Progressive Tuning: Towards Generic Sentiment Abilities for Large Language Models

Guiyang Hou, Yongliang Shen, Weiming Lu[†]
College of Computer Science and Technology, Zhejiang University
{gyhou, luwm}@zju.edu.cn

Abstract

Understanding sentiment is arguably an advanced and important capability of AI agents in the physical world. In previous works, many efforts have been devoted to individual sentiment subtasks, without considering interrelated sentiment knowledge among these subtasks. Although some recent works model multiple sentiment subtasks in a unified manner, they merely simply combine these subtasks without deeply exploring the hierarchical relationships among subtasks. In this paper, we introduce GSA-7B, an open-source large language model specific to the sentiment domain. Specifically, we deeply explore the hierarchical relationships between sentiment subtasks, proposing progressive sentiment reasoning benchmark and progressive task instructions. Subsequently, we use Llama2-7B as the backbone model and propose parameter-efficient progressive tuning paradigm which is implemented by constructing chain of LoRA, resulting in the creation of GSA-7B. Experimental results show that GSA-7B as a unified model performs well across all datasets in the progressive sentiment reasoning benchmark. Additionally, under the few-shot setting, GSA-7B also exhibits good generalization ability for sentiment subtasks and datasets that were not encountered during its training phase.

1 Introduction

Sentiment plays a crucial role in social interaction. Minsky, in his book "Society of Mind" (Minsky, 1988), mentions that equipping machines with the ability to understand the sentiment across different scenarios has consistently been the steadfast goal of researchers. With Large Language Models (LLMs) playing a growing role in our lives, developing LLMs with sentiment intelligence could be better at communicating with us, collaborating with us, and understanding us (Gandhi et al., 2023; Shu et al., 2021).

In prior literature, many efforts (Raffel et al., 2020; Zhao et al., 2022; Li et al., 2022; Hu et al., 2023; Hou et al., 2023; Qiao et al., 2023; Wang et al., 2023b) have been devoted to individual sentiment subtasks, such as Sentiment Analysis (SA), Emotion Recognition in Conversation (ERC), and Sarcasm Detection (SD). However, it has become increasingly clear that there is an interrelated sentiment knowledge among these subtasks. Therefore, integrating all subtasks into a unified model to enhance the sentiment understanding ability of the model has emerged as a significant objective. Although some recent works (Li et al., 2023; Hu et al., 2022; Shah et al., 2023) have emerged to model multiple sentiment subtasks in a unified manner, they merely simply combine various sentiment subtasks without delving into the hierarchical relationships among subtasks. Moreover, there is currently a lack of an open-source LLM with sentiment analysis capabilities that can perform well across multiple sentiment subtasks.

In this paper, we introduce an open-source LLM endowed with sentiment analysis capabilities. The complete process can be divided into three steps: (1) As shown in Figure 1, the task of SA requires reasoning based on the semantic information of sentences. The ERC task, building upon the SA task, further focuses on the contextual information of the target utterance. The SD task, building upon both SA and ERC tasks, additionally pays attention to whether the target utterance expression exhibits linguistic incongruity. We map the SA, ERC, and SD tasks to Level 1-3, respectively, forming a progressive sentiment reasoning benchmark. (2) Consistent with the perspective of progressive sentiment reasoning, we design progressive task instruction for SA, ERC, and SD tasks. (3) We propose parameter-efficient progressive tuning paradigm, which is implemented by constructing chain of LoRA (Hu et al., 2021b; Xia et al., 2024). Within this paradigm, fine-tuning for each

[†] Corresponding author.

Findings of the Association for Computational Linguistics ACL 2024, pages 14392–14402

