llama	0.753	9.12	5.31	5.06
Llama2	2.774	16.28	8.36	9.46
LoRA + InstructERC				
ChatGLM†	36.04	46.41	30.86	37.77
ChatGLM2†	67.54	65.58	39.09	57.40
Llama†	64.17	67.62	39.34	57.04
Llama2†	**71.39**	**69.15**	**41.37**	**60.64**

Table 2: The main results on three benchmarks.

over the SOTA of discriminative models on all benchmarks. Specifically, we outperform UniMSE, SACL-LSTM, and EmotionIC by 0.73%, 2.70%, and 1.36% on iemocap, meld, and EmoryNLP, respectively. Notably, we completely outperformed multimodal models on two benchmarks using only single-text modality data, 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. 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.

Dataset Models	IEMOCAP W-F1	MELD W-F1	EmoryNLP W-F1	Average W-F1
LoRA + InstructERC				
Llama2	**71.39**	**69.15**	**41.37**	**60.64**
w/o $\mathcal{L}_e$	70.50	68.97	40.78	60.08
w/o $\mathcal{L}_p$	70.70	68.76	40.59	60.01
w/o $\mathcal{L}_e + \mathcal{L}_p$	69.71	68.39	39.56	59.22
w/o $\mathcal{D}_{domain}$	70.91	68.62	40.54	60.02
w/o $_{LoRA}$	70.30	64.80	40.05	58.38

Table 3: The ablation results of Llama2 on three benchmarks.

Ablution study

We conduct an ablation study to investigate the characteristics of the main components in InstructERC. Table 3 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. Missing each part will make the semantic space more chaotic and make the emotion recognition effect worse. 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) 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, please refer to the "All Parameters vs Parameter Efficiency" section.

All Parameters vs Parameter Efficiency

In order to investigate the effect of different parameter finetuning methods on the ERC task, we conducted comparative experiments in Table 4. 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

Dataset Models	IEMOCAP W-F1	MELD W-F1	EmoryNLP W-F1	Average 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**

Table 4: The comparison results of different parameter fine-tuning settings on three benchmarks.

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 2 (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 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.

Scaling Analysis in Low-source Scenario

In this section, we gain an insight into the scaling laws of data and performance for different parameter fine-tuning settings (LoRA & All Parameter), as shown in Figure 4.

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.

Conclusion

In conclusion, our study introduces InstructERC, a novel approach that transforms the ERC task from a discriminative framework to a generative framework using LLMs. Specifically, InstructERC presents a simple and effective retrieval template with emotional domain retrieval module, adapting to different conversation lengths and providing the most similar emotional recognition demonstration for the current utterance. Futhermore, we introduce two additional tasks, speaker identification and emotion prediction, to model speaker and complex conversation relationships, enabling the model to incorporate more fine-grained ERC information. Our plug-in framework with LLMs outperform all previous models and achieve comprehensive state-of-the-art results on three ERC datasets. This demonstrates the broad prospects of LLMs in ERC applications.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant 62236005, 61936004, and U1913602.

References

Brown, T. B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; Agarwal, S.; Herbert-Voss, A.; Krueger, G.; Henighan, T.; Child, R.; Ramesh, A.; Ziegler, D. M.; Wu, J.; Winter, C.; Hesse, C.; Chen, M.; Sigler, E.; Litwin, M.; Gray, S.; Chess, B.; Clark, J.; Berner, C.; McCandlish, S.; Radford, A.; Sutskever, I.; and Amodei, D. 2020. Language Models are Few-Shot Learners. arXiv:2005.14165.

Busso, C.; Bulut, M.; Lee, C.-C.; Kazemzadeh, A.; Mower, E.; 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.

