YCW-NTUST at ROCLING-2026 Shared Task: Leaderboard-Feedback Closed-Form Ensembling with LLM Direct Scoring for Chinese Valence–Arousal Regression

Yu-Cheng Chao*, **Chuan-Jhih You***, **Guan-Hua Wen**, and **Kuan-Yu Chen**
National Taiwan University of Science and Technology
{M11515019, M11515027, B11215005, kychen}@mail.ntust.edu.tw

Abstract

We describe YCW-NTUST's system for the ROCLING-2026 shared task on Chinese Valence–Arousal (VA) prediction for new-immigrant texts. The system combines fine-tuned Chinese encoders, frozen embedding regression, target-domain synthetic data, lexicon scoring, and LLM direct scoring. To combine signals without test labels, we derive a closed-form ensemble from leaderboard PCC feedback and apply a positive affine transformation to reduce MAE while preserving PCC. Our final test-phase submission achieves V-MAE 0.5555, V-PCC 0.8951, A-MAE 0.8159, and A-PCC 0.4438. A replay on labeled ROCLING 2025 data estimates that leaderboard adaptation inflates the Arousal gain by about 0.016 PCC and 0.008 MAE at the 2026 signal strength. Although the submission was not included in the official ranking, its scores are competitive with the top system, particularly on Arousal MAE.

1 Introduction

Dimensional sentiment analysis (DSA) represents affect with continuous dimensions rather than discrete categories, following the circumplex model of affect (Russell, 1980): Valence captures how positive or negative an emotion is, and Arousal how activated or calm. Unlike categorical labels, this representation preserves the intensity of an emotion.

Recent ROCLING shared tasks have kept the Valence–Arousal (VA) rating scheme of Chinese EmoBank (Lee et al., 2022) while moving toward domain-specific text. In the ROCLING 2025 task on physicians' self-reflection texts (Lee et al., 2025), participants found that general-domain resources such as Chinese EmoBank transfer poorly to the target domain, whose emotions are often implicit, low in arousal, and spread over several sentences (Jian et al., 2025; Huang and Shao, 2025).

The ROCLING 2026 task (Lin et al., 2026) moves to Chinese texts about new-immigrant experiences, which often combine family, work, identity, cross-cultural adaptation, and interactions with institutions in a single passage. Surface vocabulary therefore does not always indicate emotional intensity: a text about waiting for administrative procedures or missing one's hometown must be read through its events, the writer's circumstances, and implicit stance. Single models struggle most with Arousal: a fine-tuned RoBERTa regressor reaches an Arousal PCC of only 0.35 on the test set, against 0.86 for Valence.

We therefore treat VA prediction as a heterogeneous ensemble problem rather than tuning a single backbone. Our system combines fine-tuned Chinese RoBERTa (Cui et al., 2021) and MacBERT (Cui et al., 2020) regressors, frozen multilingual-e5 embeddings (Wang et al., 2024) with support vector regression (SVR; Drucker et al., 1997), a specialist trained on target-style synthetic data, a CVAW lexicon signal (Yu et al., 2016a), and large language models (LLMs) as direct VA scorers. These components differ in representation, supervision, and inductive bias.

A second challenge is that the test labels are hidden, so ensemble weights cannot be fitted on them in the usual way. The leaderboard, however, returns a PCC for every submitted prediction vector. A vector p_i with leaderboard PCC r_i and locally computable standard deviation $\sigma(p_i)$ reveals $\mathrm{cov}(p_i, y)/\sigma(y) = r_i\sigma(p_i)$, where y is the hidden gold vector, and the unknown $\sigma(y)$ cancels from the PCC of any linear mixture of submitted vectors. Ensemble weights can therefore be optimized offline instead of by trial-and-error submissions. This is a form of adaptive data analysis; Section 2 relates it to known leaderboard-adaptivity results (Blum and Hardt, 2015; Dwork et al., 2015; Whitehill, 2018), and Appendix E discusses its risks.

We further exploit the invariance of PCC to pos-

*These authors contributed equally to this work.

