CYUT at ROCLING-2026 Shared Task: Perplexity-Guided Target-Domain Adaptation for Chinese Dimensional Sentiment Regression

Yi-Min Jian, Shih-Hung Wu*

Department of Computer Science and Information Engineering, Chaoyang University of Technology
Wufeng, Taichung, Taiwan

s11427601@gm.cyut.edu.tw, shwu@cyut.edu.tw

Abstract

This paper describes the CYUT system developed for the ROCLING-2026 Shared Task on Chinese dimensional sentiment regression. The target data consist of reflective narratives written by new immigrants, and these narratives differ substantially from the existing source corpora in text length, topic, and language use. We propose a perplexity-guided target-domain adaptation pipeline. We first fine-tune Qwen2-7B-Instruct and Ministral-3-8B-Instruct-2512 on 5,509 annotated instances from CVAS and CVAT. We then use each source model to calculate the perplexity of 200 unlabeled target-domain texts and select either the Top 30%, 40%, 50%, or all target texts for continued language-model training. After target-domain adaptation, we realign each model with the source-domain VA task and construct a dimension-specific weighted ensemble. The final Top-30% ensemble achieved a V-MAE of 0.5366, a V-PCC of 0.8971, an A-MAE of 0.8658, and an A-PCC of 0.4376 on the official test set, achieving first place on all four evaluation metrics and first place overall with an average rank of 1.

Keywords: dimensional sentiment regression, valence, arousal, perplexity, target-domain adaptation, continued pretraining, QLoRA, weighted ensemble

1 Introduction

1.1 Background and Task Challenges

Dimensional sentiment analysis (DSA) represents emotions in a continuous space. In this space, valence indicates the degree of pleasantness from negative to positive, whereas arousal indicates the degree of activation from calm to excited (Russell, 1980; Mehrabian, 1996). Compared with discrete sentiment classification, this representation can distinguish different levels of intensity and activation among emotions with the same polarity. EmoBank, the NRC VAD Lexicon, and Chinese EmoBank provide VAD or VA resources at the lexical and sentence levels (Buechel and Hahn, 2017; Mohammad, 2018; Lee et al., 2022).

The target texts in the ROCLING-2026 shared task consist of reflective narratives written by new immigrants (Lin et al., 2026). These texts often discuss family relationships, cultural adaptation, economic pressure, language barriers, homesickness, and related issues, with intertwined positive and negative emotions. The source dataset contains 5,509 instances from CVAS and CVAT, with an average length of 36.01 characters, whereas the development and test sets contain 200 and 1,100 texts with average lengths of 73.71 and 70.95 characters, respectively. These differences in text length, communicative purpose, and linguistic distribution require source-domain models to process unfamiliar vocabulary, syntactic patterns, and cross-sentence emotional cues.

This domain gap can also be viewed from the perspective of language contact. Reflective texts written by new immigrants may include loanwords, code-switching, non-standard spelling, and syntactic variation, resulting in a linguistic distribution different from existing Chinese sentiment corpora. Some of these features resemble language-contact phenomena discussed in studies of pidgins and creoles (Holm, 2000). Rather than classifying these texts as pidgin, we use the

*Corresponding author.

Proceedings of the 38th Conference on Computational Linguistics and Speech Processing (ROCLING 2026), pages 456–465
November 6-7, 2026 ©2026 Association for Computational Linguistics

language-contact perspective to explain the substantial domain mismatch that increases the difficulty of model adaptation.

1.2 Motivation and Method Overview

Previous studies on domain-adaptive and task-adaptive pretraining have shown that continued training on unlabeled domain-specific texts can improve downstream performance (Gururangan et al., 2020; Han and Eisenstein, 2019; Xie et al., 2024). However, this task provides only 200 unlabeled development texts, and directly using all available texts may not be the most effective strategy. We use the source model's perplexity to measure its unfamiliarity with each target-domain text and prioritize high-PPL samples for continued language-model training. A high PPL indicates that the vocabulary, syntax, or narrative pattern deviates substantially from the source distribution; it does not indicate that the text expresses stronger emotions.

