Chinchunmei at SemEval-2025 Task 11: Boosting the Large Language Model's Capability of Emotion Perception using Contrastive Learning

Tian Li[1], Yujian Sun[2], Huizhi Liang[1]
[1]School of Computing, Newcastle University, Newcastle upon Tyne, UK
[2]Shumei AI Research Institute, Beijing, China
t.li56@newcastle.ac.uk, sunyujian@ishumei.com,
huizhi.liang@newcastle.ac.uk

Abstract

The SemEval-2025 Task 11, Bridging the Gap in Text-Based Emotion Detection, introduces an emotion recognition challenge spanning over 28 languages. This competition encourages researchers to explore more advanced approaches to address the challenges posed by the diversity of emotional expressions and background variations. It features two tracks: multi-label classification (Track A) and emotion intensity prediction (Track B), covering six emotion categories: anger, fear, joy, sadness, surprise, and disgust. In our work, we systematically explore the benefits of two contrastive learning approaches: sample-based (Contrastive Reasoning Calibration) and generation-based (DPO, SimPO) contrastive learning. The sample-based contrastive approach trains the model by comparing two samples to generate more reliable predictions. The generation-based contrastive approach trains the model to differentiate between correct and incorrect generations, refining its prediction. All models are fine-tuned from LLaMa3-Instruct-8B. Our system achieves 9th place in Track A and 6th place in Track B for English, while ranking among the top-tier performing systems for other languages.

1 Introduction

Text-Based Emotion Detection (TBED) has long been a prominent research area in NLP, with widespread applications in social media analysis (Kuamri and Babu, 2017; Salam and Gupta, 2018; Cassab and Kurdy, 2020), mental health treatment (Kusal et al., 2021; Krommyda et al., 2021), and dialogue systems (Liu et al., 2022; Ide and Kawahara, 2022; Hu et al., 2021). Depending on how emotions are defined, TBED can be broadly categorized into two approaches: (1) Classification-based methods, where emotions are categorized into discrete labels (Ekman and Friesen, 1969; Plutchik, 1982). (2) Scoring-based methods, where emotions are treated as interrelated entities with varying intensity levels (Russell and Mehrabian, 1977).

However, due to the nuanced and complex nature of emotional expression, TBED faces several key challenges (Al Maruf et al., 2024): (1) The distinction between different emotions is often subtle, and emotions are often conveyed implicitly—through metaphors or situational cues rather than explicit words. (2) Cultural and linguistic differences influence emotion perception. These challenges make TBED difficult to rely solely on predefined lexicons. A robust TBED system must integrate cultural context, linguistic diversity, background knowledge, and advanced semantic understanding.

SemEval-2025, Task 11, titled "Bridging the Gap in Text-Based Emotion Detection" (Muhammad et al., 2025b), introduces a multilingual benchmark covering 28 languages (Muhammad et al., 2025a; Belay et al., 2025). The competition consists of Track A (multi-label classification) and Track B (intensity prediction). The goal is to identify the speaker's perceived emotion in a given sentence. Emotion categories follow Ekman's framework (Ekman and Friesen, 1969), encompassing six basic emotions: anger, fear, sadness, joy, disgust, and surprise. Task B further introduces four intensity levels for each emotion. This competition setup encapsulates both primary TBED methodologies while incorporating challenges in multilingual and fine-grained emotion recognition.

To participate in both tracks and support all languages predictions, we adopt the generative large language model (LLM). This decision is driven by its robust multi-task integration capabilities and strong support for cross-linguistic applications.

In this paper, we explore two alternative types of approaches - sample-based contrastive learning and generation-based contrastive learning - to address the complexity of emotional expression and the ambiguity of sentiment labels. The sample-based approach leverages Contrastive Reasoning

Proceedings of the The 19th International Workshop on Semantic Evaluation (SemEval-2025), pages 319–330

