Chinchunmei at WASSA 2024 Empathy and Personality Shared Task: Boosting LLM's Prediction with Role-play Augmentation and Contrastive Reasoning Calibration

Tian Li[1,2] **Nicolay Rusnachenko**[2] **Huizhi Liang**[2]
[1]Shumei AI Research Institute, Beijing, China
[2]Newcastle University, Newcastle Upon Tyne, England
{litianricardolee,rusnicolay}@gmail.com
Huizhi.Liang@newcastle.ac.uk

Abstract

This paper presents the Chinchunmei team's contributions to the WASSA2024 Shared-Task 1: Empathy Detection and Emotion Classification. We participated in Tracks 1, 2, and 3 to predict empathetic scores based on dialogue, article, and essay content. We choose Llama3-8b-instruct as our base model. We developed three supervised fine-tuning schemes: standard prediction, role-play, and contrastive prediction, along with an innovative scoring calibration method called Contrastive Reasoning Calibration during inference. Pearson Correlation was used as the evaluation metric across all tracks. For Track 1, we achieved 0.43 on the devset and 0.17 on the testset. For Track 2 emotion, empathy, and polarity labels, we obtained 0.64, 0.66, and 0.79 on the devset and 0.61, 0.68, and 0.58 on the testset. For Track 3 empathy and distress labels, we got 0.64 and 0.56 on the devset and 0.33 and 0.35 on the testset.

1 Introduction

Empathy refers to the ability to understand and share the feelings or experiences of others. It involves identifying, comprehending, and sharing with others' emotions, thoughts, motivations, and personality traits (Bellet and Maloney, 1991; Hall et al., 2021). As one of the essential human qualities, empathy plays an essential role not only in various academic fields such as healthcare(Decety and Fotopoulou, 2015), neuroscience(Singer and Lamm, 2009), psychology, and philosophy (Yan and Tan, 2014) but also in everyday interactions. Since empathy expression depends on human reaction and its assessment often requires nuance analysis of various features—such as underlying meanings, references, and emotional release—identifying empathy in diverse scenarios has always been a hot research topic.

For the reasons above, WASSA 2024 (Giorgi et al., 2024; Barriere et al., 2023; Omitaomu et al., 2022) has once again hosted the Empathy Detection and Emotion Classification shared task. This year's contest introduces multi-level and multimodal data, which comprises news articles, essays, and dialogs. It abandons simple classification labels in favor of a scoring system where different scores carry actual meaning. All tracks use Pearson Correlation as the evaluation metric. These factors collectively render this competition exceptionally challenging. In this competition, we participated in tracks 1, 2, and 3, all related to empathy detection.

During previous contests, most participants chose the encoder framework (Chen et al., 2022; Li et al., 2022b; Vasava et al., 2022; Li et al., 2022a; Meshgi et al., 2022). In this paper, to unify the diverse modalities and multiple labels across different tracks into a single model, we used the generative large language model (LLM) framework. However, as the training objective of LLM is the next token prediction, it can hardly carry on the discriminative training purpose. With limited samples and imbalanced label distributions, sometimes the model can only learn templated outputs rather than the logic behind the scoring. These issues are particularly severe in Track 1. To address these, we introduced various task templates to enrich the train set and incorporated the concepts of contrastive learning (Rethmeier and Augenstein, 2023; Sun et al., 2023; Li et al., 2023; Gao and Das, 2024) and contrastive chain-of-thought (Chu et al., 2023; Chia et al., 2023) to enhance the distinctiveness and reliability of the model's scoring. Additionally, our approach does not involve any external data. This further proves that the superiority of our solution stems from the technical approach itself and can be easily transferred to other similar tasks.

Our contributions are as follows:

- We introduced a role-playing template to enrich the training samples. By training the model to generate responses for a given role

Proceedings of the 14th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis, pages 385–392

based on articles, preceding dialogue history, and the provided empathy, emotion intensity, and emotion polarity scores, we aimed to help the model capture the nuanced characteristics related to empathy in different expressions. Our experiments demonstrated significant improvements in Track 3 with this approach.