The target-domain language-model training stage does not use VA labels or enforce the required JSON output format. Therefore, this stage may weaken the model's original regression capability. To address this problem, we adopt a three-stage pipeline that consists of source-task learning, target-language adaptation, and source-task realignment. We then combine Qwen and Ministral through a dimension-specific weighted ensemble to reduce model-specific bias.

1.3 Contributions

This paper makes the following contributions:

(1) We propose a target-domain adaptation pipeline that combines source-domain VA supervised fine-tuning, PPL-guided target-data selection, continued language-model training, and source-task realignment.

(2) We systematically compare the Top-30%, 40%, 50%, and 100% target-data settings and analyze how different selection ratios affect domain-adaptation performance.

(3) We propose a cross-backbone, dimension-specific weighted ensemble method that reduces model-specific bias and improves overall prediction stability.

2 Related Work

2.1 Dimensional Sentiment Analysis

Russell's circumplex model represents emotions through valence and arousal (Russell, 1980), whereas Mehrabian proposed the Pleasure–Arousal–Dominance framework (Mehrabian, 1996). Earlier studies used sentiment lexicons, distributional semantics, and regression models to predict dimensional emotion scores (Malandrakis et al., 2013; Wang et al., 2016). Later work introduced multi-task and neural models to jointly learn multiple emotion dimensions or map categorical emotions to continuous VAD representations (Buechel and Hahn, 2018; Wang et al., 2020; Park et al., 2021).

Chinese EmoBank provides Chinese VA annotations at the word, phrase, and sentence levels (Lee et al., 2022). The ROCLING 2025 shared task further showed that domain differences between source corpora and reflective texts make implicit emotions and mid-range scores difficult to predict (Lee et al., 2025; Huang and Shao, 2025). Existing systems used BERT, data augmentation, LLM embeddings, and ensemble regression (Jian et al., 2025; Li and Lin, 2025). In contrast, our method directly adapts language models with unlabeled target-domain texts.

2.2 Domain Adaptation for Language Models

Domain-adaptive and task-adaptive pretraining continue language-model training on unlabeled domain- or task-related texts to reduce source–target distribution mismatch (Gururangan et al., 2020; Xie et al., 2024). However, continued training may cause catastrophic forgetting and weaken the original task capability (Kirkpatrick et al., 2017). Because our target-adaptation stage does not use VA labels or JSON constraints, we add source VA realignment to restore regression and output-format capabilities.

2.3 Perplexity-Based Data Selection

The Moore–Lewis method ranks domain-relevant data using cross-entropy differences (Moore and Lewis, 2010), and later studies applied perplexity to bilingual-data and training-corpus selection (Axelrod et al., 2011; Toral,