Proceedings of the 38th Conference on Computational Linguistics and Speech Processing (ROCLING 2026), pages 496–510
November 6-7, 2026 ©2026 Association for Computational Linguistics

itive affine transformations to lower MAE without changing PCC. Because the official metrics are computed per dimension, both steps are solved separately for Valence and Arousal.

Every ROCLING 2026 score in this paper is an official evaluation-service reading of a submission made before the test-phase deadline; the post-competition analyses (Sections 5.3, 5.4, and 5.5) involved no further ROCLING 2026 submissions. Our final submission is not listed in the official ranking (Section 5.2).

Our contributions are as follows:

1. A heterogeneous VA prediction system that combines fine-tuned Chinese encoders, frozen embedding regression, target-domain synthetic data, lexicon features, and LLM direct scoring.

2. An adaptation of leaderboard blending, previously used with error readings in the Netflix Prize (Töscher et al., 2009), to a PCC leaderboard: a closed-form linear ensemble of previously submitted prediction vectors that needs no test labels, followed by a PCC-preserving affine rescaling that lowers MAE, both solved per dimension.

3. A replay of the procedure on labeled ROCLING 2025 data that measures how much the use of test-set feedback overstates the gains, including at the weaker signal strength of the 2026 task.

4. Ablations computed exactly from the official readings, which show that LLM direct scoring supplies most of the ensemble's Arousal gain, and a three-seed check showing that 300 synthetic target-style texts consistently raise the proxy Arousal PCC of a RoBERTa regressor, although they do not reliably widen its predictions.

2 Related Work

Chinese dimensional sentiment resources and shared tasks. Yu et al. (2016a) released the CVAW lexicon and the CVAT text corpus, later extended into Chinese EmoBank (Lee et al., 2022) with sentence-level (CVAS) and text-level (CVAT) VA annotations. Shared tasks built on these resources have moved from words (Yu et al., 2016b) to phrases (Yu et al., 2017), educational texts (Yu et al., 2021), and medical self-reflection texts (Lee et al., 2025), and models range from neural networks with linguistic dependency features (Deng et al., 2022) to MacBERT regressors (Hung et al., 2021) and transformer fusion (Deng et al., 2023). ROCLING 2026 (Lin et al., 2026) adds new-immigrant narratives, a domain for which no annotated VA resource exists.

Heterogeneous ensembles for sentiment analysis. Combining models with different inductive biases is a long-standing strategy in sentiment analysis (Fersini et al., 2014; Araque et al., 2017), with stacked generalization (Wolpert, 1992) as the general framework. The three strongest ROCLING 2025 systems all combined several models (Lee et al., 2025): the winner (Jian et al., 2025) ensembled fine-tuned BERT models and augmented its training data with retrieval-augmented LLM generation and multi-teacher pseudo-labeling, Li and Lin (2025) combined frozen LLM embeddings with SVR at the encoder level, and Huang and Shao (2025) fused encoders of different granularity to model implicit emotion. Our synthetic-data specialist builds on the generation and pseudo-labeling scheme of Jian et al. (2025) and our frozen-embedding component follows Li and Lin (2025), but our ensemble weights are solved analytically rather than tuned on labeled validation data.

LLMs as direct scorers. Large language models (Brown et al., 2020; Qwen Team, 2025; Anthropic, 2026) have been used as zero-shot evaluators that emit numeric judgements, for example in translation quality assessment (Kocmi and Federmann, 2023), NLG evaluation (Liu et al., 2023), and general response judging (Zheng et al., 2023). We adopt the same paradigm for VA regression and, for the open-weight models, follow G-Eval (Liu et al., 2023) in reading a continuous score as the probability-weighted expectation over the digit tokens rather than from decoded text.

Leaderboard feedback and adaptivity. Repeatedly querying a hidden test set is a known source of adaptive overfitting: Dwork et al. (2015) formalize the risk and propose a differentially private holdout mechanism, Blum and Hardt (2015) design a leaderboard protocol that limits the information an adaptive participant can extract, and Whitehill (2018) show that a log-loss leaderboard can be inverted to recover hidden labels. Our procedure is of the same family: it treats the returned PCC as an oracle and reconstructs the projection of the gold

vector onto the span of submitted predictions. The Netflix Prize solutions also blended prediction vectors against leaderboard error readings (Töscher et al., 2009), but those readings came from a quiz set separate from the final test set; our setting has no such separation. Unlike Whitehill (2018), we do not attempt to recover individual labels, and we used 20 of the 100 test-phase submissions allowed. Section 5.5 measures the resulting optimism.

