SCUNLP at ROCLING-2026 Shared Task: Augmenting BERT with LLM-Generated Affective Text for Valence-Arousal Prediction of New Immigrants' Texts

Hsiang-Wei Chen
Department of Data Science
Soochow University
Taipei, Taiwan
garena30927@gmail.com

He-Chang Chan
Department of Data Science
Soochow University
Taipei, Taiwan
johnnychan861@gmail.com

Jheng-Long Wu
Department of Data Science
Soochow University
Taipei, Taiwan
jlwu@gm.scu.edu.tw

Abstract

Chinese texts written by new immigrants often contain colloquial expressions, cross-cultural contexts, and implicit affective information, making valence and arousal prediction challenging. This paper presents the SCUNLP system for the ROCLING-2026 shared task. We use gpt-5.4-mini to analyze affective information in each text, concatenate the LLM-generated text with the original text, and fine-tune bert-base-chinese for joint valence-arousal regression. The training data combine Chinese EmoBank with labeled data released by the ROCLING 2025 shared task, yielding 7,094 instances after cleaning. On the official test set, the system obtains valence MAE/PCC of 0.58/0.88 and arousal MAE/PCC of 0.90/0.41, ranking second overall. Compared with a BERT baseline trained without LLM-generated text, valence MAE decreases from 0.61 to 0.58. These results show that LLM-generated text can provide useful additional semantic information for BERT in low-resource Chinese dimensional sentiment analysis.

Keywords: Dimensional Sentiment Analysis, Valence-Arousal Prediction, Large Language Models, Feature Augmentation, BERT

1 Introduction

Emotion is not a set of isolated categories, but a continuous state with differences in direction and intensity. The circumplex model of affect represents emotion along two continuous dimensions: valence and arousal (VA). Valence ranges from unpleasant to pleasant, whereas arousal ranges from calm to activated (Russell, 1980). Compared with discrete positive, negative, or neutral classification, dimensional sentiment analysis preserves more fine-grained affective distinctions. For example, "depressed" and "anxious" may both be negative, but they differ substantially in arousal. Prior work has also built VA resources and demonstrated the value of continuous dimensions for modeling subtle affective states (Buechel and Hahn, 2017; Lee et al., 2022).

However, predicting continuous affective values is highly context-dependent. When texts come from a specific community or life situation, models must also handle domain differences and indirect emotional expression. The ROCLING 2026 shared task focuses on affective texts written by new immigrants in Taiwan (Lin et al., 2026). These texts involve family relationships, work arrangements, migration adaptation, and identity experiences. Emotion may not be expressed through explicit words such as "happy," "sad," or "angry," but may instead be implied through events, negation, contrast, expectation, and resignation. The model therefore needs to infer not only affective polarity but also intensity from the full context.

Most current dimensional sentiment analysis methods fine-tune pretrained language models on manually labeled corpora, allowing the encoder to learn the mapping between text and VA scores (Devlin et al., 2019; Cui et al., 2020). Applying this approach to new-immigrant texts raises two challenges. First, target-domain annotations are limited, so models often need to rely on general affective corpora or data from other domains. Such additional data increase the number of training instances, but may also introduce source-domain vocabulary and affective patterns. Second, affective cues in new-immigrant texts are often distributed across longer narratives. A supervised encoder trained on limited data may not fully capture implicit affective cues and contextual relations. Our work therefore asks how to supplement the model's understanding of affective cues under limited target-domain supervision.

Large language models (LLMs) can perform diverse tasks from natural-language instructions and

Proceedings of the 38th Conference on Computational Linguistics and Speech Processing (ROCLING 2026), pages 477–485
November 6-7, 2026 ©2026 Association for Computational Linguistics

in-context examples (Brown et al., 2020). Their outputs can also be refined through self-generated feedback (Madaan et al., 2023), and prompts can be edited automatically or in a targeted manner (Ye et al., 2024). These capabilities allow LLMs to make dispersed or implicit affective cues more explicit in generated text. However, directly asking an LLM to predict VA scores on a 1-9 scale may introduce errors from prompt wording and scale interpretation. We therefore do not let the LLM determine final scores. Instead, we concatenate its generated text with the original text and leave numerical prediction to a regression model trained with labeled data.