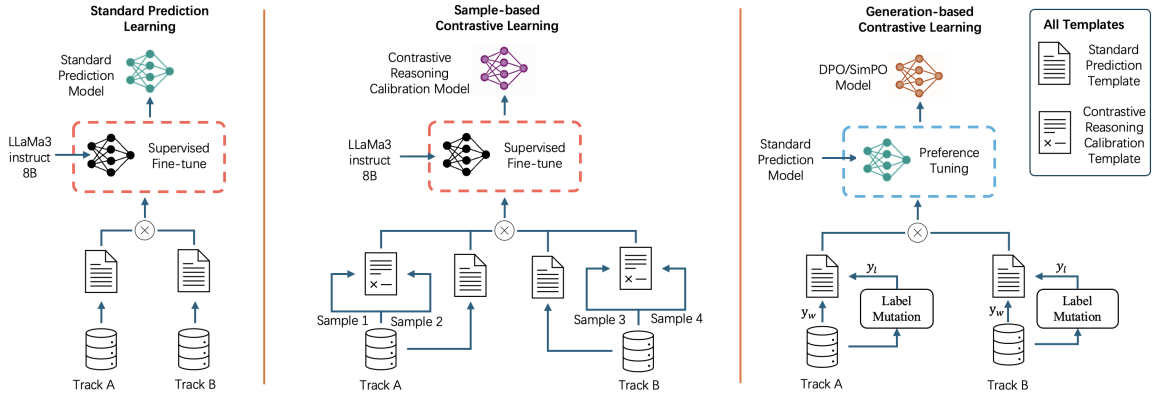

Figure 1: The 3 types of technical solutions we adopted in this competition. On the left is the Standard Prediction training. The middle is the sample-based contrastive training, where Samples 1, 2, 3, and 4 are randomly drawn from the training datasets. On the right is the generation-based contrastive training, where incorrect generations are derived through label mutation.

Calibration technology (CRC) (Li et al., 2024), which enhances prediction reliability by generating multiple predictions through sample comparisons and aggregating them via majority voting. The generation-based approaches employ preference optimization techniques (Rafailov et al., 2023; Hong et al., 2024; Meng et al., 2025), refining the model's comprehension of sentiment labels by increasing the log probability of correct outputs while reducing that of incorrect ones. Given computational constraints, we only explore DPO (Rafailov et al., 2023) and SimPO (Meng et al., 2025), two widely acknowledged methods that do not require the reward model.

Our key contributions are as follows:

- We explored the impact of non-English data on English sentiment prediction under limited computational resources. Surprisingly, experimental results indicate that incorporating non-English data degrades performance in both classification tasks (Track A) and scoring tasks (Track B).
- We conducted a comprehensive evaluation and bad case analysis of two contrastive approaches on the competition dataset. In the sample-based contrastive Learning, CRC technology yielded limited benefit. In the generation-based contrastive Learning, DPO demonstrated a significant positive effect on Track B. Although SimPO has achieved gains on some labels, the overall effect has dropped significantly.
- In the leaderboard, our approach achieved Track A top 10 in 16 languages and 9th in English; Track B top 10 across all languages and 6th in English.

2 Methodology

To reduce our workload, we integrate both Track A and Track B into the same model by using different prompt templates for each. This unified approach enables the model to dynamically switch between prediction tasks as needed.

In our explorations, we designed three distinct tasks:

- Standard Prediction (Baseline): The model is trained to directly output all emotion labels based on the input text.
- Sample-Based Contrastive Learning: The model learns to compare two samples to generate more reliable predictions, leveraging contrastive reasoning to refine label predictions.
- Generation-Based Contrastive Learning: The model learns to differentiate between correct and incorrect predictions, improving its ability to generate accurate label outputs.

Figure 1 illustrates the training workflow of these three tasks.

2.1 Standard prediction

During sample preparation, we integrate the text content into a pre-designed prompt template as input and format the label results as the ground truth output. The template details can be found in Appendix A.1. The model is then trained through supervised fine-tune (SFT) using this formatted input and corresponding output.

During inference, we apply the same prompt template to incorporate the input text for prediction. By parsing the generated output, we extract the corresponding label predictions.

2.2 Sample-based contrastive learning

In this category, we use CRC, a technique designed to enhance the model's ability to discern subtle differences in samples. By having the model compare the score variations between two samples, model can better understand their distinctions and generate calibrated predictions. The key to this task lies in sample preparation and inference process.

During sample preparation, we randomly select two samples from the training set (Figure 1 middle) and construct a contrastive pair using the template in Appendix A.1. The target output comprises two components: a contrastive summary and two samples' predictions. The contrastive summary, expressed in natural language, highlights the existence or intensity difference between two samples on a specific label. The predictions provide explicit scores for both samples on the given label.

Given that random pairwise sampling can generate a vast amount of training data, we impose an upper limit on the total number of sampled pairs for each track. Specifically, for Track A, we sample 3,000 contrastive pairs per label, while for Track B, we sample 6,000 pairs per label, as Track B involves more fine-grained scoring labels compared to Track A.