- We developed contrastive supervised fine-tune (C-SFT) and contrastive reasoning calibration (CRC) techniques for more reliable scoring generation. C-SFT not only enhances model performance but also mitigates data scarcity by creating contrastive pairs. CRC leverages chain-of-thought (COT) during inference to refine predictions, further enhancing the final performance. Our experiments showed notable improvements in both tracks 1 and 2 with these techniques.

2 Methodology

Our approach, illustrated in Fig 1, consists of two stages: the SFT stage and the inference stage. In the SFT stage, we enrich the training samples by introducing three templates: the standard prediction template, the role-play template, and the contrastive template. These are detailed in section 2.1.1, 2.1.2, and 2.1.3. In the inference stage, in addition to using the standard prediction template, we employ the contrastive template. This forces the model to compare scores of a specific label between two data points, thereby refining the prediction results. This is elaborated in section 2.2.1.

2.1 SFT Stage

2.1.1 Standard Prediction Template

In this task, the LLM performs score predictions using the corresponding standard templates. According to the input length of Llama3-8b-instruct, we concatenate the article and task content together as input and train the model to predict all tracks' results. The template is shown in B.1.

2.1.2 Role-play Template

Since parts of the data come from dialogues, performing the role-play fine-tuning based on dialogues enhances the model's perception between the roles and the empathic expression features, thereby strengthening the model's empathy detection result. Based on this assumption, we trained the model to generate the current text based on the dialogue history and the labels from Track 2. The template is shown in B.2.

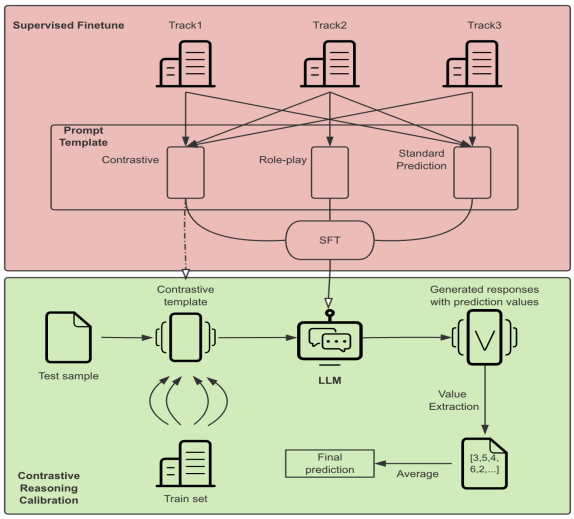

Figure 1: The overall flowchart of our method. It is divided into two stages: the red part represents the SFT stage, and the green part represents the inference stage.

2.1.3 Contrastive Supervised Templates

One of the challenges of tracks 1, 2, and 3 is that their labels are comparable values rather than isolated labels. Treating it as a traditional classification task is inappropriate, as traditional classification tasks regard all misclassifications equally. Additionally, another challenge lies in data scarcity, as tracks 1 and 3 only have 1,000 training samples.

To address these two issues, we develop a novel C-SFT approach that uses contrastive pairs to fine-tune the LLM. This not only handles the magnitude discriminative training but also solves the data scarcity problem. By randomly sampling two pieces of data to form contrastive pairs, we can:

1. Enable the model to understand that label values are comparable rather than isolated by comparing the two samples' predictions.

2. Construct a vast amount of training samples through pairwise combinations.

However, this introduces three new issues. First, it doubles the input length. Second, if the two samples in a pair have identical scores, the discriminative training will fail. Third, an excessively large training set can be a burden for training. Therefore, we discard the article content and only sample 5000 pairs for each task and each label, prioritizing data with differing scores. Taking Track 1 as an example, we retain all pairs with score differences in the range $[2, +\inf]$, keep pairs with a score difference of 1 with a probability of 30%, and only

retain pairs with identical scores with a probability of 0.1%. After sampling 9,000+ times, we obtain 5,000 contrastive pairs as training samples. Similar sampling strategies are adopted for tracks 2 and 3. The templates are list in B.3.