Dong, G.; Guo, D.; Wang, L.; Li, X.; Wang, Z.; Zeng, C.; He, K.; Zhao, J.; Lei, H.; Cui, X.; Huang, Y.; Feng, J.; and Xu,

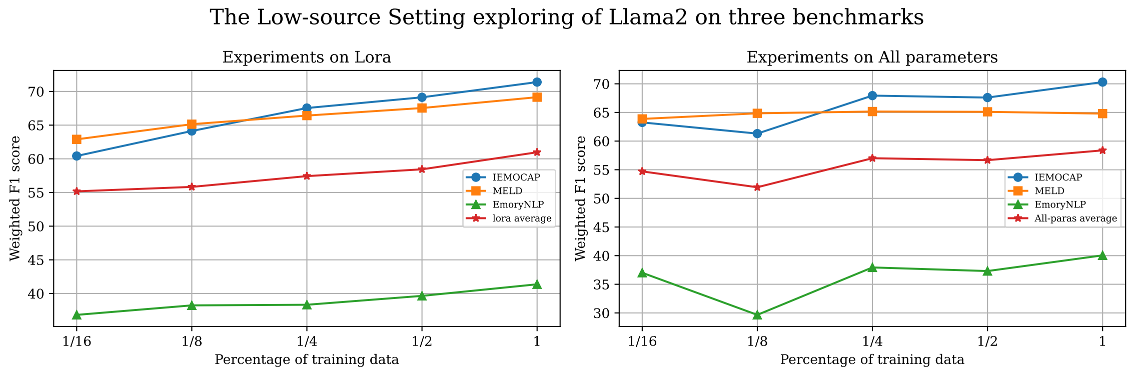

Figure 4: The scaling laws of data and performance for different parameter fine-tuning settings (LoRA & All Parameters)

W. 2022. PSSAT: A Perturbed Semantic Structure Awareness Transferring Method for Perturbation-Robust Slot Filling. In *Proceedings of the 29th International Conference on Computational Linguistics*, 5327–5334. Gyeongju, Republic of Korea: International Committee on Computational Linguistics.

Du, Z.; Qian, Y.; Liu, X.; Ding, M.; Qiu, J.; Yang, Z.; and Tang, J. 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)*, 320–335.

Ghosal, D.; Majumder, N.; Gelbukh, A.; Mihalcea, R.; and Poria, S. 2020. Cosmic: Commonsense knowledge for emotion identification in conversations. *arXiv preprint arXiv:2010.02795*.

Guo, D.; Dong, G.; Fu, D.; Wu, Y.; Zeng, C.; Hui, T.; Wang, L.; Li, X.; Wang, Z.; He, K.; Cui, X.; and Xu, W. 2023. Revisit Out-Of-Vocabulary Problem for Slot Filling: A Unified Contrastive Frameword with Multi-level Data Augmentations. *ArXiv*, abs/2302.13584.

Hu, D.; Bao, Y.; Wei, L.; Zhou, W.; and Hu, S. 2023. Supervised Adversarial Contrastive Learning for Emotion Recognition in Conversations. *arXiv preprint arXiv:2306.01505*.

Hu, D.; Wei, L.; and Huai, X. 2021. DialogueCRN: Contextual Reasoning Networks for Emotion Recognition in Conversations. In *ACL/IJCNLP (1)*, 7042–7052. Association for Computational Linguistics.

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

Li, J.; Lin, Z.; Fu, P.; and Wang, W. 2021. Past, present, and future: Conversational emotion recognition through structural modeling of psychological knowledge. In *Findings of the association for computational linguistics: EMNLP 2021*, 1204–1214.

Li, J.; Wang, X.; Lv, G.; and Zeng, Z. 2023a. Graphcfc: A directed graph based cross-modal feature complementation approach for multimodal conversational emotion recognition. *IEEE Transactions on Multimedia*.

Li, X.; Wang, L.; Dong, G.; He, K.; Zhao, J.; Lei, H.; Liu, J.; and Xu, W. 2023b. Generative Zero-Shot Prompt Learning for Cross-Domain Slot Filling with Inverse Prompting. In *Findings of the Association for Computational Linguistics: ACL 2023*, 825–834. Toronto, Canada: Association for Computational Linguistics.