We propose a two-stage architecture that combines LLM-generated text with a BERT regression model. In the first stage, gpt-5.4-mini analyzes the original text and produces natural-language text describing affective polarity, intensity, negation, contrast, and other linguistic information. In the second stage, the original text and the generated text are jointly fed into a Chinese pretrained encoder, followed by a two-output regression layer that predicts valence and arousal. All LLM-generated text is produced and cached in advance, so model training and fixed test-set inference use the same input. This design also makes it easier to compare different prompts and encoders.

In addition to building the two-stage prediction system, we compare different prompt versions, three Chinese pretrained encoders, epoch numbers, and fixed-weight model ensembling to examine the stability of LLM-generated text as additional input. The central question of this paper is not whether an LLM can directly predict VA scores, but whether its generated text can provide useful, inspectable, and reusable auxiliary information for a supervised BERT regression model. On the official test set, bert-base-chinese trained for 3 epochs with the base prompt obtains valence MAE/PCC of 0.58/0.88 and arousal MAE/PCC of 0.90/0.41, ranking second overall.

The main contributions of this paper are as follows. First, we present a two-stage architecture that combines LLM-generated text with a Chinese encoder-based regression model, allowing the LLM to organize affective information while the final VA scores are learned from labeled data. Second, we compare prompt refinement strategies, Chinese pretrained encoders, epoch numbers, and ensembling settings, and summarize how these choices affect valence and arousal prediction. Third, by analyzing prompt content, generated text, and official test results, we show that prompt rules overly tied to a source domain may create domain bias and reduce the usefulness of the generated text in the target task.

2 Related Work

2.1 Dimensional Sentiment Analysis

The valence-arousal (VA) framework originates from Russell (1980). Yu et al. (2016) constructed an early Chinese affective word lexicon and a sentence-level corpus annotated in the valence-arousal dimensions. Lee et al. (2022) later expanded Chinese EmoBank to word, phrase, sentence, and multi-sentence text levels. Chinese EmoBank provides mean and standard-deviation annotations for valence and arousal on a 1-9 scale, enabling models to learn continuous regression rather than only polarity classification. Buechel and Hahn (2017) further note that dimensional emotion annotation is affected by reader/writer perspective and representation format. Thus, under limited annotation and unstable single-point differences, introducing additional and inspectable affective signals becomes important.

2.2 LLMs for Feature and Data Augmentation

Large language models can perform diverse tasks from natural-language instructions and in-context examples (Brown et al., 2020). Using LLM-generated text as additional downstream input differs from directly asking the LLM for final answers: the supervised model still learns the task scale and data distribution, while the LLM inference cost is incurred only during one-time preprocessing. LLM-generated intermediate text has also been used as auxiliary supervision. For example, Hsieh et al. (2023) use LLM-generated rationales in a multi-task framework to train smaller models. Our method differs from rationale distillation: we concatenate the generated text with the original input and train BERT only on the downstream valence-arousal regression objective. We do not assume that the generated content is necessarily correct; it is only a possible signal that may complement the original text.

2.3 Prompt Refinement and Consistency-Based Revision

Madaan et al. (2023) iteratively refine model outputs using self-generated feedback, whereas Ye et al. (2024) use LLMs to perform automatic and targeted prompt editing. Standard self-consistency samples multiple reasoning paths and aggregates their answers (Wang et al., 2023). Our experiment does not apply that decoding procedure to final predictions. Instead, inspired by repeated generation, we compare three outputs for the same input and use the observed variation to revise the prompt. When the output is auxiliary text for a regressor, additional rules may also amplify domain bias. We therefore compare prompt revisions and this consistency-based revision to test whether more complex generation strategies improve downstream regression.