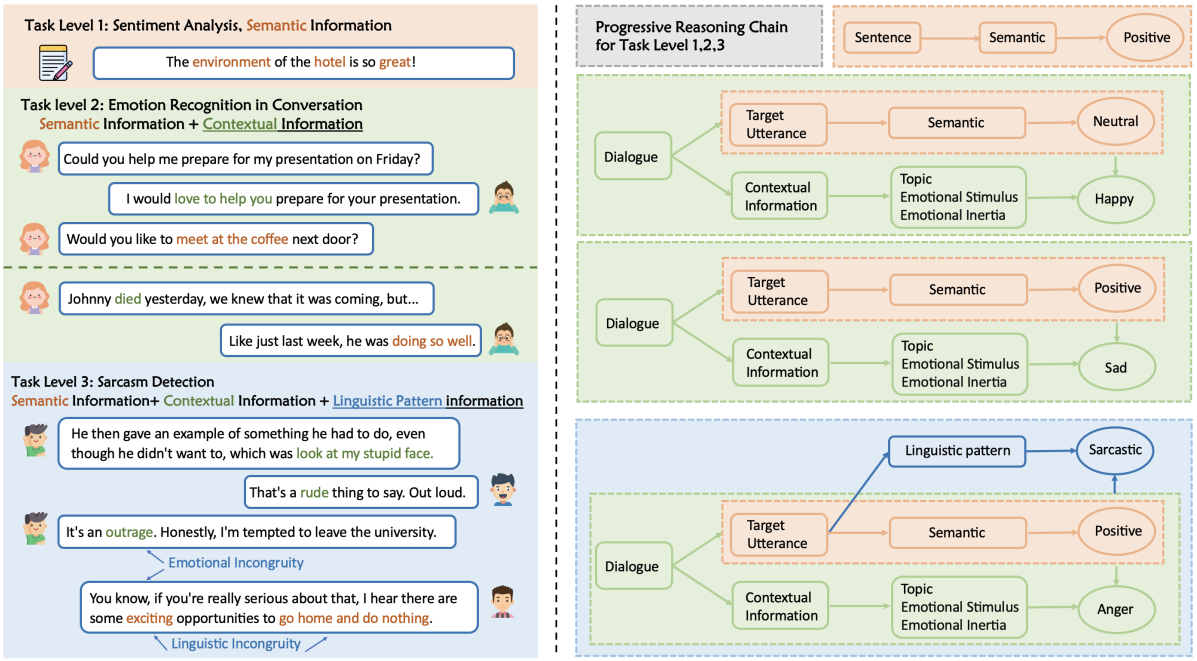

Figure 1: The key information needed for reasoning each level task (such as semantic information for sentiment analysis task) and the corresponding reasoning chain. It is worth mentioning that the reasoning chain of each level of task becomes a part of the reasoning chain for the next level of task.

level of task begins with the fine-tuned model from the previous level, leading to the residual learning between different level tasks. In Section 4.2, we provide a detailed analysis of the multiple advantages of the progressive tuning paradigm, including its effectiveness in mitigating the issue of catastrophic forgetting. We use Llama2-7B (Touvron et al., 2023) as the backbone model, and the fine-tuned model from the final level task is referred to as the result model, named GSA-7B.

The experimental results indicate that GSA-7B as a unified model achieves state-of-the-art (SOTA) results across all task types in the progressive sentiment reasoning benchmark and performs better on half of the datasets compared to SOTA models tailored to individual datasets. Additionally, in the few-shot scenario, GSA-7B demonstrates a commendable ability to generalize across sentiment subtasks and datasets that were not encountered during its training phase.

Our main contributions are summarized as follows:

- We deeply explore the hierarchical relationships between sentiment subtasks, proposing progressive sentiment reasoning benchmark that includes SA (Level 1), ERC (Level 2), and SD (Level 3) tasks.
- Consistent with the perspective of progressive sentiment reasoning, we design progressive task instruction for SA, ERC, and SD tasks.
- We propose progressive tuning paradigm, which is implemented by constructing chain of LoRA, which can effectively mitigate the issue of catastrophic forgetting.
- We introduce GSA-7B, an open-source LLM specific to the sentiment domain. Experimental results show that GSA-7B as a unified model performs well across all datasets in the progressive sentiment reasoning benchmark and exhibits good generalization ability for sentiment subtasks and datasets not seen during training.

2 Background and Related Work

2.1 Low Rank Adaptation

The study by Hu et al. (2021b) focuses on improving the fine-tuning efficiency of LLMs by training considerably smaller low-rank decomposition matrices for specific weights. It posits that weight updates during task adaptation have a low "intrinsic rank" and introduce trainable low-rank decomposition matrices into each layer of the Transformer architecture. Consider a weight matrix $W_{pretrained}$ from the pre-trained model, the weight update ΔW for task adaptation is represented with a low-rank

decomposition BA. The forward pass with LoRA is as follows:

$$W_{pretrained}x + \Delta W x = W_{pretrained}x + BAx \quad (1)$$

where $W_{pretrained}, \Delta W \in \mathbb{R}^{d\times k}$, $A \in \mathbb{R}^{r\times k}$, $B \in \mathbb{R}^{d\times r}$ and $r \ll min(d, k)$. A is typically initialized with random Gaussian initialization and B is initialized with zero to have $\Delta W = 0$ at the beginning of training. During training, $W_{pretrained}$ is frozen and only B, A are optimized. During deployment, the learned low-rank matrices can be merged with the frozen weights of the pre-trained model.

2.2 Multi-task Unified Framework

LLMs as Backbone Text-based LLMs (Brown et al., 2020; Touvron et al., 2023; Chiang et al., 2023; Du et al., 2021) have demonstrated remarkable and even human-level performance in many NLP tasks (Achiam et al., 2023). Meanwhile, instruction tuning (Wei et al., 2021; Chung et al., 2022; Peng et al., 2023), where data is organized as pairs of instruction (or prompt) and response, has emerged as an LLM training paradigm. Building on this foundation, a significant amount of research has been dedicated to developing LLMs for specific domains, including the healthcare (Yang et al., 2023) domain, the education (Lee et al., 2024) domain, the law (Cui et al., 2023) domain, the finance (Zhang and Yang, 2023) domain, and so on.

PLMs as Backbone In the sentiment domain, Yan et al. (2021) employ an improved BART (Lewis et al., 2019) architecture to solve all ABSA subtasks in an end-to-end framework. Hu et al. (2022) propose a multimodal sentiment knowledge-sharing framework that unifies MSA and ERC tasks from features, labels, and models. Shah et al. (2023) retrofit language models produce emotion-aware text representations, applying to multiple sentiment subtasks. Liu et al. (2023) propose a quantum probability framework for joint sarcasm detection and sentiment analysis.

2.3 Sentiment Analysis, Emotion Recognition in Conversation and Sarcasm Detection

Sentiment Analysis The work in this field mainly focuses on fine-tuning pre-trained language models (such as RoBERTa (Liu et al., 2019), XL-Net (Yang et al., 2019), and T5 (Raffel et al., 2020)) on specific datasets, achieving promising results.