Furthermore, due to the presence of two speakers in a conversation, for Track 1 we also constructed contrastive pairs for these two speakers. Although this dataset is limited to fewer than 1000 pairs due to the number of dialogues, it further enriches the diversity of the training set. The template is also shown in B.3.

2.2 Inference Stage

2.2.1 Contrastive Reasoning Calibration

After completing the SFT, we continue using the C-SFT templates for CRC prediction. Compared to standard prediction, predictions based on the C-SFT template are influenced by their contrastive samples. This is because before outputting the final prediction, the model first compares the two samples on a given label and then outputs the final results for both samples. This employs the COT feature, making the model's output more reliable. The algorithm is shown in Algorithm 1.

Algorithm 1 Contrastive Reasoning Calibration

Require: Test sample $\mathbf{X} = \{x_1, x_2, ..., x_n\}$, train sample $\mathbf{S} = \{s_1, s_2, ..., s_m\}$, label value $\mathbf{V} = \{v_1, v_2, ..., v_o\}$, contrastive template $\mathbf{T}$
Ensure: Prediction $\mathbf{P} = \{p_1, p_2, ..., p_n\}$

$tempP = []$
for x in $\mathbf{X}$ **do**
  for v in $\mathbf{V}$ **do**
    Sample i pieces of s with label value v, $\hat{\mathbf{S}} = \{\hat{s}_1, \hat{s}_2, ..., \hat{s}_i\}$
    for $\hat{s}$ in $\hat{\mathbf{S}}$ **do**
      Randomly apply $(x, \hat{s})$ or $(\hat{s}, x)$ to $\mathbf{T}$
      Get results $(v_x, v_{\hat{s}})$ or $(v_{\hat{s}}, v_x)$
      $tempP \leftarrow v_x$
    end for
  end for
  $p_x = AVG(tempP)$
end for

We select contrastive data from the train set because the predictions on the training set are very accurate, making them ideal benchmarks for comparison. The choice of i is constrained by our inference resources. For Track 1, i is 4. For Track 2 and 3, i is 1 to meet the competition deadline because of the limit of our computational resources. After the competition deadline, we continued experimenting with the i set to 4 on the Track 2 dev set. The results are presented in the C

3 Experiment

This section introduces the train set statistics, the base model selection, and the fine-tuning settings.

3.1 Dataset-Sample Statistics

Tracks 1 and 2 use the same dialogue data. It includes 487 dialogues corresponding to 100 articles. Each dialogue involves two speakers, with a total of 75 participants. The text length per dialogue turn ranges from 1 to 701 characters. The turn number per dialogue ranges from 13 to 44. The overall dialogue length varies from 601 to 6701 characters.

For Track 2, as it involves predicting at each dialogue turn, the actual input can be in two modes: 1). Single turn mode, referred to as Track2-single-turn. 2). Multi-turn mode with context, referred to as Track2-multi-turn.

Track 3 samples consist of individual essays. The text length ranges from 300 to 800 characters.

The article data includes the title, source, object of suffering, and content. The content length ranges from 176 to 31,784 characters.

3.2 Dataset-Label Statistics

As for Track 1, this task includes only empathy scores ranging from 1 to 9. Over half of the scores are 7, indicating a highly imbalanced issue.

As for Track 2, this task includes three types of scores: emotion intensity, empathy, and emotion polarity. The emotion and empathy scores range from 0 to 5, all values being multiples of 1/3. The emotional polarity scores range from 0 to 2.6667, with a total of 10 distinct values.

As for Track 3, this task includes two types of scores: empathy and distress. Each score ranges from 0 to 7, all values being multiples of 1/7.

Since the labels contain floating-point numbers, to prevent negative impact on LLM encoding, we mapped all label values to an integer domain starting from 0. Even though the results are evaluated using Pearson correlation, this mapping does not negatively affect the performance evaluation.