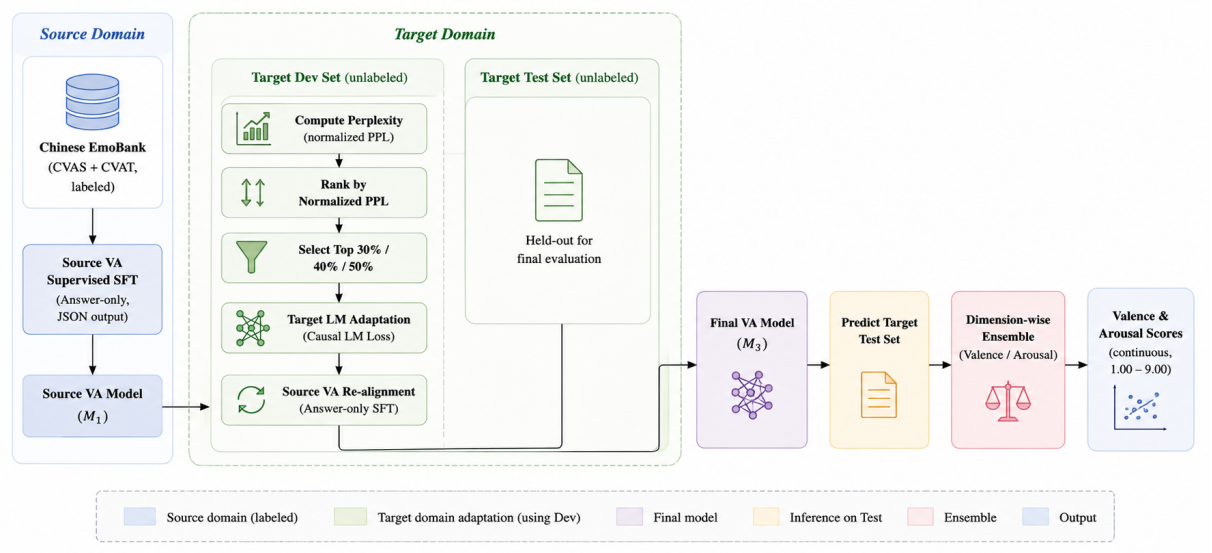

Figure 1: PPL-guided target-domain adaptation pipeline. The source model M_1 learns VA regression, high-PPL target texts are used for LM adaptation to obtain M_2, and source annotations realign the model as M_3. Predictions from different backbones are then combined through a weighted ensemble.

2013). Our method instead calculates target-text PPL with the source model and treats high PPL as an approximate signal of source–target mismatch. Accordingly, we prioritize high-PPL texts for target-domain adaptation.

2.4 Parameter-Efficient Fine-Tuning

LoRA enables task adaptation through low-rank trainable matrices (Hu et al., 2022), while QLoRA combines LoRA with 4-bit quantization to reduce memory requirements (Dettmers et al., 2023). We reuse the same LoRA adapter across source SFT, target LM adaptation, and source-task realignment. Figure 1 summarizes the complete pipeline.

3 Method

3.1 Overall Architecture

Figure 1 presents the system pipeline, which contains five stages: (1) source-domain VA supervised fine-tuning, (2) target-text PPL ranking, (3) PPL-guided target-domain LM adaptation, (4) source-domain VA realignment, and (5) cross-model weighted ensembling. For each $r \in \{30, 40, 50, 100\}$, Qwen and Ministral independently complete the entire pipeline.

3.2 Source-Domain VA Supervised Fine-Tuning

For each source-domain sample (x, y), x denotes the input text and y denotes the fixed-format VA output:

$$y = \texttt{\{"VA":"}v\#a\texttt{"\}}. \tag{1}$$

The model autoregressively generates y, and the training procedure calculates cross-entropy loss only for the assistant answer tokens:

$$\mathcal{L}_{\text{VA}} = -\sum_{t=1}^{|y|} \log p_\theta(y_t \mid x, y_{<t}). \tag{2}$$

We assign a label of -100 to the system instruction and user input tokens, so the loss only supervises the assistant answer tokens. The resulting source model M_1 serves as both the PPL scorer and the Source-only baseline.

3.3 Target-Domain Perplexity Estimation

For each unlabeled target-domain text $x = (w_1, \ldots, w_T)$, we use the source model M_1 to calculate its length-normalized negative log-likelihood:

$$\text{NLL}(x) = -\frac{1}{T-1} \sum_{t=2}^{T} \log p_{M_1}(w_t \mid w_{<t}). \tag{3}$$

Perplexity is defined as:

$$\text{PPL}(x) = \exp(\text{NLL}(x)). \tag{4}$$

A higher PPL indicates that M_1 is less familiar with the text. We rank the 200 target-domain texts in descending order of PPL and

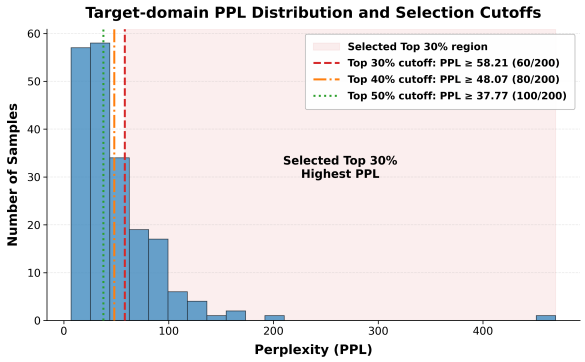

Figure 2: PPL distribution of the 200 unlabeled target-domain texts computed by the source Qwen model. Dashed lines indicate the Top 30%, 40%, and 50% selection cutoffs.

select the highest-ranked $r\%$ as the adaptation dataset:

$$\mathcal{D}_{\text{target}}^{(r)} = \text{Top}_{r\%}\left(\mathcal{D}_{\text{target}}; \text{PPL}\right). \quad (5)$$

We compare four ratios, namely 30%, 40%, 50%, and 100%, which correspond to 60, 80, 100, and 200 texts, respectively. The 100% target-adaptation setting serves as a comparison baseline that does not apply PPL-based filtering.