Emotion Recognition in Conversation Unlike the basic SA task, ERC is a more practical endeavor that involves predicting the emotion label of each utterance based on the surrounding context. Existing works can be roughly divided into sequence-, graph-, and Transformer-based methods. Sequence-based methods: Hu et al. (2021a) propose a cognitive-inspired network that uses multi-turn reasoning modules to capture implicit emotional clues in conversations. Zhao et al. (2022) utilize GRUs to fuse commonsense knowledge and capture complex interactions in the dialogue. Graph-based methods: Ghosal et al. (2019) treats the dialogue as a directed graph, where each utterance is connected with the surrounding utterances. Ishiwatari et al. (2020) introduces a positional encoding module to simultaneously consider speaker interactions and sequence information. Transformer-based methods: Shen et al. (2021) adopt a modified XLNet to deal with longer context and multi-party structures. Li et al. (2022) utilize supervised contrastive learning and a response generation task to enhance BART's ability for ERC.

Sarcasm Detection The work in this field mainly focuses on capturing incongruity patterns. For example, Pan et al. (2020) utilizes deep neural networks augmented by attention mechanisms for explicitly exploring the contrast and incongruity on word-level or snippet-level. Wang et al. (2023b) introduce iterative incongruity graph structure learning to augment affective dependency graphs for sarcasm detection.

3 Progressive Sentiment Reasoning Benchmark

As shown in Figure 1, different sentiment subtasks form a progressive reasoning chain, leading to a renewed understanding of each sentiment subtask from a reasoning perspective. This section provides a description of the datasets that constitute the sentiment reasoning benchmark. More details can be found in Table 1.

Level 1: Sentiment Analysis SST-2 (Socher et al., 2013), MOSI (Zadeh et al., 2016) and MOSEI (Zadeh et al., 2016) are three widely used datasets for sentiment analysis focusing on sentiment polarity at the sentence (single-turn utterance) level. For the MOSI and MOSEI datasets, we only use their textual modal information.

Task Type	Dataset	Reasoning Level	Modality	Average Length	Total Size	Source
Sentiment Analysis	SST-2	1	T	10	11k	Socher et al. (2013)
Sentiment Analysis	MOSI	1	T	12	2k	Zadeh et al. (2016)
Sentiment Analysis	MOSEI	1	T	20	20k	Zadeh et al. (2018)
Emotion Recognition in Conversation	IEMOCAP	2	T	12	7k	Busso et al. (2008)
Emotion Recognition in Conversation	MELD	2	T	8	13k	Poria et al. (2018)
Emotion Recognition in Conversation	EmoryNLP	2	T	10	12k	Zahiri and Choi (2018)
Sarcasm Detection	MUStARD	3	T	14	0.7k	Castro et al. (2019)
Sarcasm Detection	MUStARD++	3	T	15	1.2k	Ray et al. (2022)

Table 1: Details of our progressive sentiment reasoning benchmark.

Level 2: Emotion Recognition in Conversation IEMOCAP (Busso et al., 2008), MELD (Poria et al., 2018) and EmoryNLP (Zahiri and Choi, 2018) are three widely used datasets for emotion recognition in conversation, which aims to identify emotions conveyed in each utterance within the dialogue context (Poria et al., 2019). For the IEMOCAP and MELD datasets, we only use their textual modal information.

Level 3: Sarcasm Detection MUStARD (Castro et al., 2019) and MUStARD++ (Ray et al., 2022) are two widely used datasets for sarcasm detection in a conversational setting, which involves finding linguistic expression and emotional states incongruity. For the MUStARD and MUStARD++ datasets, we only use their textual modal information.

4 Progressive Tuning

In this section, we first introduce the task instruction used for progressive tuning, followed by a detailed description of the progressive tuning process.

4.1 Progressive Task Instruction

Consistent with the perspective of progressive sentiment reasoning, for basic SA tasks, we design instructions to focus on the semantic information of sentence (single-turn utterance):

Instruction for SA task

```
{sentence}
Consider the semantic information of
sentence to determine its sentiment
polarity.
{label list}
```

For the ERC task, we design instructions to focus on both the semantic information of the target utterance and the contextual information of the dialogue it is situated in, to determine the emotion of the target utterance accurately:

Instruction for ERC task

```
{dialogue}
{target utterance}
Consider both the semantic information of
the target utterance itself and the
contextual information of the dialogue it
is situated in.
{label list}
```

For the SD task, we design instructions to focus on if there's an incongruity between the emotional states conveyed by the target utterance's semantics information and contextual information, and if the target utterance's expression shows any linguistic incongruity:

Instruction for SD task

```
{dialogue}
{target utterance}
Consider whether there is an incongruity
between the emotional states of the target
utterance's semantic information and its
contextual information, and whether the
expression of the target utterance
involves linguistic incongruity.
{label list}
```

Examples of progressive task instruction and a comparison of instruction for different level of tasks are shown in Figure 2 and Table 2.

Instruction	Semantic	Contextual	Linguistic Pattern
Instruction$_{SA}$	✔	✘	✘
Instruction$_{ERC}$	✔	✔	✘
Instruction$_{SD}$	✔	✔	✔

Table 2: Comparison of instruction for different level of tasks.