2.4 Regression Fine-Tuning of Pretrained Encoders

We compare bert-base-chinese (Devlin et al., 2019), hfl/chinese-roberta-wwm-ext (Liu et al., 2019; Cui et al., 2020), and hfl/chinese-macbert-base (Cui et al., 2020). All three models use Transformer encoders (Vaswani et al., 2017), take the [CLS] representation, and feed it into the same dropout and two-output linear regression head. This design separates backbone differences from prompt and data differences.

2.5 Domain Transfer and Prompt Bias

The training signals in this task come from three distributions: general sentences and web texts from Chinese EmoBank, medical self-reflection texts released by the ROCLING 2025 shared task (Lee et al., 2025), and new-immigrant life narratives from the current task. Extra data can reduce estimation variance, but may also introduce source-domain vocabulary and affective patterns. When prompts are revised from previous-task examples, domain bias may remain in natural-language rules and becomes easier to inspect than ordinary model parameters.

We therefore distinguish between whether a prompt is more detailed and whether it is suitable for the target domain. If a prompt asks the model to focus on illness, general wards, ventilators, or caregivers, it may appear precise on the source data while missing family concerns, work expectations, and adaptation pressure in new-immigrant texts. This perspective shapes our interpretation of negative results: later prompt versions failing does not mean prompt engineering is useless; it more likely indicates a mismatch between the data used for prompt revision and the target domain.

3 Method

Figure 1 illustrates the final system. The workflow consists of constructing LLM-generated text in advance and fine-tuning a BERT regression model. The main submission uses text generated by the base prompt (v1) and trains bert-base-chinese for 3 epochs. The LLM expresses implicit affective information in text, while BERT maps the original and generated text to the task scale.

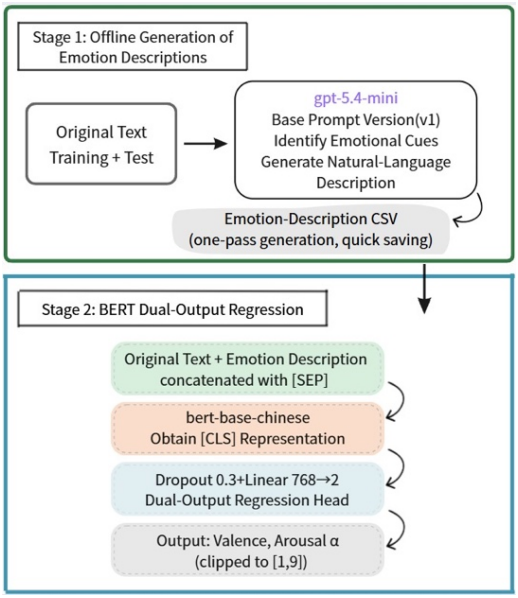

Figure 1. SCUNLP system workflow: enhancing BERT dual-output regression with LLM-generated text.

3.1 Problem Definition

Given a Chinese text x, the model jointly predicts valence and arousal y=(v,a), both real-valued scores on a 1-9 scale. We use a shared encoder and a two-output regression head so that the two dimensions share semantic representations. At inference time, predictions are clipped to [1,9] to match the official output range.

3.2 LLM-Generated Text

We call the first prompt the base prompt (v1). This version uses gpt-5.4-mini with temperature set to 0. The prompt asks the model to identify explicit and implicit affective information and the speaker's affective state, and to return plain text without Markdown formatting or heading symbols. The actual

prompt used in the experiments was written in Chinese; Figure 2 presents a larger English translation together with the smaller original Chinese prompt. In the figure, [original Chinese text] represents the input placeholder {text}, and 'Affective features:' translates the final output cue.



Figure 2. The base prompt (v1), shown as an English translation in larger type and the original Chinese prompt in smaller type. The original Chinese prompt was used in the experiments.

The LLM does not return VA scores, but instead produces readable natural-language text. The training and test sets are processed once and stored as CSV files; all generations succeed and contain no empty values. The output produced by the base prompt is much longer than the original text, averaging about 460 characters for training and 570 characters for test, compared with about 45 and 71 characters for the original texts.