3 Methods

Figure 1 gives an overview of the prediction and feedback pipeline.

3.1 Task and Data

Given a text, the system predicts a Valence and an Arousal score on a 1–9 scale, where higher values mean more positive and more activated affect. The official evaluation (Lin et al., 2026) scores each dimension by mean absolute error (MAE) and Pearson correlation coefficient (PCC), giving four metrics: V-MAE, V-PCC, A-MAE, and A-PCC.

For supervision we use the sentence-level CVAS and text-level CVAT corpora of Chinese EmoBank (Lee et al., 2022), and the ROCLING 2025 validation data (Lee et al., 2025), whose shared-task setting is closer to the target task than general-domain resources. We also generate 300 fictional texts that match the topics and register of the ROCLING 2026 data (Section 3.4), and the CVAW lexicon (Yu et al., 2016a) provides a dictionary-based Arousal signal.

No labeled target-domain data was available to us: the 200-text validation set and the 1,100-text test set (Lin et al., 2026) were released without labels and could be scored only through the evaluation service, which pairs predictions with gold labels by row position rather than by ID, so the official row order must be preserved.

3.2 Fine-Tuned Chinese Encoders

Our supervised regressors are based on `hfl/chinese-roberta-wwm-ext` (Cui et al., 2021), a whole-word-masking variant of BERT (Devlin et al., 2019) pre-trained on Chinese, and `hfl/chinese-macbert-base` (Cui et al., 2020). A linear head on the pooled `[CLS]` representation jointly predicts Valence and Arousal under a mean squared error loss; the targets are standardized with the training mean and standard deviation, and predictions are mapped back to the 1–9 scale and clipped.

Both models use CVAS (2,583 sentences), CVAT (2,954 texts), and the ROCLING 2025 validation data (994 texts), 6,531 examples in total, with a 10% held-out split for checkpoint selection (training details in Appendix C). We also train a RoBERTa regressor on the ROCLING 2025 validation data alone; this 2025-only RoBERTa sees only shared-task supervision.

3.3 Frozen Embedding Regression

Following the embedding-plus-SVR approach of Li and Lin (2025), we encode each text with the frozen `intfloat/multilingual-e5-large-instruct` model (Wang et al., 2024) and train separate radial basis function (RBF) support vector regressors (Drucker et al., 1997) for Valence and Arousal on all 6,531 examples, with the hyperparameters of Li and Lin (2025) (Appendix C).

3.4 Synthetic-Data Specialist

None of our supervised data comes from the target domain, so we train a specialist on synthetic texts in the register and on the topics of the target domain, with extra examples at the extremes of the VA space.

We generated 300 fictional Chinese texts in a new-immigrant register with subagents of Claude Code, Anthropic's agentic command-line tool (Claude Opus 4.8 according to our records of the period): 150 high-Arousal and 150 low- to mid-Arousal texts, each requested for a specific target Valence and Arousal. Register and topics were anchored on the unlabeled ROCLING 2026 validation texts, and no text shares a 12-gram with any ROCLING 2026 validation or test text. Appendix C gives the generation details and Appendix D a reconstruction of the unarchived prompt.

Following the multi-teacher pseudo-labeling of Jian et al. (2025), the fine-tuned RoBERTa and MacBERT regressors and the e5-SVR model serve as teachers (Hinton et al., 2015): each synthetic text is labeled, for Valence and Arousal separately, with the equal-weight average of its generation target and the mean of the three teacher predictions. The teachers rank the synthetic texts consistently with the generation targets (V-PCC 0.88, A-PCC 0.91) but compress the Arousal scale (standard deviation 1.13 versus 2.03), and the averaged label keeps a wider spread (1.50). The specialist is a RoBERTa regressor trained with the recipe above on the original training data plus the synthetic texts. We also tested teacher-consistency filtering, which