As Figure 2 shows, the target-domain PPL values calculated by Qwen follow a right-skewed distribution. Most texts have relatively low perplexity, whereas only a small number of samples differ substantially from the source distribution. The Top-30%, 40%, and 50% cutoff values are 58.21, 48.07, and 37.77, respectively.

3.4 PPL-Guided Target-Domain Adaptation

For each selected target-text set $\mathcal{D}_{\text{target}}^{(r)}$, we continue training the same LoRA adapter from M_1. This stage does not use VA annotations. Instead, it calculates the causal language-modeling loss over all input tokens:

$$\mathcal{L}_{\text{target}} = -\sum_{x \in \mathcal{D}_{\text{target}}^{(r)}} \sum_{t=2}^{|x|} \log p_\theta(w_t \mid w_{<t}). \quad (6)$$

This stage produces $M_2^{(r)}$. The model learns the vocabulary, syntax, and narrative characteristics of the target-domain distribution without generating VA pseudo-labels.

3.5 Source-Domain VA Realignment

Because $\mathcal{L}_{\text{target}}$ differs from the VA task objective, target-domain LM adaptation may weaken the model's numerical regression and format-following capabilities. Therefore, we perform one epoch of answer-only SFT on the complete source-domain VA dataset using $M_2^{(r)}$. This process produces the final model $M_3^{(r)}$:

$$M_3^{(r)} = \text{Realign}\left(M_2^{(r)}, \mathcal{D}_{\text{source}}\right). \quad (7)$$

This step reinforces the VA scale, numerical generation, and JSON output constraints using only source-domain annotations.

3.6 Weighted Ensemble

Qwen2 and Ministral differ in their pretraining corpora, tokenization procedures, and parameterization. Therefore, their prediction errors may be complementary. We construct dimension-specific weighted ensembles for Valence and Arousal using the test-set performance of the individual backbone models reported in Table 3. For model m and emotional dimension $d \in \{V, A\}$, we define the reliability score as:

$$S_{m,d} = \frac{\text{PCC}_{m,d}}{\text{MAE}_{m,d}}. \quad (8)$$

We calculate the final prediction by directly normalizing the reliability scores:

$$\hat{y}_{i,d}^{\text{ens}} = \sum_{m=1}^{M} \frac{S_{m,d}}{\sum_{j=1}^{M} S_{j,d}} \hat{y}_{i,d}^{(m)}. \quad (9)$$

For each selection ratio, the ensemble combines the predictions of Qwen2 and Ministral under the same ratio. We calculate separate weights for Valence and Arousal to account for both prediction error and correlation (Dietterich, 2000; Jian et al., 2025).

3.7 Inference

Both models use greedy decoding with `do_sample=False` and `max_new_tokens=64`. The system first attempts to parse the `VA` field from the generated JSON output. When JSON parsing fails, the system uses a regular expression to extract the `V#A` pattern. The system constrains each score to the interval $[1, 9]$ and rounds it to two decimal places. When all parsing procedures fail, the system returns `5.00#5.00`.

Dataset	Instances	Mean length	Median length
CVAS+CVAT training set	5,509	36.01	20
ROCLING development set	200	73.71	64
ROCLING test set	1,100	70.95	64

Table 1: Dataset sizes and character-length statistics.

Parameter	Source SFT	LM Adapt.	Re-align
Epochs	3	1	1
Learning rate	10^{-4}	10^{-5}	2×10^{-5}
Batch/GPU	2	1	2
Grad. accumulation	8	8	8
Effective batch	16	8	16
Max length	1024	1024	1024
Loss scope	Assistant	Full text	Assistant
Seed	42	42	42

Table 2: Shared hyperparameters across training stages.

4 Experimental Setup

4.1 Datasets

Table 1 summarizes the three data splits. The training dataset combines CVAS and CVAT and contains 5,509 instances. The development and test datasets contain 200 and 1,100 reflective texts written by new immigrants, respectively. The 200 development texts are unlabeled and are used as target-domain data for perplexity estimation, PPL-based selection, and continued language-model adaptation; they are not used for supervised evaluation. All MAE and PCC results reported in Tables 3–5 are obtained on the 1,100-text test set. The source-domain texts are substantially shorter on average than the target-domain texts. This difference indicates that the task contains distributional shifts in both text length and narrative structure.