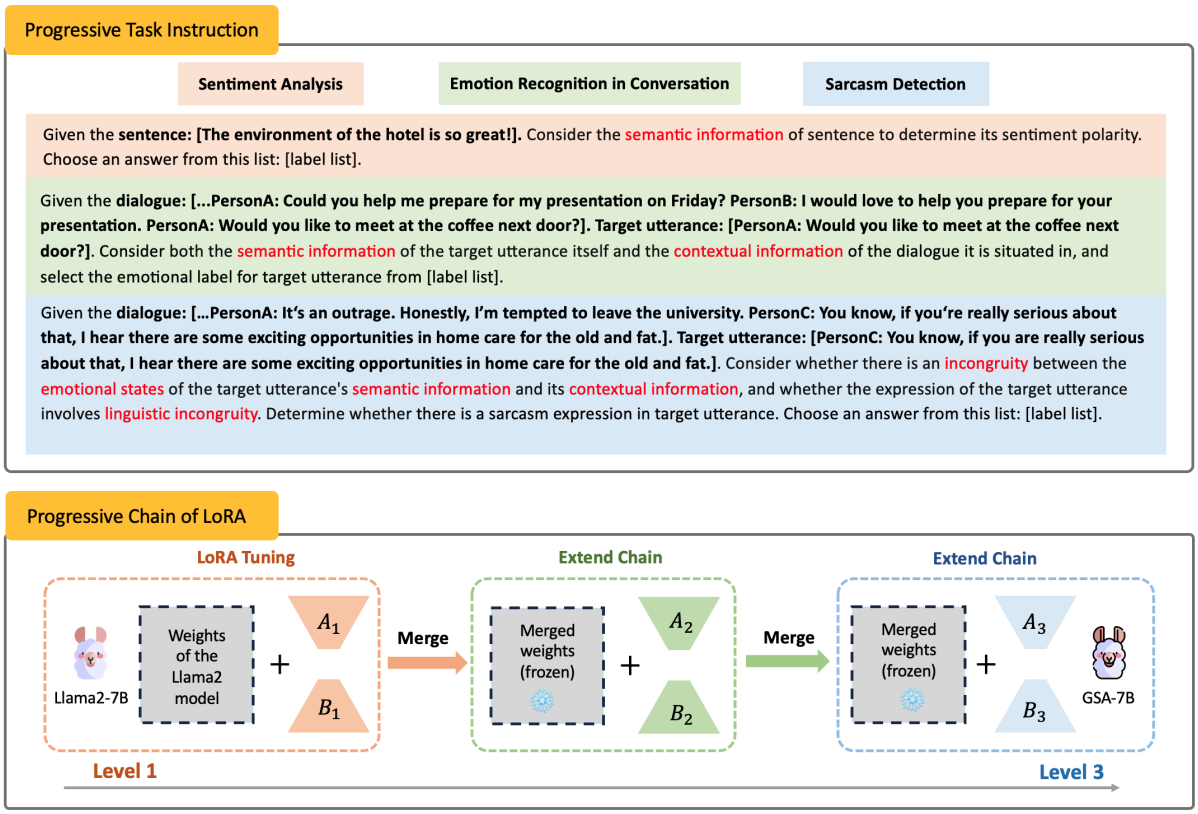

Figure 2: Top: Examples of progressive task instructions. Bottom: Pipeline overview of parameter-efficient progressive tuning, which starts with the frozen Llama2-7B model and consists of three steps:(1) LoRA Tuning, (2) Merge, and (3) Extend the chain. The resulting model is named GSA-7B, a specialized open-source LLM for the sentiment domain.

4.2 Progressive Chain of LoRA

The key idea of our method is to form a chain of LoRA and progressively learn the low-rank adaptation LoRA modules. As illustrated in Figure 2, we first utilize the Instruction$_{SA}$ to fine-tune the Llama2-7B model on the SA task:

$$W_{SA} = W_{pretrained} + \Delta W_{SA} \tag{2}$$

where $W_{pretrained} \in \mathbb{R}^{d\times k}$ represents the pre-trained weights matrix of the Llama2-7B model and $\Delta W_{SA} = B_1 A_1$ represents the weight update occurred during fine-tuning. Then, starting from W_{SA}, we continue to fine-tune on the ERC task using Instruction$_{ERC}$:

$$\begin{aligned} W_{ERC} &= W_{SA} + \Delta W_{ERC} \\ &= W_{pretrained} + \Delta W_{SA} + \Delta W_{ERC} \end{aligned} \tag{3}$$

where $\Delta W_{ERC} = B_2 A_2$. Similarly, starting from W_{ERC}, we continue to fine-tune on the SD task using Instruction$_{SD}$:

$$\begin{aligned} W_{SD} &= W_{ERC} + \Delta W_{SD} \\ &= W_{pretrained} + \Delta W_{SA} + \Delta W_{ERC} + \Delta W_{SD} \end{aligned} \tag{4}$$

where $\Delta W_{SD} = B_3 A_3$. Each low-rank tuple (A_i, B_i) is obtained by optimizing:

$$\arg\min_{B_i A_i} \mathcal{L}\left(W_{pretrained} + \sum_{j=1}^{i} B_j A_j\right) \tag{5}$$

where $\mathcal{L}$ is the task-specific objective function.

Within the progressive tuning paradigm, fine-tuning (A_2, B_2) can be viewed as learning the residual of $W_{ERC} - W_{SA}$, which is not only an easier optimization problem compared to learning W_{ERC} from scratch but also can better coordinate semantic and contextual information for accurate emotion recognition in conversation. Furthermore, we think that the residual of $W_{ERC} - W_{SA}$ represents the learning of contextual information and a broader, more comprehensive understanding of semantic information. Our method ensures that during the fine-tuning process for the ERC (Level 2) task, attention is also given to the semantic information relevant to the SA (Level 1) task, effectively mitigating the problem of catastrophic forgetting. The same analytical perspective is also applicable to fine-tuning (A_3, B_3).