3.3 Selection of Base Model

We compare the suitability of Llama2-7b-chat[1] and Llama3-8b-instruct[2] for all three tracks and finally choose the later one. This is because: 1). Llama3-8b-instruct has up to 8192 input length that can

[1]https://huggingface.co/meta-llama/Llama-2-7b-chat-hf
[2]https://huggingface.co/meta-llama/Meta-Llama-3-8B-Instruct

Table 1: The performance comparison of different SFT types combinations on devset

Pearson	Track 1	Track 2 single turn				Track 2 multi turn				Track 3		
Correlation	Empathy	Emotion	Empathy	Polarity	AVG	Emotion	Empathy	Polarity	AVG	Empathy	Distress	AVG
Baseline	-0.037	**0.634**	0.576	**0.733**	**0.648**	**0.637**	0.624	0.745	0.668	0.563	0.448	0.505
+Role-play	0.127	0.628	0.580	0.700	0.636	0.624	0.638	0.738	0.667	**0.639**	**0.559**	**0.599**
+Role-play +C-SFT	**0.270**	0.618	**0.587**	**0.733**	0.646	0.625	**0.636**	**0.747**	**0.669**	0.542	0.336	0.439

accept article content. 2). With the standard prediction template, Llama3-8b-instruct outperforms Llama2-7B-chat. Therefore, all subsequent results are based on the tuning of Llama3-8B-Instruct.

3.4 Training Configuration

Due to resource limits, all training processes use LoRA technique (Hu et al., 2021). The rank is 8, alpha is 16, and dropout is 0.

The epoch number is 3, and the learning rate (LR) is 2e-4. The LR scheduler employs the cosine strategy with 0.1 warmup ratio and 128 batch size.

4 Result and Analysis

4.1 Baseline

The baseline model is obtained by standard prediction fine-tuning. The devset results are in Table 1. Notably, the result of Track 1 is quite poor, probably due to the severe imbalance in sample labels.

4.2 Standard Prediction + Role Play

After incorporating the role-play template, tracks 1 and 3 on devset show significant improvement, proving the effectiveness of the role-play fine-tuning. However, according to Table 1, the Track 2 results fluctuate among all labels, making it difficult to distinguish any clear benefit. This is because the role-play and the Track2-multi-turn standard prediction are similar tasks, with the content and prediction swapped.

4.3 Standard Prediction + Role Play + Contrastive Tuning

According to Table 1, this approach further improves the devset results of Track 1 and Track2-multi-turn. However, for Track2-single-turn, there is a noticeable decline in emotion intensity. After analyzing the cases, we find that the dataset contains instances of similar data with different scores, causing confusion to the model and leading to tuning failures. We suspect that Track 2 labeling was based on the current and historical dialogue turns, resulting in similar texts with different labels. To save inference costs, we discard the Track2-single-turn in subsequent experiments.

Besides, the Track 3 devset results also declined. Our analysis suggests that this decline is due to Track 3's complex labeling system and its dependence on article content. Track 3 has up to 43 values for each label, greatly increasing the learning difficulty for the model. Moreover, skipping the article content may lost key semantic features. Thus, we submit our Track 3 testset results based on section 4.2 method and drop the Track 3 task in future experiments to reduce costs.

4.4 Contrastive Reasoning Calibration

To demonstrate that the success of our approach is not coincidental, we prepare an extra model using LR 8e-4 and apply the CRC method to both for comparison. The devset results in Table 2 show significant improvements across all labels for both models. This validates the effectiveness and robustness of our approach. Thus, for Track 1, we submit the testset results obtained from the model using LR 8e-4 with CRC. For Track 2, we submit the testset results from the model with LR 2e-4 using the same method.

Table 2: The performance comparison between standard prediction and CRC on devset

Pearson	Track 1	Track 2 multi turn			
Correlation	Empathy	Emotion	Empathy	Polarity	AVG
LR: 2e-5	0.270	0.625	0.636	0.747	0.669
+CRC	0.395	0.641	**0.664**	**0.790**	**0.698**
LR: 8e-5	0.360	0.632	0.622	0.729	0.661
+CRC	**0.434**	**0.662**	0.645	0.773	0.693

5 Conclusion

Our experiments demonstrate the significant potential of the current LLM in empathy detection. Firstly, with a modest amount of SFT data preparation, we successfully created an 8B scale strong baseline LLM. Secondly, we improve it by introducing the Role-play and C-SFT tasks. Thirdly, we further enhance the performance using contrastive reasoning to refine scoring outputs. Finally, our solution secured third place in Track 1 and second place in both tracks 2 and 3. The testset results are shown in A. Since our techniques require no external data, they can be widely applied to similar classification or scoring tasks.