The source-domain Valence and Arousal scores have means of 4.73 and 4.90 and standard deviations of 1.37 and 1.17, respectively. The correlation between the two dimensions is −0.20.

4.2 Hyperparameters

We use Qwen/Qwen2-7B-Instruct (Yang et al., 2024)[1] and mistralai/Ministral-3-8B-Instruct-2512[2] as the two backbone models. Both models use the same training pipeline, prompt content, generation settings, and hyperparameters, as shown in Table 2. We apply QLoRA with 4-bit NF4 quantization, double quantization, and bfloat16 computation. The LoRA configuration uses $r = 16$, $\alpha = 32$, dropout $= 0.05$, and targets `q_proj`, `k_proj`, `v_proj`, `o_proj`, `gate_proj`, `up_proj`, and `down_proj`.

[1] Qwen/Qwen2-7B-Instruct

[2] mistralai/Ministral-3-8B-Instruct-2512

4.3 Prompt Design

As Figure 3 shows, both backbone models use the same prompt during source SFT, source-task realignment, and inference. The system instruction defines reflective texts written by new immigrants, Russell's VA task, and the 1.00–9.00 scale. The user input appends the original text after `Text:`. The assistant returns the two scores with two decimal places in the format `{"VA":"V#A"}`. During training, the system calculates the loss only for the assistant response tokens. During testing, the model directly generates the prediction.

4.4 Evaluation Metrics

We evaluate Valence and Arousal using MAE and PCC:

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|a_i - p_i|, \tag{10}$$

$$\mathrm{PCC} = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{a_i-\mu_A}{\sigma_A}\right)\left(\frac{p_i-\mu_P}{\sigma_P}\right). \tag{11}$$

A lower MAE and a higher PCC indicate better performance.

5 Experimental Results

5.1 Overall Performance of Individual Models

Table 3 compares the Source-only SFT baseline and four PPL-TDA settings for the two backbone models on the test set. This table addresses two questions. First, it examines whether target-domain adaptation outperforms supervised fine-tuning that uses only source-domain annotations. Second, it examines whether the adaptation effect changes



Figure 3: Prompt templates used for training and inference. The left panel shows the Chinese version and the right panel shows the English translation. The upper panel presents training with a reference VA response, whereas the lower panel presents inference with unknown values to avoid revealing annotations.

with the proportion of selected high-PPL samples. Every PPL-TDA model first performs target-domain language-model training and subsequently performs source-domain VA realignment. Therefore, the differences in the table reflect the complete adaptation pipeline rather than LM adaptation alone.

For Qwen2-7B, the Top-30% PPL-TDA setting achieves the best Valence performance. Compared with Source-only SFT, this setting reduces V-MAE from 0.5699 to 0.5496 and increases V-PCC from 0.8838 to 0.8932. It also increases A-PCC from 0.3887 to 0.4343. However, A-MAE increases from 0.8793 to 0.9005. This result indicates that the setting improves the relative ranking capability for Arousal but does not simultaneously improve absolute numerical calibration. In other words, the model becomes more capable of determining which texts express higher arousal, but its predicted values remain shifted relative to the reference scale.

When the Qwen selection ratio increases to 40%, 50%, or 100%, the improvement in Valence gradually decreases. These results suggest a conservative conclusion: among the ratios evaluated in this study, the 60 texts with the highest PPL values contain the most discriminative target-domain signals for Qwen. Adding lower-PPL samples does not provide further improvements. This finding does not imply that 30% is a universally optimal ratio for other datasets. Instead, it shows that PPL-based ranking affects the efficiency with which a limited amount of target-domain data is used.