There are two reasons for not allowing the LLM to decide the scores directly. First, the generated text can identify evidence such as rhetorical questions, negation, expectation, or restraint, while scale calibration is learned from labeled data. Second, the generated content can be cached, inspected, and replaced rather than being treated as a final prediction. At the same time, long outputs or incorrect background assumptions may interfere with the encoder, so later experiments compare different prompt versions.

3.3 Feature-Augmented BERT Regression Model

The model input is "original text [SEP] LLM-generated text." The tokenizer pads or truncates the sequence to 256 tokens. bert-base-chinese outputs the [CLS] representation $h \in \mathbb{R}^{768}$, which is passed through dropout (0.3) and a linear layer to obtain $\hat{y} = (\hat{v}, \hat{a})$. The plain-text baseline uses a maximum length of 128 because it does not include LLM-generated text and its sequences are shorter; the augmented models use 256 tokens to preserve the additional content. RoBERTa and MacBERT variants use the same input format, [CLS]+dropout+linear regression head, and training settings, changing only the backbone.

$$L = \frac{1}{2N} \sum_{i} \left[(\hat{v}_i - v_i)^2 + (\hat{a}_i - a_i)^2 \right] \quad (1)$$

where (v_i, a_i) are the gold labels and $(\hat{v}_i, \hat{a}_i)$ are the predictions. This is the mean squared error averaged over N instances and both output dimensions, matching the default mean reduction used in training. The two dimensions therefore contribute equally, without a separate loss weight for arousal. The main submission uses bert-base-chinese with text generated by the base prompt, AdamW, a learning rate of 2e-5, batch size 16, and 3 epochs. We do not use a learning-rate scheduler or early stopping.

Because all LLM-generated text is stored in CSV files, every encoder training run reads the same content, and official-test inference does not call the LLM again. This allows different backbones to be compared using identical LLM-generated inputs while limiting API costs to one-time data construction. We also use fixed-weight linear combinations of BERT, RoBERTa, and MacBERT predictions as an ablation. The weights are manually fixed rather than learned from validation data: valence uses 0.3/0.3/0.4 and arousal uses 0.3/0.4/0.3 for BERT/RoBERTa/MacBERT.

3.4 Prompt Variants

In addition to the base prompt (v1), we retain four increasingly elaborate variants. For the single-round revision (v2), the system generates outputs for a shuffled batch of 300 labeled ROCLING 2025 texts; up to the first 20 text-label-output cases are assembled for prompt revision, and the assembled case block is truncated to 3,000 characters. The three-round revision (v3) repeats this procedure for three rounds, using the next batch of 300 texts in each round. The consistency-based revision (v4) starts from the three-round prompt, generates three outputs for each of the 200 current-task validation texts, and supplies a 3,000-character excerpt of the collected comparisons to one final prompt-revision call. It uses repeated outputs only to revise the prompt and does not aggregate final predictions. The new-immigrant background version (v5) finally adds the statement that the texts are written by Taiwanese new immigrants. Versions v2-v5 also limit max_completion_tokens to 300, resulting in average output lengths of about 320-350 characters.

These versions do not merely add general rules. The saved v4 prompt still contains many criteria related to "illness," "general wards," "ventilators," and "caregivers" from the previous medical self-reflection corpus. Although v5 adds the new-immigrant background, it still inherits the same long prompt. This creates a clear cross-domain risk, whose impact is analyzed in Section 5.

3.5 Training and Inference

The data are first converted to numeric labels and rows with unparseable labels are removed. We then use random_state=42 for a fixed 80/20 split. Model parameters are updated only on the 80% training subset, while the 20% subset is used for internal checking. We do not save the best checkpoint according to validation performance; after each epoch, the final weights are directly used to generate official test submissions.

The training scripts do not fully fix all PyTorch, NumPy, and data-loader random states, and we do not repeat experiments over multiple seeds. Therefore, small differences among settings are treated as single-run development results rather than statistically significant findings. This limitation also applies to prompt and backbone comparisons.