Regarding the performance decline for Track 3 after using C-SFT, we plan to conduct further investigations with long-context LLM to validate if the decline is due to the lack of article information.

References

Valentin Barriere, João Sedoc, Shabnam Tafreshi, and Salvatore Giorgi. 2023. Findings of wassa 2023 shared task on empathy, emotion and personality detection in conversation and reactions to news articles. In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 511–525.

Paul S Bellet and Michael J Maloney. 1991. The importance of empathy as an interviewing skill in medicine. *Jama*, 266(13):1831–1832.

Yue Chen, Yingnan Ju, and Sandra Kübler. 2022. IUCL at WASSA 2022 shared task: A text-only approach to empathy and emotion detection. In *Proceedings of the 12th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis*, pages 228–232, Dublin, Ireland. Association for Computational Linguistics.

Yew Ken Chia, Guizhen Chen, Luu Anh Tuan, Soujanya Poria, and Lidong Bing. 2023. Contrastive chain-of-thought prompting. *Preprint*, arXiv:2311.09277.

Zheng Chu, Jingchang Chen, Qianglong Chen, Weijiang Yu, Tao He, Haotian Wang, Weihua Peng, Ming Liu, Bing Qin, and Ting Liu. 2023. A survey of chain of thought reasoning: Advances, frontiers and future. *arXiv preprint arXiv:2309.15402*.

Jean Decety and Aikaterini Fotopoulou. 2015. Why empathy has a beneficial impact on others in medicine: unifying theories. *Frontiers in behavioral neuroscience*, 8:457.

Xiang Gao and Kamalika Das. 2024. Customizing language model responses with contrastive in-context learning. *arXiv preprint arXiv:2401.17390*.

Salvatore Giorgi, João Sedoc, Valentin Barriere, and Shabnam Tafreshi. 2024. Findings of wassa 2024 shared task on empathy and personality detection in interactions. In *Proceedings of the 14th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*.

Judith A Hall, Rachel Schwartz, and Fred Duong. 2021. How do laypeople define empathy? *The Journal of Social Psychology*, 161(1):5–24.

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

Bin Li, Yixuan Weng, Qiya Song, Fuyan Ma, Bin Sun, and Shutao Li. 2022a. Prompt-based pre-trained model for personality and interpersonal reactivity prediction. In *Proceedings of the 12th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis*, pages 265–270, Dublin, Ireland. Association for Computational Linguistics.

Bin Li, Yixuan Weng, Qiya Song, Bin Sun, and Shutao Li. 2022b. Continuing pre-trained model with multiple training strategies for emotional classification. In *Proceedings of the 12th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis*, pages 233–238, Dublin, Ireland. Association for Computational Linguistics.

Zongxia Li, Paiheng Xu, Fuxiao Liu, and Hyemi Song. 2023. Towards understanding in-context learning with contrastive demonstrations and saliency maps. *arXiv preprint arXiv:2307.05052*.

Kourosh Meshgi, Maryam Sadat Mirzaei, and Satoshi Sekine. 2022. Uncertainty regularized multi-task learning. In *Proceedings of the 12th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis*, pages 78–88, Dublin, Ireland. Association for Computational Linguistics.

Damilola Omitaomu, Shabnam Tafreshi, Tingting Liu, Sven Buechel, Chris Callison-Burch, Johannes Eichstaedt, Lyle Ungar, and João Sedoc. 2022. Empathic conversations: A multi-level dataset of contextualized conversations. *Preprint*, arXiv:2205.12698.