During inference, each test sample is paired with a randomly selected training sample to form a contrastive input. Since the model achieves highly accurate predictions on training samples, these trained samples serve as reliable reference points, reducing prediction uncertainty. The two samples are integrated into the CRC prompt template, with the test sample randomly assigned to either position 1 or position 2. This process is repeated N times to generate N different input instances, producing N predictions. The final score is determined through a voting mechanism, where the most frequently predicted score is selected as the final output.

2.3 Generation-based contrastive learning

To enhance the model's sensitivity to scoring, we incorporate preference optimization. However, due to computational constraints, we only explore techniques that do not require a reward model, such as DPO and SimPO.

DPO optimizes the language model by maximizing the relative probability ratio to favor preferred outputs (Eq 1). Here, π_θ and π_{ref} represent the target and reference models, respectively, while y_w and y_l denote the correct and incorrect outputs. β is a scaling hyperparameter. This optimization process is applied after SFT.

$$-log\sigma(\beta log\frac{\pi_\theta(y_w|x)}{\pi_{ref}(y_w|x)} - \beta log\frac{\pi_\theta(y_l|x)}{\pi_{ref}(y_l|x)}) \quad (1)$$

SimPO adopts a similar optimization objective (Eq 2), directly enhancing the probability of the target model's preferred outputs. However, it simplifies DPO by discarding the reference model and using sequence length to normalize the loss. Additionally, it adopts a hyperparameter γ to ensure that the likelihood for the correct response exceeds the incorrect response by γ. Like DPO, SimPO is also applied after SFT.

$$-log\sigma(\frac{\beta}{|y_w|}log\pi_\theta(y_w|x) - \frac{\beta}{|y_l|}log\pi_\theta(y_l|x) - \gamma) \quad (2)$$

During sample preparation, to enhance the model's sensitivity to label scoring, we mute only the label scores to generate y_l (Figure 1 right, the Label Mutation block). The y_w corresponds to the original SP ground truth, while the y_l follows the same SP template but with muted scores. The probabilities of muting one, two, three, four, and five labels are [63.8%, 26.1%, 8.3%, 1.6%, 0.1%], respectively. When mutation is triggered, a random value is selected from the remaining label values as the incorrect score.

For Track A, each training sample executes five mutations, generating five contrastive samples for generation-based contrastive learning. For Track B, mutation is repeated 15 times, creating 15 contrastive samples. These settings ensured that the preference optimization dataset was comparable in size to the CRC dataset, allowing for a fair comparison between sample-based and generation-based approaches.

3 Experiment

In this competition, we participate in Track A and Track B. Track A covers 28 languages, while Track B includes 11 languages. Each language contains thousands of samples. Some languages use the same dataset for both Track A and Track B. Both

Model	Training Language	Track A - eng						
		Macro	**Micro**	**Anger**	**Fear**	**Joy**	**Sadness**	**Surprise**
SP	English	**0.828**	**0.808**	0.738	**0.876**	0.824	0.818	0.784
SP	Multilingual	0.820	0.802	0.742	0.865	0.814	0.808	0.780
CRC	English	0.819	0.802	**0.752**	0.865	0.811	0.819	0.765
DPO	English	0.827	0.806	0.736	0.875	0.816	0.821	**0.785**
SimPO	English	0.748	0.741	0.700	0.737	**0.827**	**0.832**	0.607

Table 1: English test results of all models in Track A

Model	Training Language	Track B - eng						
		Macro	**Micro**	**Anger**	**Fear**	**Joy**	**Sadness**	**Surprise**
SP	English	0.845	0.823	**0.812**	0.841	**0.850**	0.847	0.763
SP	Multilingual	0.835	0.815	0.807	0.816	0.845	0.840	0.765
CRC	English	0.828	0.805	0.796	0.807	0.836	0.837	0.750
DPO	English	**0.846**	**0.824**	0.810	**0.844**	0.846	**0.849**	**0.773**
SimPO	English	0.770	0.741	0.700	0.737	0.827	0.832	0.607

Table 2: English test set results of all models in Track B

tracks share the same sentiment labels: anger, fear, joy, sadness, surprise, and disgust. However, some languages have missing labels (e.g., English does not include the disgust label). All sample contents are single-turn dialogue without prior dialogue. Track A is evaluated using the F1 score, while Track B uses Pearson Correlation, with both metrics computed in Macro and Micro modes. Macro calculates the overall score across all labels, while Micro first computes per-label scores and then averages them. Although training, development, and test sets are provided, the development set is too small, leading to unstable evaluation results. Therefore, all reported results are based on the test set. Please see Appendix A.2 for the dev set results and discussion.