4 Experimental Setup

4.1 Datasets

Our experiments use the data released by the organizers of the ROCLING 2026 shared task (Lin et al., 2026). The released Chinese EmoBank resources (Lee et al., 2022) span four granularities: words (CVAW), phrases (CVAP), sentences (CVAS), and texts (CVAT). We use the sentence-level (CVAS, 2,583 instances) and text-level (CVAT, 2,971 instances) portions for training. The current task additionally provides 200 new-immigrant validation texts and 1,100 test texts.

Data	Use	Instances
CVAS	Training	2,583
CVAT	Training	2,970
DSA-MST 2025	Training	1,541
DSA-NIFT validation	Development	200
DSA-NIFT test	Evaluation	1,100

Table 1. Summary of the datasets used in our experiments.

To construct training supervision, we combine CVAS, CVAT, and 1,541 labeled instances released by the ROCLING 2025 shared task (DSA-MST; Lee et al., 2025). Numeric conversion removes one unparseable CVAT row, resulting in 7,094 training instances. With the fixed split, the training and internal validation subsets contain 5,675 and 1,419 instances, respectively.

4.2 Evaluation Metrics

Following the official evaluation protocol of the ROCLING 2026 shared task, we compute mean absolute error (MAE) and Pearson correlation coefficient (PCC) independently for valence and arousal. Let $a_i \in A$ denote the gold value of the i-th instance, $p_i \in P$ denote the model prediction, and n denote the number of test instances. The two metrics are defined as follows.

MAE measures the average distance between predicted and gold values on the 1-9 scale; lower values indicate smaller absolute prediction error:

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|a_i - p_i| \tag{2}$$

PCC measures the linear correlation between predicted and gold values; higher values indicate that the model better preserves the relative affective ordering among instances:

$$\mathrm{PCC} = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{a_i-\mu_A}{\sigma_A}\right)\left(\frac{p_i-\mu_P}{\sigma_P}\right) \tag{3}$$

Here μ_A and σ_A are the mean and standard deviation of the gold-value set A, while μ_P and σ_P are the mean and standard deviation of the prediction set P. We use MAE to examine absolute error and PCC to examine relative ordering. Because the two metrics reflect different aspects of model behavior, we report both metrics for valence and arousal rather than interpreting performance from a single score.

4.3 Implementation Details

Our LLM-augmented models use a maximum sequence length of 256, batch size 16, the AdamW optimizer, and a learning rate of 2e-5. The regression head consists of dropout 0.3 followed by a 768-to-2 linear layer. The final system uses bert-base-chinese with text generated by the base prompt.

We compare training for 3, 5, and 8 epochs for the base-prompt version, and fix 3 epochs for the main comparisons of prompt versions (v2-v5) and the three backbones. All experiments are run on a single GPU in Google Colab.

Item	Setting
LLM	gpt-5.4-mini, temperature=0
Input	Original text [SEP] LLM-generated text; max length 256
Regression head	Dropout 0.3; Linear 768→2
Final system	AdamW; lr 2e-5; batch 16; 3 epochs
Output	Valence/Arousal clipped to [1,9]

Table 2. Main implementation settings of the final BERT system using LLM-generated text.

4.4 Data Usage Boundaries

We distinguish the roles of different text batches. Prompt revisions v2-v4 use the publicly released text and labels from ROCLING 2025, and these instances are also included in regression training as external public supervision. The v4 consistency-based step additionally reads the 200 validation texts from the current task, generates three outputs for each text, and uses their variation only for prompt revision; it does not use the validation labels or aggregate final predictions. The current test texts are used only to generate LLM text and perform ordinary inference, and are not fed back into the prompt or model parameters.

5 Results and Analysis

5.1 Main Results

The main result shows that LLM-generated text primarily improves valence prediction. We first compare three major development stages to examine the effects of additional supervised data and LLM-generated text. When BERT is trained only on CVAS and CVAT, valence MAE is 0.63. After adding the 1,541 labeled instances released by the ROCLING 2025 shared task, valence MAE drops to 0.61. Adding text produced by the base prompt further reduces valence MAE to 0.58, and valence PCC increases from 0.85 to 0.88. The improvement for arousal is smaller: MAE decreases from 0.92 to 0.90, and PCC increases from 0.38 to 0.41. The final system obtains valence MAE/PCC of 0.58/0.88 and arousal MAE/PCC of 0.90/0.41 on the official test set, ranking second overall with an average rank of 2.25. Overall, the generated text is most useful for supplementing semantic information about affective polarity, while its contribution to arousal ranking is more limited.