The fine-tuned model from the final level is referred to as the result model, named GSA-7B.

$$W_{GSA} = W_{SD} = W_{pretrained} + \sum_{j=1}^{3} B_j A_j \quad (6)$$

where W_{GSA} represents the weights of GSA-7B model.

5 Experimental Setup

5.1 Baselines

In this section, we report the unified model for sentiment subtasks and current SOTA models for each dataset in the progressive sentiment reasoning benchmark, which we use as baselines to compare with the performance of GSA-7B.

Unified Models KEA(Suresh and Ong, 2021) and RobertaEmo(Shah et al., 2023) retrofit language models produce emotion-aware text representations, applying to multiple sentiment subtasks. TLearn (Shah et al., 2023) is a transfer learning paradigm that initially fine-tunes on the emotion recognition task, followed by further fine-tuning on end tasks using a linear classification head. Quiet(Liu et al., 2023) is a quantum probability framework for joint sarcasm detection and sentiment analysis. UniMSE (Hu et al., 2022) is a sentiment knowledge-sharing framework that unifies sentiment analysis and emotion recognition in conversation tasks from features, labels, and models. UniSA (Li et al., 2023) unifies multiple sentiment subtasks under a single generative framework.

SOTA Models T5-11B (Raffel et al., 2020), serving as a pre-trained model, is fine-tuned on the SST-2 and achieves SOTA results. UniMSE is the current SOTA model for MOSI and MOSEI. InstructERC(Lei et al., 2023) reformulates the ERC task from a discriminative framework to a generative framework based on LLMs. It is the current SOTA model for IEMOCAP, MELD and EmoryNLP. MIL(Zhang et al., 2023) proposes a multi-task learning mechanism to capture correlations and differences across sarcasm detection and sentiment analysis tasks. It is the current SOTA model for MUStARD. Gaze(Tiwari et al., 2023) proposes the utilization of synthetic gaze data to improve the task performance for sarcasm detection. It is the current SOTA model for MUStARD++.

5.2 Evaluation Metrics

We use the same evaluation metrics as employed in the original paper of the dataset. Weighted Accuracy is used for SST-2, 2-category Accuracy is used for MOSI and MOSEI, and F1 scores are used for IEMOCAP, MELD, EmoryNLP, MUStARD and MUStARD++.

5.3 Implementation Details

We develop our approach using the PyTorch framework and the Transformers library (Wolf et al., 2020). We use Llama2-7B (Touvron et al., 2023) as our backbone model. In implementing LoRA, we adhere to the practice outlined in (Hu et al., 2021b), introducing trainable low-rank modules to the self-attention layer. We use a random Gaussian initialization for A_i and set B_i to zero, resulting in the value of $B_i A_i$ being zero. We use Adam optimizer (Kingma and Ba, 2014) to update LoRA parameters and set the learning rate of {8e-5, 2e-4, 1e-4} for progressive tuning of different level tasks. Greedy search is used during inference if not specified.

6 Result and Analysis

6.1 Main Results

We report the test performance of our method and baseline across all datasets in the progressive sentiment reasoning benchmark. The experimental results are shown in Table 3.

Compared to unified models, our method brings average absolute performance improvements of +0.28, +3.34, and +7.43 on SA, ERC, and SD tasks, respectively.

Compared to SOTA models tailored to individual dataset, although our method can be viewed as a unified model, it brings absolute performance improvements of +0.39, +0.64, +2.11, +1.43 on IEMOCAP (Level 2), EmoryNLP (Level 2), MUStARD (Level 3), and MUStARD++ (Level 3) datasets, respectively.

6.2 The Effect of Key Components

In this section, we analyze the knowledge transfer between tasks of different levels in the progressive sentiment reasoning benchmark and the effect of the progressive tuning paradigm. We analyze based on the model's performance on the ERC task.

Knowledge Transfer To explore knowledge transfer between tasks of different levels, we per-

Models	SST-2	MOSI	MOSEI	IEMOCAP	MELD	EmoryNLP	MUStARD	MUStARD++
Unified Models								
KEA(Suresh and Ong, 2021)	93.68	-	-	-	-	-	-	58.28
RobertaEmo(Shah et al., 2023)	94.54	-	-	-	-	-	-	61.40
TLearn(Shah et al., 2023)	93.74	-	-	-	-	-	-	57.34
Quiet(Liu et al., 2023)	-	-	-	-	41.88	-	72.13	-
UniMSE*(Hu et al., 2022)	-	85.85	85.86	70.66	65.51	-	-	-
UniSA*(Li et al., 2023)	90.71	84.11	84.93	64.46	62.22	34.95	-	-
SOTA Models								
T5-11B(Raffel et al., 2020)	97.50	-	-	-	-	-	-	-
UniMSE*(Hu et al., 2022)	-	85.85	85.86	-	-	-	-	-
InstructERC(Lei et al., 2023)	-	-	-	71.39	69.15	41.37	-	-
MIL*(Zhang et al., 2023)	-	-	-	-	-	-	72.64	-
Gaze(Tiwari et al., 2023)	-	-	-	-	-	-	-	72.20
Ours	96.04 ↑1.50	85.46 ↓0.39	85.58 ↓0.28	**71.78** ↑1.12	67.35 ↑1.84	**42.01** ↑7.06	**74.75** ↑2.62	**73.63** ↑12.23
Avg		↑0.28			↑3.34			↑7.43

Table 3: Experimental results on the progressive sentiment reasoning benchmark. $*$ represents this method utilizing multimodal information. We present absolute performance difference between our method and unified models in the baselines. **Bold** indicates that our method brings better performance compared to the SOTA models. The average absolute performance improvement of our model for different level tasks is statistically significant with $p < 0.05$ under one sample t-test.

form LoRA fine-tuning on the Llama2 model using only three datasets from the ERC task. As shown in Table 4, compared to models fine-tuned on tasks across all levels, the models fine-tuned only on ERC tasks exhibited a noticeable average absolute performance decrease.