Liang, C.; Yang, C.; Xu, J.; Huang, J.; Wang, Y.; and Dong, Y. 2021. S+ page: A speaker and position-aware graph neural network model for emotion recognition in conversation. *arXiv preprint arXiv:2112.12389*.

Lin, J.; Men, R.; Yang, A.; Zhou, C.; Zhang, Y.; Wang, P.; Zhou, J.; Tang, J.; and Yang, H. 2021. M6: Multi-Modality-to-Multi-Modality Multitask Mega-Transformer for Unified Pretraining. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery Data Mining*, KDD '21, 3251–3261. New York, NY, USA: Association for Computing Machinery. ISBN 9781450383325.

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 preprint arXiv:1907.11692*.

Lu, Y.; Liu, Q.; Dai, D.; Xiao, X.; Lin, H.; Han, X.; Sun, L.; and Wu, H. 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)*, 5755–5772. Dublin, Ireland: Association for Computational Linguistics.

Ma, Y.; Nguyen, K. L.; Xing, F. Z.; and Cambria, E. 2020. A survey on empathetic dialogue systems. *Information Fusion*, 64: 50–70.

MacKenzie, I. S. 2012. Human-computer interaction: An empirical research perspective.

Minsky, M. 1988. *Society of mind*. Simon and Schuster.

OpenAI. 2023. GPT-4 Technical Report. arXiv:2303.08774.

Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C. L.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; Schulman, J.; Hilton, J.; Kelton, F.; Miller, L.; Simens, M.; Askell, A.; Welinder, P.; Christiano, P.; Leike, J.; and Lowe, R. 2022. Training language models to follow instructions with human feedback. arXiv:2203.02155.

Pontiki, M.; Galanis, D.; Papageorgiou, H.; Androutsopoulos, I.; Manandhar, S.; AL-Smadi, M.; Al-Ayyoub, M.; Zhao, Y.; Qin, B.; De Clercq, O.; et al. 2016. Semeval-2016 task 5: Aspect based sentiment analysis. In *ProWorkshop on Semantic Evaluation (SemEval-2016)*, 19–30. Association for Computational Linguistics.

Poria, S.; Hazarika, D.; Majumder, N.; Naik, G.; Cambria, E.; and Mihalcea, R. 2018. Meld: A multimodal multi-party dataset for emotion recognition in conversations. *arXiv preprint arXiv:1810.02508*.

Qixiang, G.; Dong, G.; Mou, Y.; Wang, L.; Zeng, C.; Guo, D.; Sun, M.; and Xu, W. 2022. Exploiting domain-slot related keywords description for Few-Shot Cross-Domain Dialogue State Tracking. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 2460–2465. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics.

Reimers, N.; and Gurevych, I. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Shen, W.; Wu, S.; Yang, Y.; and Quan, X. 2021. Directed acyclic graph network for conversational emotion recognition. *arXiv preprint arXiv:2105.12907*.

Shen, Y.; Song, K.; Tan, X.; Li, D.; Lu, W.; and Zhuang, Y. 2023. HuggingGPT: Solving AI Tasks with ChatGPT and its Friends in HuggingFace. arXiv:2303.17580.

Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; Rodriguez, A.; Joulin, A.; Grave, E.; and Lample, G. 2023. LLaMA: Open and Efficient Foundation Language Models. arXiv:2302.13971.

Wang, L.; Li, R.; Yan, Y.; Yan, Y.; Wang, S.; Wu, W.; and Xu, W. 2022. InstructionNER: A Multi-Task Instruction-Based Generative Framework for Few-shot NER. arXiv:2203.03903.

Wang, X.; Zhou, W.; Zu, C.; Xia, H.; Chen, T.; Zhang, Y.; Zheng, R.; Ye, J.; Zhang, Q.; Gui, T.; Kang, J.; Yang, J.; Li, S.; and Du, C. 2023. InstructUIE: Multi-task Instruction Tuning for Unified Information Extraction. arXiv:2304.08085.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Ichter, B.; Xia, F.; Chi, E.; Le, Q.; and Zhou, D. 2023. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. arXiv:2201.11903.