All models in our experiments are fine-tuned from Llama3-Instruct-8B (Dubey et al., 2024), selected for its strong multilingual capabilities and acceptable computational cost. To validate its multilingual capabilities in each language, we check the consistency between original and recovered text using its tokenizer to encode and decode the text content. This experiment confirmed that the model supports all competition languages.

We first evaluate the impact of multilingual data on English-only performance in the standard prediction (SP) task. Then, we experiment with CRC, DPO, and SimPO on the English dataset to examine the effectiveness of sample-based and generation-based contrastive learning for TBED. The CRC model is trained via SFT on a combination of SP and CRC datasets. DPO and SimPO models are obtained by applying preference tuning to the English-only SP model.

To reduce training costs, we use LoRA (Hu et al., 2022) for all fine-tuning tasks, with its rank and alpha set as 8 and 16. Each training lasts for 3 epochs with a batch size of 128. Regarding the learning rate (LR), SP and CRC employ 4e-4, DPO adopts 5e-6, and SimPO uses 1e-6. All models are trained using the AdamW optimizer, with $\beta_1 = 0.8, \beta_2 = 0.99$. All LR scheduler employs cosine decay with a warmup ratio of 0.1. For CRC inference, we set $N = 3$ for Track A and $N = 7$ for Track B.

All experiments are conducted using 8-GPU distributed training. Due to limited computational resources, different GPUs are used across various training runs. However, all GPUs are based on the Ada Lovelace or Hopper architecture, allowing us to leverage a wide range of existing acceleration techniques to enhance training efficiency.

4 Results

Tables 1 and 2 present the performance of all models on the English test set. Table 1 corresponds to Track A, the multi-emotion classification task, while Table 2 corresponds to Track B, the emotion intensity prediction task.

4.1 Multilingual influence

We observe that multilingual training underperforms compared to English-only training in both Track A and Track B. This finding highlights the significant differences in emotion perception across languages and cultural contexts, which can introduce conflicts in the model's understanding of sentiment labels. Therefore, we submit the non-English results generated by the multilingual SP model to

ID	Text	Pred	True
eng_test_track_a_02136	It was growing harder for him to keep balanced with my legs shifting every few seconds, until finally, I found the right position and his back a good shove with my knee.	0	1
eng_test_track_a_01815	So you let it shatter, breaking at my feet.	1	0
eng_test_track_a_01594	Dad thought it was just my imagination.	0	1
eng_test_track_a_00071	There's just something about watching an acne-ridden high school dropout cooking my food that just doesn't sit well in my stomach.	0	1

Table 3: Bad cases of the CRC model on the Anger label of Track A

the leaderboard and discard the multilingual setting in subsequent comparison experiments. For the multilingual SP results of Track A and Track B in all languages, please see Appendix A.3.

4.2 Sample-based vs generation-based contrastive learning

In Track A, the SP model performed the best overall. Specifically, it outperformed other approaches on the fear label, while the DPO model achieved a slight advantage on surprise. For anger, the CRC model ranked first, whereas SimPO achieved the best performance on joy and sadness. However, in Track B, DPO secures the top position across all evaluation metrics except for anger and joy. Meanwhile, SimPO lags significantly behind other systems in both macro and micro performance metrics. To further investigate the underlying reasons for these results, we conduct the bad case analysis for different techniques.

For the CRC model, we analyze Track A anger's misclassifications. We find that most misclassifications—accounting for 70% of the errors—are due to the model incorrectly predicting a neutral emotional state. Table 3 presents 4 randomly sampled bad cases, illustrating that these instances are on the borderline of the anger definition. Comparing them with other samples can easily influence their predictions, causing uncertainty and error. This suggests that the dataset may not be well-suited for sample-based comparison approach.

For the DPO model, we analyze sadness's wrong cases in Track B. We observe that over 90% of the errors differs from the ground truth by only one intensity level. This indicates that the model still has room for improvement in its recognition of label intensity.

Regarding the SimPO model, we figure out that its poor performance primarily stemmed from the loss of output formatting, leading to frequent content parsing errors. This highlights the critical role of the reference model in preference tuning. While SimPO effectively increases the probability margin between correct and incorrect outputs, it may also make correct outputs no longer rank as the top generation.