Ministral-3-8B exhibits a different trend. Its Source-only SFT baseline already achieves a V-PCC of 0.8938 and an A-PCC of 0.4394, which are better than the results of most PPL-TDA settings. The Top-50% PPL-TDA setting only slightly reduces V-MAE to 0.5481. This result demonstrates that the benefit of domain adaptation depends on the backbone model. When a source model already represents the target-domain texts adequately, additional continued training may not produce consistent improvements and may even intro-

Backbone	Adaptation setting	Valence		Arousal	
		MAE↓	PCC↑	MAE↓	PCC↑
Qwen2-7B	Source-only SFT	0.5699	0.8838	**0.8793**	0.3887
	Top-30% PPL-TDA	**0.5496**	**0.8932**	0.9005	**0.4343**
	Top-40% PPL-TDA	0.5643	0.8887	0.8865	0.3969
	Top-50% PPL-TDA	0.5665	0.8872	0.8822	0.3983
	100% target PPL-TDA	0.5645	0.8878	0.8823	0.3986
Ministral-3-8B	Source-only SFT	0.5509	**0.8938**	**0.8863**	**0.4394**
	Top-30% PPL-TDA	0.5639	0.8880	0.8842	0.3958
	Top-40% PPL-TDA	0.5502	0.8929	0.8991	0.4333
	Top-50% PPL-TDA	**0.5481**	0.8932	0.8989	0.4382
	100% target PPL-TDA	0.5497	0.8935	0.8991	0.4343

Table 3: Test-set results of individual backbone models. Bold indicates the best result within each backbone.

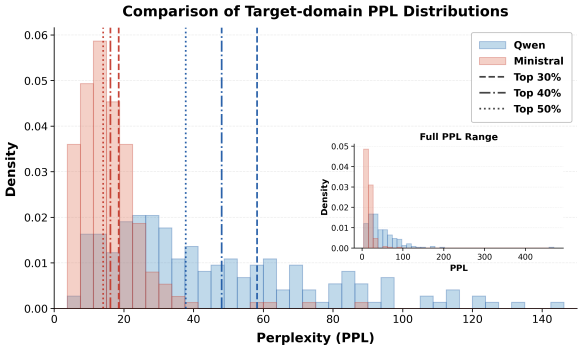

Figure 4: Target-domain PPL distributions of Qwen and Ministral. Line styles denote the Top 30%, 40%, and 50% cutoffs; the inset shows the full range.

Target-text subset	Valence		Arousal	
	MAE↓	PCC↑	MAE↓	PCC↑
Top 30%	0.5366	0.8971	0.8658	0.4376
Top 40%	**0.5366**	**0.8972**	0.8661	0.4383
Top 50%	0.5373	0.8969	**0.8639**	**0.4404**
100% target	0.5372	0.8971	0.8656	0.4387

Table 4: Test-set results of weighted ensembles under different target-data ratios.

duce calibration shifts.

5.2 PPL Distributions and Backbone Differences

Figure 4 presents the PPL distributions calculated by the two source models for the same target-domain texts. The Qwen distribution is more dispersed and contains a pronounced right tail. This pattern indicates that the linguistic modeling difficulty of the target-domain texts varies substantially for Qwen. In contrast, the Ministral distribution is concentrated within a lower and narrower range, which indicates that its prediction difficulty is more consistent across most target-domain texts.

This difference in distribution is consistent with the results in Table 3. For Qwen, the high-PPL samples provide stronger discrimination. Therefore, continued training on the Top-30% texts is more likely to supplement target-domain linguistic patterns that the model has not adequately covered. For Ministral, the PPL distribution is more concentrated, which indicates that the linguistic-difficulty differences among the ranked subsets are limited. Consequently, it is more difficult to obtain stable improvements from a particular selection ratio. The absolute PPL values of different models are affected by their tokenizers, vocabulary segmentation, and probability scales. Therefore, this paper compares distributional shapes and the discriminative capacity of sample rankings rather than using absolute PPL values to directly compare model capability.

5.3 Cross-Backbone Weighted Ensemble

Table 4 reports the test-set performance of the four weighted ensembles. For each selection ratio, the Qwen2 and Ministral test-set scores in Table 3 are converted into dimension-specific reliability weights using Equation (8), and these weights are then applied to their predictions using Equation (9). Table 4 shows that all four ensembles outperform the best