Progressive Tuning Paradigm Compared to multi-task tuning, progressive tuning has achieved better results across all datasets in the ERC task, as shown in Table 4. This represents that the progressive tuning strategy has established more accurate representations of semantic and contextual information for the target utterance in the ERC task, which also reflects its effective mitigation of the catastrophic forgetting problem to a certain extent.

Methods	IEMOCAP	MELD	EmoryNLP	Avg
Individual Domain				
ERC only	68.63	67.55	40.04	58.74
Different Training Strategies				
Multi-task Tuning	70.45	66.71	41.16	59.44
		Knowledge Transfer		
Progressive Tuning	71.78	67.35	42.01	60.38
		Better Tuning Strategy		

Table 4: Comparison of model performance on the ERC task (Level 2) under different settings.

6.3 Few-shot Generalization

Quantitative Analysis To demonstrate the generalizability of our method in sentiment subtasks: (1) we conduct experiments on hateful detection (Basile et al., 2019) and emoji prediction (Barbieri et al., 2018) tasks, (2) we conduct experiments on sentiment analysis (Rosenthal et al., 2019) and emotion recognition (Mohammad et al., 2018) datasets beyond the progressive sentiment reasoning benchmark. All experiments are conducted under the few-shot setting.

The experimental results in Table 5 demonstrate that our method, under the few-shot setting, performs well on sentiment subtasks and datasets not included in the progressive sentiment reasoning benchmark. This result reveals that our method has learned general sentiment knowledge during the training phase.

Methods	Hateval	Twitter emoji	Twitter emotion	Twitter sentiment
SOTA_{train}	65.10	32.20	76.10	72.07
$\text{T5}_{few-shot}$	47.94	7.58	59.56	66.13
$\text{UniSA}_{few-shot}$	49.61	14.82	65.98	64.17
$\text{Ours}_{few-shot}$	57.65	20.10	68.82	68.03
$\text{Ratio}_{few-shot/train}$	0.03	0.07	0.18	0.01

Table 5: Generalization results of our method on sentiment subtasks. The SOTA models learn on the entire training set. The ratio represents the proportion of the number of samples in the few-shot setting to the number of samples in the entire training set.

Qualitative Analysis Rooting in the nature of LoRA, Wang et al. (2023a) propose an insight that

Task Type	Sentence/Target Utterance	Context	Model 1	Model 2
SA	I mean I don't regret seeing it.	-	GSA-7B (Positive ✔)	SA-7B (Non-Positive ✘)
SD	Person A: It has a **bad** memory but a **great** battery life.	Person A: I just got a new phone. Person B: How is it?	GSA-7B (Non-sarcastic ✔)	Multi-Task Tuning (Sarcastic ✘)

Table 6: Case studies show that the GSA-7B model provides correct answers for SA and SD tasks, highlighting its advantages effective mitigation of disaster amnesia and its advantages in effectively mitigating the problem of catastrophic forgetting and possessing better cognitive patterns. SA-7B represents the model fine-tuned on the SA task.

LoRA parameters are not mere numerical adjustments but encapsulate crucial directions for model gradient updates. Under the progressive tuning paradigm, fine-tuning for each level task starts with the model fine-tuned on the previous level task, learning the residuals between different level tasks. Furthermore, as illustrated in Figure 3, considering that various sentiment subtasks share similar gradient directions, when faced with a new sentiment subtask, the model can quickly and easily learn the patterns of the new task, thereby achieving competitive performance even in a few-shot setting.

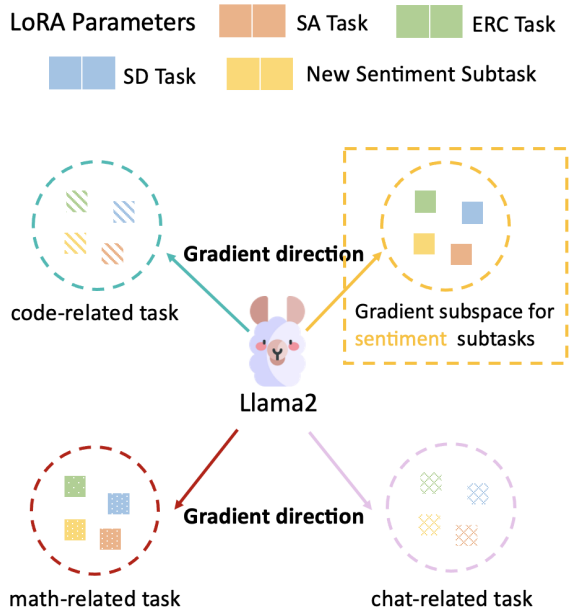

Figure 3: Qualitative analysis of the generalization of our method on sentiment subtasks.

6.4 Case Study

In Table 6, we exemplify two cases from the SA and SD tasks.