5 Conclusion

The SemEVAL-2025 organizers introduce a highly challenging text-based emotion detection dataset, covering multi-label classification and intensity prediction across 28 languages. The dataset reflects the complexity of emotional expression and the diversity introduced by different linguistic and cultural backgrounds.

We explore three types of approaches in this competition. The baseline approach is the standard prediction task. The two enhanced types of approaches were sample-based contrastive learning and generation-based contrastive learning. For the sample-based approach, we try CRC. For the generation-based approach, we explore both DPO and SimPO. They all aim to improve model prediction through contrastive training—one by comparing samples and the other by discriminating between correct and incorrect generations.

Our experiments reveal that different languages reflect distinct cultural backgrounds, so multilingual training does not improve English emotion detection. Meanwhile, due to the inherent ambiguity of sentiment expressions, sample-based contrastive learning raises additional uncertainty, ultimately reducing prediction accuracy. On the other hand, generation-based contrastive learning provides consistent improvements in intensity prediction, though its effectiveness varies significantly across different techniques. Notably, reference model constraint is crucial in stabilizing generation-based contrastive optimization process. It prevents excessive deviation from the original model distribution, and preserves key capabilities such as structured output generation.

References

Abdullah Al Maruf, Fahima Khanam, Md Mahmudul Haque, Zakaria Masud Jiyad, Firoj Mridha, and Zeyar Aung. 2024. Challenges and opportunities of text-based emotion detection: A survey. *IEEE Access*.

Tadesse Destaw Belay, Israel Abebe Azime, Abinew Ali Ayele, Grigori Sidorov, Dietrich Klakow, Philip Slusallek, Olga Kolesnikova, and Seid Muhie Yimam. 2025. Evaluating the capabilities of large language models for multi-label emotion understanding. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 3523–3540, Abu Dhabi, UAE. Association for Computational Linguistics.

S Cassab and Mohamad-Bassam Kurdy. 2020. Ontology-based emotion detection in arabic social media. *International Journal of Engineering Research & Technology (IJERT)*, 9(08):1991–2013.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Paul Ekman and Wallace V Friesen. 1969. The repertoire of nonverbal behavior: Categories, origins, usage, and coding. *semiotica*, 1(1):49–98.

Jiwoo Hong, Noah Lee, and James Thorne. 2024. Orpo: Monolithic preference optimization without reference model. *arXiv preprint arXiv:2403.07691*.

Dou Hu, Lingwei Wei, and Xiaoyong Huai. 2021. 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, Weizhu Chen, and 1 others. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

Tatsuya Ide and Daisuke Kawahara. 2022. Building a dialogue corpus annotated with expressed and experienced emotions. *arXiv preprint arXiv:2205.11867*.

Maria Krommyda, Anastasios Rigos, Kostas Bouklas, and Angelos Amditis. 2021. An experimental analysis of data annotation methodologies for emotion detection in short text posted on social media. In *Informatics*, volume 8, page 19. MDPI.

Santoshi Kuamri and C Narendra Babu. 2017. Real time analysis of social media data to understand people emotions towards national parties. In *2017 8th international conference on computing, communication and networking technologies (ICCCNT)*, pages 1–6. IEEE.

Sheetal Kusal, Shruti Patil, Ketan Kotecha, Rajanikanth Aluvalu, and Vijayakumar Varadarajan. 2021. Ai based emotion detection for textual big data: Techniques and contribution. *Big Data and Cognitive Computing*, 5(3):43.

Tian Li, Nicolay Rusnachenko, and Huizhi Liang. 2024. Chinchunmei at wassa 2024 empathy and personality shared task: Boosting llm's prediction with role-play augmentation and contrastive reasoning calibration. In *Proceedings of the 14th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 385–392.

Yuchen Liu, Jinming Zhao, Jingwen Hu, Ruichen Li, and Qin Jin. 2022. Dialogueein: Emotion interaction network for dialogue affective analysis. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 684–693.

Yu Meng, Mengzhou Xia, and Danqi Chen. 2025. Simpo: Simple preference optimization with a reference-free reward. *Advances in Neural Information Processing Systems*, 37:124198–124235.