Setting	V-MAE↓	V-PCC↑	A-MAE↓	A-PCC↑
EmoBank only	0.63	0.85	0.92	0.38
+ 2025	0.61	0.86	0.92	0.39
+ base-prompt LLM	0.58	0.88	0.90	0.41

Table 3. Three development stages of bert-base-chinese (3 epochs); scores are from single official test evaluations.

5.2 Prompt Variant Analysis

To test whether more complex prompts produce more effective textual input, we compare several prompt revisions. Table 4 shows that longer and more explicit prompts do not lead to lower valence MAE. The single-round revision (v2) and three-round revision (v3) both obtain valence MAE of 0.62; the consistency-based revision (v4) obtains 0.61; and the new-immigrant background version (v5) improves to 0.60, still behind the base prompt at 0.58. Because v2-v5 use a 300-token generation limit, their average output lengths fall around 320-350 characters, much shorter than the base prompt. However, valence MAE does not improve, suggesting that output length alone cannot explain model performance.

Version	V-MAE↓	V-PCC↑	A-MAE↓	A-PCC↑
v1 base prompt	0.58	0.88	0.90	0.41
v2 single revision	0.62	0.87	0.89	0.40
v3 three revisions	0.62	0.86	0.89	0.35
v4 consistency-based revision	0.61	0.87	0.88	0.35
v5 immigrant context	0.60	0.87	0.89	0.40

Table 4. Prompt version comparison (bert-base-chinese, 3 epochs, with 2025 labels).

From the prompt content, the main issue in later versions appears not to be insufficient rules, but misplaced rule sources. The consistency-based revision retains criteria from the previous medical corpus, including illness, caregivers, and response intensity. Such rules are reasonable for medical self-reflection, but may not correspond to family concerns, work expectations, and adaptation pressure in new-immigrant texts. Although the new-immigrant background version appends target-domain context, it does not fully offset the earlier medical-oriented rules and therefore only partially recovers the performance loss. Notably, the arousal PCC of multi-round revision versions (v3 and v4) also drops from 0.41 to 0.35, consistent with the stagnation in valence and suggesting that source-domain rules may interfere with arousal

ranking.

5.3 Backbone and Other Ablations

If LLM-generated text is the main source of additional signal, simply replacing the encoder may not provide a stable improvement. For model architecture, we further compare different Chinese encoders using the same base-prompt output. Table 5 shows that, under the 3-epoch setting, BERT achieves the lowest valence MAE (0.58), compared with 0.62 for RoBERTa and 0.60 for MacBERT; all three obtain valence PCC of 0.88. MacBERT has a lower arousal MAE (0.87), and its arousal PCC equals BERT at 0.41, indicating a trade-off between the two dimensions rather than a uniformly superior backbone. Since this study uses valence MAE as the main selection criterion and BERT is not clearly worse on the other metrics, we retain bert-base-chinese for the final system.

Epoch and ensemble experiments lead to the same conclusion. For the base-prompt BERT, training for 3 epochs gives the lowest valence MAE among the 3-, 5-, and 8-epoch settings. Although the 5-epoch setting gives a lower arousal MAE, the preferred setting differs by metric. Fixed-weight ensembling also does not provide a stable advantage: the BERT/RoBERTa/MacBERT linear combination with 2025 labels obtains valence MAE of 0.62, behind the single base-prompt BERT at 0.58. We therefore treat ensembling and alternative epoch numbers as auxiliary ablations rather than final submission settings.

Backbone	V-MAE↓	V-PCC↑	A-MAE↓	A-PCC↑
BERT	0.58	0.88	0.90	0.41
RoBERTa	0.62	0.88	0.93	0.38
MacBERT	0.60	0.88	0.87	0.41

Table 5. Complete official test metrics of three backbones under the 3-epoch setting.