In the first case, we employ the model fine-tuned on the SA task to determine that the sentiment polarity of the sentence "I mean I don't regret seeing it." as incorrect, whereas the GSA-7B model provides the correct answer. We believe that this is due to during the fine-tuning process for the ERC (Level 2) and SD (Level 3) task, attention is also given to the semantic information relevant to the SA (Level 1) task, leading to a broader, more comprehensive understanding of semantic information, effectively mitigating the problem of catastrophic forgetting.

In the second case, we compare the result of multi-task tuning and the GSA-7B model on an instance of SD task. Due to the presence of both 'good' and 'bad' words in the target utterance, the multi-task tuning model classified it as sarcastic. For the GSA-7B model obtained by our progressive fine-tuning, it not only focuses on the linguistic expression of the target utterance but also takes into full consideration the emotional state corresponding to the contextual and semantic information, possessing a better cognitive pattern.

7 Conclusion

In this paper, we deeply explore the hierarchical relationships between sentiment subtasks, proposing progressive reasoning benchmark that includes SA (Level 1), ERC (Level 2), and SD (Level 3) tasks and progressive task instruction. Subsequently, we propose parameter-efficient progressive tuning paradigm to finetune the Llama2-7B model. The resulting model is named GSA-7B, a specialized open-source LLM for the sentiment domain. Experimental results show that GSA-7B performs well across all datasets in the progressive sentiment reasoning benchmark and exhibits good generalization ability for sentiment subtasks and datasets that were not encountered during its training phase, which reveals the feasibility of building AI that can understand sentiment.

Limitations

There are two major limitations in this study. Firstly, there are relatively few task types in the progressive sentiment reasoning benchmark, which can be further enriched by conducting more fine-grained and in-depth exploration of tasks in the sentiment domain. For example, incorporating generative tasks such as empathetic responses into the benchmark. Secondly, in constructing the progressive sentiment reasoning benchmark, we only utilize the information from the textual modality in each dataset. Towards sentiment analysis capabilities for Multimodal LLM is also important, which we treat as future work.

Acknowledgements

This work is supported by the National Natural Science Foundation of China (No. 62376245), the National Natural Science Foundation of Zhejiang Province (LY24C090001) and the Fundamental Research Funds for the Central Universities.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Francesco Barbieri, Jose Camacho-Collados, Francesco Ronzano, Luis Espinosa Anke, Miguel Ballesteros, Valerio Basile, Viviana Patti, and Horacio Saggion. 2018. Semeval 2018 task 2: Multilingual emoji prediction. In *Proceedings of the 12th international workshop on semantic evaluation*, pages 24–33.

Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. 2019. Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter. In *Proceedings of the 13th international workshop on semantic evaluation*, pages 54–63.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

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.

Santiago Castro, Devamanyu Hazarika, Verónica Pérez-Rosas, Roger Zimmermann, Rada Mihalcea, and Soujanya Poria. 2019. Towards multimodal sarcasm detection (an _obviously_ perfect paper). *arXiv preprint arXiv:1906.01815*.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*.

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*.

Jiaxi Cui, Zongjian Li, Yang Yan, Bohua Chen, and Li Yuan. 2023. Chatlaw: Open-source legal large language model with integrated external knowledge bases. *arXiv preprint arXiv:2306.16092*.

Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. 2021. Glm: General language model pretraining with autoregressive blank infilling. *arXiv preprint arXiv:2103.10360*.

Kanishk Gandhi, Jan-Philipp Fränken, Tobias Gerstenberg, and Noah D Goodman. 2023. Understanding social reasoning in language models with language models. *arXiv preprint arXiv:2306.15448*.

Deepanway Ghosal, Navonil Majumder, Soujanya Poria, Niyati Chhaya, and Alexander Gelbukh. 2019. Dialoguegcn: A graph convolutional neural network for emotion recognition in conversation. *arXiv preprint arXiv:1908.11540*.

Guiyang Hou, Yongliang Shen, Wenqi Zhang, Wei Xue, and Weiming Lu. 2023. Enhancing emotion recognition in conversation via multi-view feature alignment and memorization. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12651–12663.

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*.

Dou Hu, Lingwei Wei, and Xiaoyong Huai. 2021a. Dialoguecrn: Contextual reasoning networks for emotion recognition in conversations. *arXiv preprint arXiv:2106.01978*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021b. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Guimin Hu, Ting-En Lin, Yi Zhao, Guangming Lu, Yuchuan Wu, and Yongbin Li. 2022. Unimse: Towards unified multimodal sentiment analysis and emotion recognition. *arXiv preprint arXiv:2211.11256*.

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.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Unggi Lee, Minji Jeon, Yunseo Lee, Gyuri Byun, Yoorim Son, Jaeyoon Shin, Hongkyu Ko, and Hyeoncheol Kim. 2024. Llava-docent: Instruction tuning with multimodal large language model to support art appreciation education. *arXiv preprint arXiv:2402.06264*.

Shanglin Lei, Guanting Dong, Xiaoping Wang, Keheng Wang, and Sirui Wang. 2023. Instructerc: Reforming emotion recognition in conversation with a retrieval multi-task llms framework. *arXiv preprint arXiv:2309.11911*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Shimin Li, Hang Yan, and Xipeng Qiu. 2022. Contrast and generation make bart a good dialogue emotion recognizer. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 11002–11010.

Zaijing Li, Ting-En Lin, Yuchuan Wu, Meng Liu, Fengxiao Tang, Ming Zhao, and Yongbin Li. 2023. Unisa: Unified generative framework for sentiment analysis. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 6132–6142.