Nils Rethmeier and Isabelle Augenstein. 2023. A primer on contrastive pretraining in language processing: Methods, lessons learned, and perspectives. *ACM Computing Surveys*, 55(10):1–17.

Tania Singer and Claus Lamm. 2009. The social neuroscience of empathy. *Annals of the new York Academy of Sciences*, 1156(1):81–96.

Weiwei Sun, Zhengliang Shi, Shen Gao, Pengjie Ren, Maarten de Rijke, and Zhaochun Ren. 2023. Contrastive learning reduces hallucination in conversations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 13618–13626.

Himil Vasava, Pramegh Uikey, Gaurav Wasnik, and Raksha Sharma. 2022. Transformer-based architecture for empathy prediction and emotion classification. In *Proceedings of the 12th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis*, pages 261–264, Dublin, Ireland. Association for Computational Linguistics.

Lu Yan and Yong Tan. 2014. Feeling blue? go online: An empirical study of social support among patients. *Information Systems Research*, 25(4):690–709.

A Test Results

Table 3 presents the final performance of our submitted test results. It is noteworthy that since the testset for Track 2 contains some missing data, some preprocessing is required to obtain the final results. The organizers' official approach involves removing all rows with missing data before calculating the evaluation metrics. However, since the missing data in Track 2's three labels are not consistent, an alternative method is to remove missing data separately for each label before calculating the evaluation metrics. These two approaches yielded different results. The results presented here use the official method.

B Task Templates

B.1 Standard Prediction Template

In the Standard prediction task, we used the following template for Track 1, 2, and 3:

```
<|start_header_id|>system<|end_header_id|>
You are a helpful assistant.<|eot_id|>

<|start_header_id|>user<|end_header_id|>
{taskDescription}

You must strictly follow the following template to generate
your the prediction:
Emotion intensity: {{ your prediction }}, empathy level:
{{ your prediction }}, emotion polarity: {{ your prediction }}.

Article title: {title}
Article source: {source}
Object of suffering: {objectOfSuffering}
Article content: {articleContent}

Speaker1: {content}
<|eot_id|>

<|start_header_id|>assistant<|end_header_id|>
Emotion intensity: {emotionValue}, empathy level:
{empathyValue}, emotion polarity: {emotionPolarity}.
<|eot_id|>
```

Figure 2: The template of Track 2's Standard Prediction

B.2 Role-play Template

In the Role-play task, we used the conversation data from Track 2, the label results, and the following template to construct the training set:

```
<|start_header_id|>system<|end_header_id|>
You are a helpful assistant.<|eot_id|>

<|start_header_id|>user<|end_header_id|>
{taskDescription}

Meanwhile, the response should present the following
traits:
The emotion intensity is {emotionValue}, the empathy l
evel is {empathyValue}, the emotion polarity is {emotio
nPolarity}, and the self-disclosure status is {selfDisclos
ure}.

Article title: {title}
Article source: {source}
Object of suffering: {objectOfSuffering}
Article content: {articleContent}

Conversation history: {convHistory}
<|eot_id|>

<|start_header_id|>assistant<|end_header_id|>
{response}
<|eot_id|>
```

B.3 Contrastive templates

In contrastive SFT, we built two sets of contrast templates for Track 1.

For the comparison of two speakers in the same conversation, we used the template shown in Fig 4.

For the comparison of two people in different dialogues, we used the template in Fig 5.

```
<|start_header_id|>system<|end_header_id|>
You are a helpful assistant.<|eot_id|>

<|start_header_id|>user<|end_header_id|>
This is a Perceived Empathy Level Comparison task.
You are asked to compare the perceived empathy lev
els of two speakers in a dialogue and predict the per
ceived empathy level for each person. In this dialogu
e, each turn is tagged with "Speaker1" or "Speaker2"
to indicate the speaker. In prediction, please use the
speaker tag to refer to the corresponding speaker's id
entity. You need to first provide the comparison result
 and then give the perceived empathy level for each
speaker. The empathy levels are divided into 9 levels.
All predictions must be within the range between 0 to
 8 and must be made using integers only.

Dialogue:
{content}
<|eot_id|>

<|start_header_id|>assistant<|end_header_id|>
Speaker1's perceived empathy level in this conversatio
n is [higher than/lower than/equal to] Speaker2's. Spe
aker1's perceived empathy level is {sp1pel}, and Spea
ker2's perceived empathy level is {sp2pel}.
<|eot_id|>
```

Figure 4: The contrastive template of Track1's two speakers within one dialogue

For Track2's single-turn situation, we create 3 templates for 3 labels. Taking Emotion Polarity as an example, the template is in Fig 6.