5.4 Metric-Level and Qualitative Analysis

Finally, the metric-level pattern shows that the model is better at capturing affective polarity than ranking arousal. Across the reported settings, valence PCC remains high at 0.86-0.88, while arousal PCC ranges from 0.35 to 0.41 and arousal MAE ranges from 0.87 to 0.93. New-immigrant texts often involve responsibility, adaptation, and resignation, which may imply negative valence without necessarily indicating high activation. If the LLM-generated text maps event seriousness directly to high arousal, it may disrupt arousal ranking.

The generated outputs show a similar pattern. For example, the training sentence "This is the most amazing performance I have ever watched" has labels of valence 7.00 and arousal 7.75. The text produced by the base prompt explicitly identifies superlative evaluation, admiration, shock, and excitement, separating positive affect from high intensity; such information is useful for the regressor. In contrast, later prompts often apply fixed rules before explaining the text. Their outputs are more regular, but may introduce unnecessary medical or background assumptions.

Instance-level gold labels for the current test set were not available to us. We therefore do not conduct error grouping by negation, rhetorical questions, family events, or long texts. Our interpretation is limited to the available aggregate metrics and generated files. If instance-level labels become available, future analysis should first examine high/low arousal confusions and sequence truncation, rather than simply adding more prompt clauses.

6 Limitations and Future Work

The limitations of this study mainly concern data, generated content, and evaluation. First, LLM-generated text is much longer than the original text. Under the base prompt, the average output lengths are about 460 and 570 characters for training and test data, respectively, while the model input is truncated to at most 256 tokens. Longer generated text can supplement implicit affective information, but also increases truncation risk, so parts of the original or generated content may not enter the encoder. In addition, the plain-text baseline uses a 128-token limit whereas the augmented model uses 256 tokens; consequently, the 0.61-to-0.58 comparison does not isolate generated text from the possible effect of the longer input limit. Future work should compare both inputs under the same length limit and can also control output length or encode the original and generated text separately before fusion.

Second, prompt revision is not a fully neutral form of general optimization. Versions v2-v4 use medical self-reflection data from the previous shared task for prompt adjustment, so some rules carry source-domain context, such as illness, caregivers, and response intensity. These rules may be appropriate for medical texts but not for new-

immigrant life narratives. Future work should separate source-domain and target-domain prompt design and reserve validation data that are not involved in prompt revision, so that source-task rules are not mistaken for general affective knowledge.

Third, the reported results come from single training runs and official feedback during shared-task development. We do not verify small differences with multiple seeds or confidence intervals. A small number of duplicate texts is also retained, and fixed ensemble weights are not recalibrated on an independent validation set. Future work should fix all random states, repeat training, and report averaged performance. Since arousal is clearly more difficult than valence, dimension-specific heads, variance-aware loss weighting, or auxiliary objectives that better preserve affective-intensity ranking are also worth exploring.

7 Conclusion

This paper describes the SCUNLP system for the ROCLING-2026 shared task on Chinese dimensional sentiment analysis. We use gpt-5.4-mini to analyze affective information in each text, concatenate its generated text with the original text, and fine-tune bert-base-chinese for joint valence-arousal regression. With 7,094 training instances, 256-token inputs, and 3 training epochs, the final system obtains valence MAE/PCC of 0.58/0.88 and arousal MAE/PCC of 0.90/0.41 on the official test set, ranking second overall.

Overall, LLM-generated text can supplement dispersed or implicit affective information in the original text, especially for valence prediction. However, multi-round prompt refinement, consistency-based prompt revision, new-immigrant background prompting, alternative backbones, and fixed-weight ensembling do not consistently outperform the base prompt. This suggests that, in limited-annotation settings, the usefulness of LLM-generated text depends not only on the amount of information it contains, but also on whether the prompt matches the target domain. Future work will focus on controlling output length, reducing cross-domain prompt bias, and improving arousal modeling and reproducibility. Taken together, these findings suggest that the effective role of LLMs in low-resource Chinese dimensional sentiment analysis may not be to directly generate scores, but to provide additional textual information that can be used by a supervised model.