Yaochen Liu, Yazhou Zhang, and Dawei Song. 2023. A quantum probability driven framework for joint multi-modal sarcasm, sentiment and emotion analysis. *IEEE Transactions on Affective Computing*.

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*.

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

Saif Mohammad, Felipe Bravo-Marquez, Mohammad Salameh, and Svetlana Kiritchenko. 2018. Semeval-2018 task 1: Affect in tweets. In *Proceedings of the 12th international workshop on semantic evaluation*, pages 1–17.

Hongliang Pan, Zheng Lin, Peng Fu, and Weiping Wang. 2020. Modeling the incongruity between sentence snippets for sarcasm detection. In *ECAI 2020*, pages 2132–2139. IOS Press.

Baolin Peng, Chunyuan Li, Pengcheng He, Michel Galley, and Jianfeng Gao. 2023. Instruction tuning with gpt-4. *arXiv preprint arXiv:2304.03277*.

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*.

Soujanya Poria, Navonil Majumder, Rada Mihalcea, and Eduard Hovy. 2019. Emotion recognition in conversation: Research challenges, datasets, and recent advances. *IEEE Access*, 7:100943–100953.

Yang Qiao, Liqiang Jing, Xuemeng Song, Xiaolin Chen, Lei Zhu, and Liqiang Nie. 2023. Mutual-enhanced incongruity learning network for multi-modal sarcasm detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 9507–9515.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Anupama Ray, Shubham Mishra, Apoorva Nunna, and Pushpak Bhattacharyya. 2022. A multimodal corpus for emotion recognition in sarcasm. *arXiv preprint arXiv:2206.02119*.

Sara Rosenthal, Noura Farra, and Preslav Nakov. 2019. Semeval-2017 task 4: Sentiment analysis in twitter. *arXiv preprint arXiv:1912.00741*.

Sapan Shah, Sreedhar Reddy, and Pushpak Bhattacharyya. 2023. Retrofitting light-weight language models for emotions using supervised contrastive learning. *arXiv preprint arXiv:2310.18930*.

Weizhou Shen, Junqing Chen, Xiaojun Quan, and Zhixian Xie. 2021. Dialogxl: All-in-one xlnet for multi-party conversation emotion recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 13789–13797.

Tianmin Shu, Abhishek Bhandwaldar, Chuang Gan, Kevin Smith, Shari Liu, Dan Gutfreund, Elizabeth Spelke, Joshua Tenenbaum, and Tomer Ullman. 2021. Agent: A benchmark for core psychological reasoning. In *International Conference on Machine Learning*, pages 9614–9625. PMLR.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. 2013. Recursive deep models for

semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1631–1642.

Varsha Suresh and Desmond C Ong. 2021. Using knowledge-embedded attention to augment pretrained language models for fine-grained emotion recognition. In *2021 9th International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 1–8. IEEE.

Divyank Tiwari, Diptesh Kanojia, Anupama Ray, Apoorva Nunna, and Pushpak Bhattacharyya. 2023. Predict and use: Harnessing predicted gaze to improve multimodal sarcasm detection. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15933–15948.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Xiao Wang, Tianze Chen, Qiming Ge, Han Xia, Rong Bao, Rui Zheng, Qi Zhang, Tao Gui, and Xuanjing Huang. 2023a. Orthogonal subspace learning for language model continual learning. *arXiv preprint arXiv:2310.14152*.

Xiaobao Wang, Yiqi Dong, Di Jin, Yawen Li, Longbiao Wang, and Jianwu Dang. 2023b. Augmenting affective dependency graph via iterative incongruity graph learning for sarcasm detection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pages 4702–4710.

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*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pages 38–45.

Wenhan Xia, Chengwei Qin, and Elad Hazan. 2024. Chain of lora: Efficient fine-tuning of language models via residual learning. *arXiv preprint arXiv:2401.04151*.

Hang Yan, Junqi Dai, Xipeng Qiu, Zheng Zhang, et al. 2021. A unified generative framework for aspect-based sentiment analysis. *arXiv preprint arXiv:2106.04300*.

Songhua Yang, Hanjia Zhao, Senbin Zhu, Guangyu Zhou, Hongfei Xu, Yuxiang Jia, and Hongying Zan. 2023. Zhongjing: Enhancing the chinese medical capabilities of large language model through expert feedback and real-world multi-turn dialogue. *arXiv preprint arXiv:2308.03549*.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32.

Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. 2016. Mosi: multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos. *arXiv preprint arXiv:1606.06259*.

AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2018. Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2236–2246.

Sayyed M Zahiri and Jinho D Choi. 2018. Emotion detection on tv show transcripts with sequence-based convolutional neural networks. In *Workshops at the thirty-second aaai conference on artificial intelligence*.

Xuanyu Zhang and Qing Yang. 2023. Xuanyuan 2.0: A large chinese financial chat model with hundreds of billions parameters. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pages 4435–4439.

Yazhou Zhang, Yang Yu, Dongming Zhao, Zuhe Li, Bo Wang, Yuexian Hou, Prayag Tiwari, and Jing Qin. 2023. Learning multi-task commonness and uniqueness for multi-modal sarcasm detection and sentiment analysis in conversation. *IEEE Transactions on Artificial Intelligence*.

Weixiang Zhao, Yanyan Zhao, and Xin Lu. 2022. Cauain: Causal aware interaction network for emotion recognition in conversations. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI*, pages 4524–4530.