Team	V-MAE	V-PCC	A-MAE	A-PCC
CYUT	**0.5366 (1)**	**0.8971 (1)**	**0.8658 (1)**	**0.4376 (1)**
SCUNLP	0.5793 (2)	0.8815 (2)	0.8951 (2)	0.4050 (3)
aof123	0.6018 (3)	0.8688 (3)	0.9403 (7)	0.4052 (2)
DonkiDonki	0.6200 (4)	0.8663 (4)	0.9259 (4)	0.3566 (6)
BEDO	0.6418 (5)	0.8619 (5)	0.8992 (3)	0.3204 (7)
wuhongyi	0.6495 (6)	0.8586 (6)	0.9294 (5)	0.3366 (6)
845310	0.6684 (7)	0.8542 (7)	0.9341 (6)	0.3581 (5)

Table 5: Official ROCLING-2026 test results. Numbers in parentheses indicate metric ranks. CYUT ranked first overall with an average rank of 1.00.

individual model in V-MAE. Although Qwen and Ministral perform differently across dimensions and adaptation settings, their prediction errors remain complementary. Consequently, combining the two backbones consistently improves overall performance. The ensemble results range from 0.5366 to 0.5373, whereas the best individual model achieves 0.5481.

The differences among the four ratios are small. Top-40% achieves the best V-PCC and ties for the best V-MAE, while Top-50% achieves the best A-MAE and A-PCC. The Top-30% ensemble was used for the official submission. The additional ratios provide a post-hoc comparison and suggest that ensemble performance is not highly sensitive to the selection ratio.

For each dimension, we define model reliability as PCC/MAE and normalize the resulting scores into ensemble weights. This approach gives greater influence to models with lower error and higher correlation, while allowing Qwen and Ministral to contribute differently to Valence and Arousal (Jian et al., 2025).

Tables 3 and 4 both report results on the 1,100-text test set. Table 3 presents the performance of the individual backbone models, and these scores are used to derive the dimension-specific ensemble weights reported in Table 4. Table 5 separately presents the official ROCLING-2026 leaderboard for the submitted Top-30% ensemble. Therefore, the Top-30% results in Table 4 are identical to the CYUT results in Table 5.

The CYUT system achieves a V-MAE of 0.5366, a V-PCC of 0.8971, an A-MAE of 0.8658, and an A-PCC of 0.4376. The submitted system ranked first on all four evaluation metrics and achieved the best overall ranking with an average rank of 1.00. These results show that PPL-guided target-domain adaptation does not consistently improve every backbone model or every emotional dimension. Nevertheless, source-task realignment and cross-backbone ensembling allow the complete system to achieve stable and competitive performance.

6 Discussion and Conclusion

This paper presents a perplexity-guided target-domain adaptation pipeline that combines source-domain VA fine-tuning, PPL-guided target-text selection, continued LM adaptation, source-task realignment, and cross-backbone ensembling. The loanwords, code-switching, and syntactic variation found in new-immigrant narratives provide one possible explanation for the distribution mismatch between the source and target corpora from the perspective of language contact.

The results show that the benefit of PPL-guided adaptation depends on the backbone model. Qwen improves under the Top-30% setting, whereas Ministral already performs strongly with Source-only SFT and gains less from additional adaptation. The ensemble reduces V-MAE from 0.5481 to 0.5366, a relative improvement of 2.10%, and reduces A-MAE from 0.8793 to 0.8639, an improvement of 1.75%. The submitted Top-30% ensemble achieved first place on all four official evaluation metrics and ranked first overall.

This study uses only one random seed, evaluates four fixed selection ratios, and does not isolate the effect of source-task realignment. Future work will investigate multi-model PPL aggregation, embedding distance, model uncertainty, and dynamic data selection.

Overall, the results suggest that perplexity provides a practical signal for target-domain adaptation when labeled target-domain data are limited.

Acknowledgment

This study was supported by the National Science and Technology Council under Grant NSTC 115-2221-E-324-014.

References

Amittai Axelrod, Xiaodong He, and Jianfeng Gao. 2011. Domain adaptation via pseudo in-domain data selection. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 355–362, Edinburgh, Scotland, UK. Association for Computational Linguistics.

Sven Buechel and Udo Hahn. 2017. EmoBank: Studying the impact of annotation perspective and representation format on dimensional emotion analysis. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 578–585, Valencia, Spain. Association for Computational Linguistics.