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

Yingjian, L.; Jiang, L.; Xiaoping, W.; and Zhigang, Z. 2023. EmotionIC: Emotional Inertia and Contagion-driven Dependency Modelling for Emotion Recognition in Conversation. *arXiv preprint arXiv:2303.11117*.

Yuan, Z.; Yuan, H.; Li, C.; Dong, G.; Tan, C.; and Zhou, C. 2023. Scaling Relationship on Learning Mathematical Reasoning with Large Language Models. arXiv:2308.01825.

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

Zhang, Y.; Wang, S.; Li, P.; Dong, G.; Wang, S.; Xian, Y.; Li, Z.; and Zhang, H. 2023. Pay Attention to Implicit Attribute Values: A Multi-modal Generative Framework for AVE Task. In *Findings of the Association for Computational Linguistics: ACL 2023*, 13139–13151. Toronto, Canada: Association for Computational Linguistics.

Zhao, G.; Dong, G.; Shi, Y.; Yan, H.; Xu, W.; and Li, S. 2022. Entity-level Interaction via Heterogeneous Graph for Multimodal Named Entity Recognition. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, 6345–6350. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics.

A Appendix

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.

A.1 Unified dataset labeling

We continue to use the previous datasets IEMOCAP, MELD, and EmoryNLP. In accordance with The Feeling Wheel (Willcox 1982) proposed in 1982, as shown in Figure 5, we align all emotional labels of three datasets under this standard, the details of which are shown in Tabel 5. After completing the label mapping, there are a total of 9 kinds of emotional labels, which are *joyful, sad, neutral, mad, excited, powerful, fear, peaceful, and disgust*.

A.2 Unified dataset Experiment

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. Meanwhile, we design total mix and ratio mix experiments to explore the impact of different data mixing strategies and data quantities on the model. On below basis, we further explore the impact of data sampling ratio on the model's performance. The details are shown in the Table 5, a more intutive presentation is shown in Figure 6.

The robustness of InstructERC As depicted in the first row of Table 6, upon finetuning InstructERC using the unified dataset, there is a slight decline in the performance of the three benchmarks compared to the SOTA under single dataset training. However, a relatively high Weighted F1-score (W-F1) can still be maintained simultaneously on these three benchmarks, particularly the performance of MELD, which continues to surpass the SOTA level of all small models. Consequently, it is evident that our approach to dataset processing is simple yet efficient. Furthermore, InstructERC, grounded on the Llama2-7B large language model base, exhibits exceptional robustness, capable of concurrently acquiring emotional paradigms from a multitude of distinct distributions, a feat previously unattainable by small models.

The data scaling law Large language models possess formidable learning capabilities, thus validating the data scaling law is a crucial part of our framework. We conduct data scaling experiments 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. This is consistent with the findings of some existing explorations in large models.

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 5: Unified Label Mapping

The Low resource mutual gain We also surprised to discover that during the final stage of training data reduction from 1/32 to 1/64, the Total Mix and Ratio Mix 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 6. 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).

The exploring of different mixing strategies We have further investigated the impact of different mixing strategies on data scaling. Under the Total Mix setting, all datasets are combined for unified sampling. Conversely, under the Ratio Mix setting, datasets are sampled individually before being combined. These two strategies maintain consistency in the number of training data. However, due to the larger absolute quantity of training data from MELD and EmoryNLP, there are more training samples from these two datasets under the Total Mix setting. Therefore, when comparing Total Mix and Ratio Mix, we can observe that IEMOCAP, MELD and EmoryNLP show certain performance biaes due to the varying quantities of their training data.