Shamsuddeen Hassan Muhammad, Nedjma Ousidhoum, Idris Abdulmumin, Jan Philip Wahle, Terry Ruas, Meriem Beloucif, Christine de Kock, Nirmal Surange, Daniela Teodorescu, Ibrahim Said Ahmad, David Ifeoluwa Adelani, Alham Fikri Aji, Felermino D. M. A. Ali, Ilseyar Alimova, Vladimir Araujo, Nikolay Babakov, Naomi Baes, Ana-Maria Bucur, Andiswa Bukula, and 29 others. 2025a. Brighter: Bridging the gap in human-annotated textual emotion recognition datasets for 28 languages. *Preprint*, arXiv:2502.11926.

Shamsuddeen Hassan Muhammad, Nedjma Ousidhoum, Idris Abdulmumin, Seid Muhie Yimam, Jan Philip Wahle, Terry Ruas, Meriem Beloucif, Christine De Kock, Tadesse Destaw Belay, Ibrahim Said Ahmad, Nirmal Surange, Daniela Teodorescu, David Ifeoluwa Adelani, Alham Fikri Aji, Felermino Ali, Vladimir Araujo, Abinew Ali Ayele, Oana Ignat, Alexander Panchenko, and 2 others. 2025b. SemEval task 11: Bridging the gap in text-based emotion detection. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, Vienna, Austria. Association for Computational Linguistics.

R Plutchik. 1982. A psycho evolutionary theory of emotions. *Social Science Information*.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36:53728–53741.

James A Russell and Albert Mehrabian. 1977. Evidence for a three-factor theory of emotions. *Journal of research in Personality*, 11(3):273–294.

Shaikh Abdul Salam and Rajkumar Gupta. 2018. Emotion detection and recognition from text using machine learning. *Int. J. Comput. Sci. Eng*, 6(6):341–345.

A Appendix

A.1 Prompt templates

Table 4, 5, 6, and 7 show the prompt templates used by Standard Prediction and Contrastive Reasoning Calibration on Track A and B, respectively. Preference Tuning also uses the Standard Prediction template.

A.2 Development set discussion

Table 8 and 9 present our system's performance on the English development set. However, the results from this dataset are somewhat misleading due to the following reasons:

- In Track A, the development set suggests that multilingual training significantly benefits English performance, which contradicts the findings on the test set.
- In Track A, the CRC technique outperforms all other models by a substantial margin, which is inconsistent with its performance on the test set.

During the competition's evaluation phase, we submitted our results based on these misleading insights. Then, we discovered that our system's actual performance on the test set did not meet expectations. Upon further analysis, we identified that this discrepancy primarily stems from the development set's limited size. With only 116 samples and highly imbalanced label distributions, certain labels had extremely sparse scores, making the computed metrics unreliable.

To ensure the reliability of our conclusions, we have removed the development set results from the main text to prevent any potential misinterpretation.

A.3 Test results of SP multilingual model on all languages

Table 10 shows the performance of the multilingual model on all languages, including Track A and B.

Input	Output
Task Description: You are tasked with determining the perceived emotion(s) of a speaker based on a conversation. Specifically, your goal is to predict the emotions that most people would associate with the speaker's last utterance. The possible emotions are: joy, sadness, fear, anger, surprise, and disgust. The conversation may be in any of the following languages: Afrikaans, Algerian Arabic, Amharic, Emakhuwa, Hausa, Igbo, Kinyarwanda, Moroccan Arabic, Mozambican Portuguese, Nigerian-Pidgin, Oromo, Setswana, Somali, Swahili, Sundanese, Tigrinya, Xitsonga, IsiXhosa, Yoruba, isiZulu Arabic, Chinese, Hindi, Indonesian, Javanese, Marathi English, German, Romanian, Russian, Latin American Spanish, Tatar, Ukrainian, Swedish, Mozambican Portuguese, and Brazilian Portuguese. Instructions: 1. The language of the conversation will be explicitly indicated at the first place. 2. Each turn in the conversation will be marked with "Speaker1" or "Speaker2" to indicate the speaker. 3. You need to predict the emotions based on the last utterance from "Speaker1" (and any additional context or dialogue history if provided). 4. For each emotion, indicate whether it applies using binary labels: 1 (emotion is present) or 0 (emotion is absent). Example Output Format: joy: {{ 1 or 0 }}, sadness: {{ 1 or 0 }}, fear: {{ 1 or 0 }}, anger: {{ 1 or 0 }}, (optional) surprise: {{ 1 or 0 }}, (optional) disgust: {{ 1 or 0 }}. Language: {lan} Content: Speaker1: {text}	joy: {joy}, sadness: {sadness}, fear: {fear}, anger: {anger}, surprise: {surprise}, disgust: {disgust}.