Sven Buechel and Udo Hahn. 2018. Word emotion induction for multiple languages as a deep multi-task learning problem. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1907–1918, New Orleans, Louisiana. Association for Computational Linguistics.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. QLoRA: Efficient finetuning of quantized LLMs. In *Advances in Neural Information Processing Systems*, volume 36, pages 10088–10115.

Thomas G. Dietterich. 2000. Ensemble methods in machine learning. In *Multiple Classifier Systems*, volume 1857 of *Lecture Notes in Computer Science*, pages 1–15, Berlin, Heidelberg. Springer.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online. Association for Computational Linguistics.

Xiaochuang Han and Jacob Eisenstein. 2019. Unsupervised domain adaptation of contextualized embeddings for sequence labeling. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4238–4248, Hong Kong, China. Association for Computational Linguistics.

John Holm. 2000. *An Introduction to Pidgins and Creoles*. Cambridge University Press, Cambridge.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Sieh-Chuen Huang and Hsuan-Lei Shao. 2025. NTULAW at ROCLING-2025 shared task: Domain-adaptive modeling of implicit emotions in medical reflections. In *Proceedings of the 37th Conference on Computational Linguistics and Speech Processing (ROCLING 2025)*, pages 390–398, National Taiwan University, Taipei City, Taiwan. Association for Computational Linguistics.

Yi-Min Jian, An Yu Hsiao, and Shih-Hung Wu. 2025. CYUT-NLP at ROCLING-2025 shared task: Valence–arousal prediction in physicians' texts using BERT, RAG, and multi-teacher pseudo-labeling. In *Proceedings of the 37th Conference on Computational Linguistics and Speech Processing (ROCLING 2025)*, pages 381–389, National Taiwan University, Taipei City, Taiwan. Association for Computational Linguistics.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwińska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. 2017. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526.

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).

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, National Taiwan University, Taipei City, Taiwan. Association for Computational Linguistics.

Hsin-Chieh Li and Wen-Cheng Lin. 2025. TCU at ROCLING-2025 shared task: Leveraging LLM embeddings and ensemble regression for Chinese dimensional sentiment analysis. In *Proceedings of the 37th Conference on Computational Linguistics and Speech Processing (ROCLING 2025)*, pages 399–406, National Taiwan University, Taipei City, Taiwan. Association for Computational Linguistics.

Tzu-Mi Lin, Wen-Ni Liu, Yen-Chen Tang, and 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*. Association for Computational Linguistics.

Nikolaos Malandrakis, Alexandros Potamianos, Elias Iosif, and Shrikanth Narayanan. 2013. Distributional semantic models for affective text analysis. *IEEE Transactions on Audio, Speech, and Language Processing*, 21(11):2379–2392.

Albert Mehrabian. 1996. Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in temperament. *Current Psychology*, 14(4):261–292.

Saif Mohammad. 2018. Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 English words. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 174–184, Melbourne, Australia. Association for Computational Linguistics.

Robert C. Moore and William Lewis. 2010. Intelligent selection of language model training data. In *Proceedings of the ACL 2010 Conference Short Papers*, pages 220–224, Uppsala, Sweden. Association for Computational Linguistics.

Sungjoon Park, Jiseon Kim, and Jaegul Choo. 2021. Dimensional emotion detection from categorical emotion. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4367–4380, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

James A. Russell. 1980. A circumplex model of affect. *Journal of Personality and Social Psychology*, 39(6):1161–1178.

Antonio Toral. 2013. Hybrid selection of language model training data using linguistic information and perplexity. In *Proceedings of the Second Workshop on Hybrid Approaches to Translation*, pages 8–12, Sofia, Bulgaria. Association for Computational Linguistics.

Jin Wang, Liang-Chih Yu, K. Robert Lai, and Xuejie Zhang. 2016. Community-based weighted graph model for valence-arousal prediction of affective words. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(11):1957–1968.

Jin Wang, Liang-Chih Yu, K. Robert Lai, and Xuejie Zhang. 2020. Tree-structured regional CNN-LSTM model for dimensional sentiment analysis. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:581–591.

Yong Xie, Karan Aggarwal, and Aitzaz Ahmad. 2024. Efficient continual pre-training for building domain-specific large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 10184–10201, Bangkok, Thailand. Association for Computational Linguistics.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jianxin Yang, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Xuejing Liu, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, Zhifang Guo, and Zhihao Fan. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.