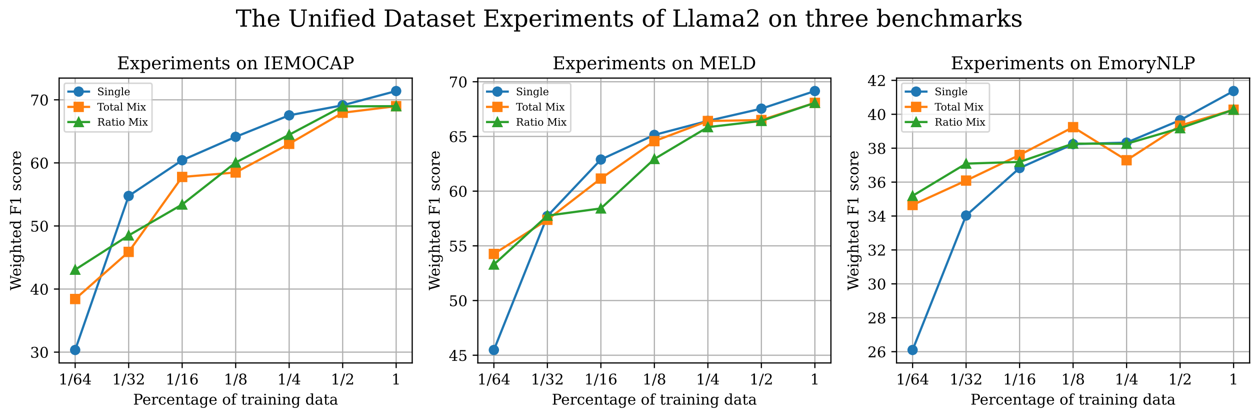

Figure 5: The data scaling law demonstrated on three benchmarks using different data mixing strategies

Data Precent	**IEMOCAP W-F1**			**MELD W-F1**			**EmoryNLP W-F1**		
	Total Mix	Ratio Mix	Single	Total Mix	Ratio Mix	Single	Total Mix	Ratio Mix	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

Table 6: The Unified Dataset Experiments of Llama2 on three benchmarks

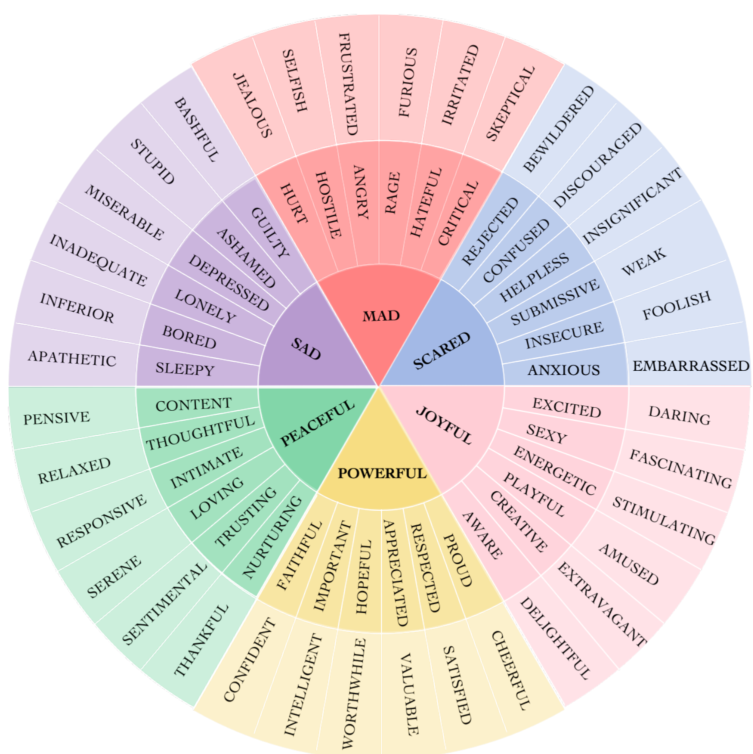

Figure 6: The Feeling Wheel
(Willcox 1982)