Acknowledgments

This research was partially supported by the National Science and Technology Council, Taiwan (Grant Nos.: NSTC 115-2813-C-031-079-E and NSTC 115-2813-C-031-080-E).

References

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. In Advances in Neural Information Processing Systems (NeurIPS).

Sven Buechel and Udo Hahn. 2017. EmoBank: Studying the Impact of Annotation Perspective and Representation Format on Dimensional Emotion Analysis. In Proceedings of EACL, pages 578–585. https://doi.org/10.18653/v1/E17-2092

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Shijin Wang, and Guoping Hu. 2020. Revisiting Pre-Trained Models for Chinese Natural Language Processing. In Findings of EMNLP, pages 657–668. https://doi.org/10.18653/v1/2020.findings-emnlp.58

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of NAACL-HLT, pages 4171–4186. https://doi.org/10.18653/v1/N19-1423

Cheng-Yu Hsieh, Chun-Liang Li, Chih-kuan Yeh, Hootan Nakhost, Yasuhisa Fujii, Alex Ratner, Ranjay Krishna, Chen-Yu Lee, and Tomas Pfister. 2023. Distilling Step-by-Step! Outperforming Larger Language Models with Less Training Data and Smaller Model Sizes. In Findings of the Association for Computational Linguistics: ACL 2023, pages 8003–8017. https://doi.org/10.18653/v1/2023.findings-acl.507

Lung-Hao Lee, Jian-Hong Li, and Liang-Chih Yu. 2022. Chinese EmoBank: Building Valence-Arousal Resources for Dimensional Sentiment Analysis. ACM Transactions on Asian and Low-Resource Language Information Processing, 21(4): Article 65. https://doi.org/10.1145/3489141

Lung-Hao Lee, Tzu-Mi Lin, Hsiu-Min Shih, Kuo-Kai Shyu, Anna S. Hsu, and Peih-Ying Lu. 2025. ROCLING-2025 Shared Task: Chinese Dimensional Sentiment Analysis for Medical Self-Reflection Texts. In Proceedings of the 37th Conference on Computational Linguistics and Speech Processing (ROCLING

2025), pages 375–380. https://aclanthology.org/2025.rocling-main.41/

Tzu-Mi Lin, Wen-Ni Liu, Yen-Chen Tang, Lung-Hao Lee. 2026. ROCLING-2026 Shared Task: Chinese Dimensional Sentiment Analysis for New Immigrants' Feeling Texts. In Proceedings of the 38th Conference on Computational Linguistics and Speech Processing.

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. Computing Research Repository, arXiv:1907.11692. https://arxiv.org/abs/1907.11692

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. 2023. Self-Refine: Iterative Refinement with Self-Feedback. In Advances in Neural Information Processing Systems (NeurIPS). https://doi.org/10.52202/075280-2019

James A. Russell. 1980. A circumplex model of affect. Journal of Personality and Social Psychology, 39(6):1161–1178. https://doi.org/10.1037/h0077714

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. In Advances in Neural Information Processing Systems (NeurIPS). https://papers.nips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-Consistency Improves Chain of Thought Reasoning in Language Models. In International Conference on Learning Representations (ICLR). https://openreview.net/forum?id=1PL1NIMMrw

Qinyuan Ye, Mohamed Ahmed, Reid Pryzant, and Fereshte Khani. 2024. Prompt Engineering a Prompt Engineer. In Findings of the Association for Computational Linguistics: ACL 2024, pages 355–385. https://doi.org/10.18653/v1/2024.findings-acl.21

Liang-Chih Yu, Lung-Hao Lee, Shuai Hao, Jin Wang, Yunchao He, Jun Hu, K. Robert Lai, and Xuejie Zhang. 2016. Building Chinese Affective Resources in Valence-Arousal Dimensions. In Proceedings of NAACL-HLT, pages 540–545. https://doi.org/10.18653/v1/N16-1066