For Track2's multi-turn situation, we also create 3 templates for 3 labels. Taking Emotion Polarity as an example, the template is in Fig 7.

For Track3's Empathy/Distress prediction tasks, we used the templates shown in Fig 8 and 9.

For Track3's Empathy/Distress prediction tasks, we used the templates shown in Fig 8 and 9.

C Post-competition experiments

Table 4, 5 show the result comparison of different i on Track 2's devset and test set. It can be seen that increasing i can further enhances the performance

Table 3: The results of our submission on test set

Pearson	Track 1	Track 2				Track 3		
Correlation	Empathy	Emotion	Empathy	Polarity	AVG	Empathy	Distress	AVG
Our method	0.172	0.607	0.582	0.680	0.623	0.474	0.311	0.393

<|start_header_id|>system<|end_header_id|>
You are a helpful assistant.<|eot_id|>

<|start_header_id|>user<|end_header_id|>

This is a Perceived Empathy Level Comparison task across multiple dialogues. You are asked to compare the perceived empathy levels of a speaker in two dialogues and predict the perceived empathy levels of this speaker in both dialogues. In both conversations, the speaker to be predicted will be tagged as "Speaker1", while other speakers will be tagged as "Speaker2". The two dialogues will be tagged as "Conv1" and "Conv2". In prediction, please use the corresponding conversation tags to refer to each conversation. You only need to compare and predict the perceived empathy levels of Speaker1 in both conversations. There is no need to annotate Speaker2. You should first provide the comparison result, followed by the perceived empathy levels of Speaker1 in both conversations. The perceived empathy levels are divided into 9 levels. All annotations are in the range between 0 to 8 and must be made using integers only.

Conv1:
{conv1content}

Conv2:
{conv2content}
<|eot_id|>

<|start_header_id|>assistant<|end_header_id|>
Speaker1's perceived empathy level in Conv1 is [higher than/lower than/equal to] Conv2. The perceived empathy level in Conv1 is {conv1l}, and in Conv2 it is {conv2l}.
<|eot_id|>

Figure 5: The contrastive template of Track1's two speakers with two dialogues

<|start_header_id|>system<|end_header_id|>
You are a helpful assistant.<|eot_id|>

<|start_header_id|>user<|end_header_id|>
This is an Emotional Polarity Comparison task at the sentence level. You are asked to compare the emotional polarities of these two sentences and predict the emotional polarities for each sentence. The two sentences are tagged as "Sent1" and "Sent2". In prediction, please use the corresponding sentence tag to refer to each sentence. You should first provide the comparison result, then give the emotional polarity for each sentence. The emotional polarity is divided into 10 levels. All annotations are in the range between 0 to 9 and must be made using integers only.

Sent1:
{sent1}

Sent2:
{sent2}
<|eot_id|>

<|start_header_id|>assistant<|end_header_id|>
Sent1's emotional polarity is [higher than/lower than/equal to] Sent2. The emotional polarity of Sent1 is {s1polarity}, and Sent2 is {s2polarity}.
<|eot_id|>

Figure 6: The contrastive template of Track2's single-turn Emotion Polarity

<|start_header_id|>system<|end_header_id|>
You are a helpful assistant.<|eot_id|>

<|start_header_id|>user<|end_header_id|>
This is an Emotional Polarity Comparison task at the conversation-turn level. You are asked to compare the emotional polarities of the last conversation turn of two dialogues and predict the emotional polarities for both. You need to focus on the last turn of the dialogue, which starts with the "Speaker1: " string at each dialogue's last line. The two dialogues are tagged as "Dia1" and "Dia2". In prediction, please use the corresponding dialogue tag to refer to each dialogue. You should first provide the comparison result, then give the perceived emotional polarity for each dialogue's last turn. The emotional polarity is divided into 10 levels. All annotations are in the range between 0 to 9 and must be made using integers only.