Table 4: Standard prediction template for Track A

Input	Output
Task Description: You are tasked with predicting the intensity for each of the perceived emotion classes of a speaker based on a conversation. Specifically, your prediction should represent the emotional intensity most people associate with the speaker's last utterance. The possible emotion classes are: joy, sadness, fear, anger, surprise, and disgust. The conversation may be in any of the following languages: Afrikaans, Algerian Arabic, Amharic, Emakhuwa, Hausa, Igbo, Kinyarwanda, Moroccan Arabic, Mozambican Portuguese, Nigerian-Pidgin, Oromo, Setswana, Somali, Swahili, Sundanese, Tigrinya, Xitsonga, IsiXhosa, Yoruba, isiZulu Arabic, Chinese, Hindi, Indonesian, Javanese, Marathi English, German, Romanian, Russian, Latin American Spanish, Tatar, Ukrainian, Swedish, Mozambican Portuguese, and Brazilian Portuguese. Instructions: 1. The language of the conversation will be explicitly indicated at the first place. 2. Each turn in the conversation will be marked with "Speaker1" or "Speaker2" to indicate the speaker. 3. You need to predict the emotion intensity based on the last utterance from "Speaker1" (and any additional context or dialogue history if provided). 4. For each emotion class, the ordinal intensity levels include: 0 for no emotion, 1 for a low degree of emotion, 2 for a moderate degree of emotion, and 3 for a high degree of emotion. Example Output Format: joy: {{ 0, 1, 2, or 3 }}, sadness: {{ 0, 1, 2, or 3 }}, fear: {{ 0, 1, 2, or 3 }}, anger: {{ 0, 1, 2, or 3 }}, (optional) surprise: {{ 0, 1, 2, or 3 }}, (optional) disgust: {{ 0, 1, 2, or 3 }}. Language: {lan} Content: Speaker1: {text}	joy: {joy}, sadness: {sadness}, fear: {fear}, anger: {anger}, surprise: {surprise}, disgust: {disgust}.

Table 5: Standard prediction template for Track B

Input	Output
Task Description: Your task is to compare and predict the perceived emotional label exhibited by the speaker in two separate conversations. The target emotion for comparison is "{label}". The conversation may be in any of the following languages: Afrikaans, Algerian Arabic, Amharic, Emakhuwa, Hausa, Igbo, Kinyarwanda, Moroccan Arabic, Mozambican Portuguese, Nigerian-Pidgin, Oromo, Setswana, Somali, Swahili, Sundanese, Tigrinya, Xitsonga, IsiXhosa, Yoruba, isiZulu Arabic, Chinese, Hindi, Indonesian, Javanese, Marathi English, German, Romanian, Russian, Latin American Spanish, Tatar, Ukrainian, Swedish, Mozambican Portuguese, and Brazilian Portuguese. Instructions: 1. The two conversations will be marked as "Conversation1" and "Conversation2". Each turn in the conversation will be marked as "Speaker1" or "Speaker2" to indicate the speaker. 2. The language of the conversation will be explicitly stated at the beginning of each conversation. 3. You only need to predict the emotions of "Speaker1" in both conversations. No predictions are required for "Speaker2". 4. Your comparison and prediction should be based on the last utterance of "Speaker1" in each conversation, while also considering any additional background or dialogue history if provided. 5. First, provide a brief summary of the comparison result between the two conversations. Then, use binary labels to indicate whether the specified emotion ("{label}") is present in each conversation: 1 (emotion is present) or 0 (emotion is absent). Example Output Format: For emotion label "{label}", {{Brief summary of the comparison result}}. Conversation1: {{1 or 0}}, Conversation2: {{1 or 0}}. Conversation1: Language: {lan1} Speaker1: {text1} Conversation2: Language: {lan2} Speaker1: {text2} For emotion label "{label}", {brief}. Conversation1: {conv1Value}, Conversation2: {conv2Value}.	For emotion label "{label}", {{Brief summary of the comparison result}}. Conversation1: {{1 or 0}}, Conversation2: {{1 or 0}}.

Table 6: Contrastive reasoning calibration prompt template for Track A

Input	Output
Task Description: Your task is to compare and predict the intensity of the specific perceived emotion class in two separate conversations. The target preceived emotion class for comparison is "{label}". The conversation may be in any of the following languages: Afrikaans, Algerian Arabic, Amharic, Emakhuwa, Hausa, Igbo, Kinyarwanda, Moroccan Arabic, Mozambican Portuguese, Nigerian-Pidgin, Oromo, Setswana, Somali, Swahili, Sundanese, Tigrinya, Xitsonga, IsiXhosa, Yoruba, isiZulu Arabic, Chinese, Hindi, Indonesian, Javanese, Marathi English, German, Romanian, Russian, Latin American Spanish, Tatar, Ukrainian, Swedish, Mozambican Portuguese, and Brazilian Portuguese. Instructions: 1. The two conversations will be marked as "Conversation1" and "Conversation2". Each turn in the conversation will be marked as "Speaker1" or "Speaker2" to indicate the speaker. 2. The language of the conversation will be explicitly stated at the beginning of each conversation. 3. You only need to predict the emotional intensity of "Speaker1" in both conversations. No predictions are required for "Speaker2". 4. Your comparison and prediction should be based on the last utterance of "Speaker1" in each conversation, while also considering any additional background or dialogue history if provided. 5. First, provide a brief summary of the comparison result between the two conversations. Then, use one of the four levels to indicate the target ordinal intensity: 0 for no emotion, 1 for a low degree of emotion, 2 for a moderate degree of emotion, and 3 for a high degree of emotion. Example Output Format: For emotion label "{label}", {{Brief summary of the comparison result}}. Conversation1: {{ 0, 1, 2, or 3 }}, Conversation2: {{ 0, 1, 2, or 3 }}. Conversation1: Language: {lan1} Speaker1: {text1} Conversation2: Language: {lan2} Speaker1: {text2} For emotion label "{label}", {brief}. Conversation1: {conv1Value}, Conversation2: {conv2Value}.	For emotion label "{label}", {{Brief summary of the comparison result}}. Conversation1: {{1 or 0}}, Conversation2: {{1 or 0}}.

Table 7: Contrastive reasoning calibration prompt template for Track B

Model	Training Language	Track A - eng Dev Set						
		Macro	Micro	Anger	Fear	Joy	Sadness	Surprise
SP	English	0.824	0.812	0.788	0.871	**0.793**	0.812	0.794
SP	Multilingual	0.829	0.825	0.813	0.862	0.833	0.824	0.767
CRC	English	**0.839**	**0.832**	**0.848**	**0.884**	0.778	0.817	**0.831**
DPO	English	0.817	0.806	0.788	0.864	0.767	0.824	0.781
SimPO	English	0.749	0.726	0.643	0.810	0.774	**0.831**	0.538

Table 8: Development set english results of all models in Track A

Model	Training Language	Track B - eng Dev Set						
		Macro	Micro	Anger	Fear	Joy	Sadness	Surprise
SP	English	0.834	0.825	0.836	0.782	0.811	0.874	**0.822**
SP	Multilingual	0.818	0.809	0.824	0.776	**0.830**	0.887	0.680
CRC	English	0.793	0.776	0.778	0.721	0.798	**0.905**	0.714
DPO	English	**0.840**	**0.831**	**0.851**	**0.784**	0.820	0.886	0.814
SimPO	English	0.756	0.731	0.721	0.717	0.815	0.885	0.468

Table 9: Development set english results of all models in Track B

Language	Track A		Track B	
	Macro	Micro	Macro	Micro
eng	0.820	0.802	0.835	0.815
ptbr	0.722	0.607	0.758	0.638
ary	0.591	0.553	-	-
afr	0.669	0.580	-	-
ptmz	0.581	0.523	-	-
kin	0.520	0.475	-	-
pcm	0.693	0.632	-	-
amh	0.778	0.766	0.764	0.726
tat	0.767	0.767	-	-
chn	0.756	0.664	0.781	0.661
ukr	0.690	0.652	0.671	0.633
vmw	0.234	0.188	-	-
yor	0.511	0.353	-	-
orm	0.626	0.522	-	-
rus	0.887	0.887	0.902	0.900
sun	0.708	0.457	-	-
arq	0.594	0.561	0.525	0.481
deu	0.755	0.714	0.761	0.721
esp	0.819	0.823	0.769	0.773
mar	0.862	0.868	-	-
hau	0.680	0.670	0.719	0.692
swa	0.368	0.313	-	-
swe	0.762	0.591	-	-
som	0.469	0.437	-	-
tir	0.565	0.474	-	-
ron	0.770	0.765	0.778	0.682
ibo	0.634	0.576	-	-
hin	0.896	0.896	-	-

Table 10: SP multilingual model results on all tracks and languages