Dia1:
{dia1}

Dia2:
{dia2}
<|eot_id|>

<|start_header_id|>assistant<|end_header_id|>
The emotional polarity of the Dia1's last turn is [higher than/lower than/equal to] than Dia2. The emotional polarity of Dia1's last turn is {s1polarity}, and Dia2's last turn is {s2polarity}.
<|eot_id|>

Figure 7: The contrastive template of Track2's multi-turn Emotion Polarity

<|start_header_id|>system<|end_header_id|>
You are a helpful assistant.<|eot_id|>

<|start_header_id|>user<|end_header_id|>
This is an Empathy Concern Level Comparison task at the essay level. You are asked to compare the empathy concern levels of two essays and predict the empathy concern levels for each essay. The two essays will be tagged as "Essay1" and "Essay2". In prediction, please use the corresponding essay tag to refer to each essay. You should first provide the comparison result, then give the empathy concern level for each essay. The empathy concern level is divided into 43 levels. All annotations are in the range between 0 to 42 and must be made using integers only.

Essay1:
{e1content}

Essay2:
{e2content}
<|eot_id|>

<|start_header_id|>assistant<|end_header_id|>
Essay1's empathy concern level is [higher than/lower than/equal to] Essay2's. Essay1's empathy concern level is {e1value}, and Essay2's empathy concern level is {e2value}.
<|eot_id|>

Figure 8: The contrastive template of Track3's Empathy

<|start_header_id|>system<|end_header_id|>
You are a helpful assistant.<|eot_id|>

<|start_header_id|>user<|end_header_id|>
This is an Personal Distress Level Comparison task at the essay level. You are asked to compare the personal distress levels of two essays and predict the personal distress levels for each essay. The two essays will be tagged as "Essay1" and "Essay2". In prediction, please use the corresponding essay tag to refer to each essay. You should first provide the comparison result, then give the personal distress level for each essay. The personal distress level is divided into 43 levels. All annotations are in the range between 0 to 42 and must be made using integers only.

Essay1:
{e1content}

Essay2:
{e2content}
<|eot_id|>

<|start_header_id|>assistant<|end_header_id|>
Essay1's personal distress level is [higher than/lower than/equal to] Essay2's. Essay1's personal distress level is {e1value}, and Essay2's personal distress level is {e2value}.
<|eot_id|>

Figure 9: The contrastive template of Track3's Distress

of three label predictions.

Table 4: The performance comparison with different i on Track 2 devset

Pearson	Track 2 multi turn			
Correlation	Emotion	Empathy	Polarity	AVG
LR: 2e-5	0.625	0.636	0.747	0.669
+CRC (i = 1)	0.641	0.664	**0.790**	0.698
+CRC (i = 4)	**0.650**	**0.672**	**0.790**	**0.704**

Table 5: The performance comparison with different i on Track 2 testset

Pearson	Track 2 multi turn			
Correlation	Emotion	Empathy	Polarity	AVG
CRC (i = 1)	**0.607**	0.582	0.680	0.623
CRC (i = 4)	0.606	**0.586**	**0.685**	**0.626**

D Models

All models are released in our Huggingface team website[3]:

- **baseline model**: WASSA2024 EmpathyDetection Chinchunmei EXP300
- **baseline + role-play**: WASSA2024 EmpathyDetection Chinchunmei EXP302
- **baseline + role-play + C-SFT (LR 2e-5)**: WASSA2024 EmpathyDetection Chinchunmei EXP304
- **baseline + role-play + C-SFT (LR 8e-5)**: WASSA2024 EmpathyDetection Chinchunmei EXP305

[3]https://huggingface.co/collections/RicardoLee/chinchunmei-on-wassa2024-shared-task-1-66853bab4fd43e12c535efa8