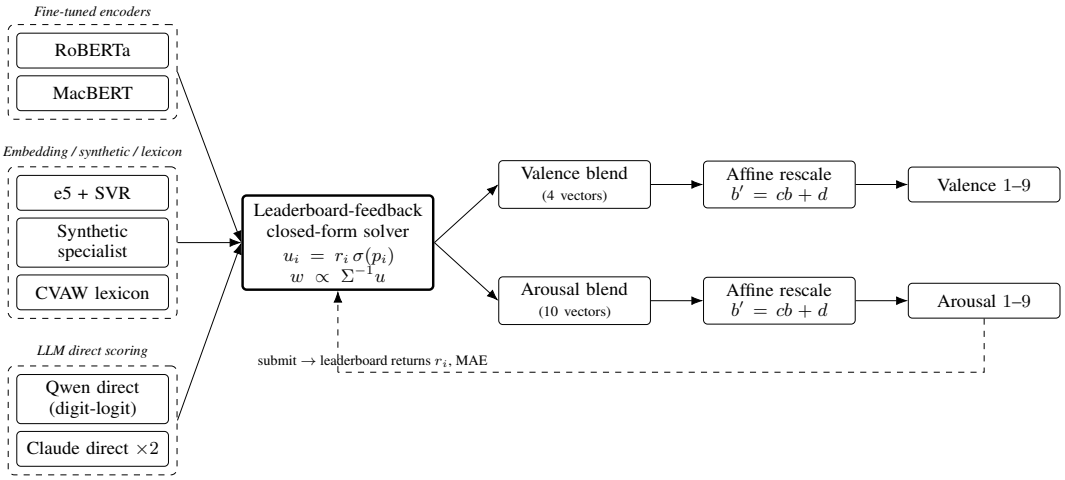

Figure 1: System overview. The closed-form solver turns leaderboard readings into per-dimension blend weights without gold labels; each blend is then rescaled to lower MAE without changing PCC (Section 3.7.2). Dashed: each submission returns a PCC and an MAE reading.

keeps 207 of the 300 texts (Appendix C); the final system uses all 300.

3.5 Lexicon Scoring

For the CVAW Arousal signal, forward maximum matching finds the lexicon entries in a text, and the prediction is the mean Arousal of the three highest-Arousal matches, or the global Arousal mean if nothing matches. The signal is weak on its own; we include it because its errors may differ from those of the neural models, although its contribution to the final blend turns out to be at the noise level (Section 5.3).

3.6 LLM Direct Scoring

We also ask LLMs to score each text directly, without fine-tuning, following their use as zero-shot numeric evaluators (Kocmi and Federmann, 2023; Liu et al., 2023; Zheng et al., 2023). We explored several prompt framings on the ROCLING 2025 proxy (positive-versus-negative polarity, reversed scales, emotional intensity, internal agitation, physiological arousal, and read-aloud tone); only the configurations described below enter the final system.

For the Qwen models (Qwen Team, 2025), instead of the greedily decoded score we use the expectation of the next-token distribution restricted to the digits 1–9, $\hat{y} = \sum_{k=1}^{9} k\,P(k \mid x) / \sum_{k=1}^{9} P(k \mid x)$, a continuous score on the task's 1–9 scale. In our runs these digits receive almost all of the next-token probability mass (mean above 0.999), so the renormalization is a safeguard rather than a correction.

Both Qwen models score Arousal zero-shot in a single forward pass with 4-bit quantization: `Qwen3-4B-Instruct-2507` with a system prompt that defines the 1–9 scale, and `Qwen3-8B`, with thinking mode disabled, under the same prompt and a reversed calmness scale whose output is flipped, the two z-normalized scores being averaged (Appendix C). We use the Qwen scorers for Arousal only, because development effort and submissions were concentrated on Arousal, where the supervised encoders were weakest.

For Claude (Anthropic, 2026), Claude Code subagents (Claude Opus 4.8 according to our records) score all 1,100 test texts zero-shot in batches of 50, with a rating rubric based on the Self-Assessment Manikin (Bradley and Lang, 1994), returning Valence and Arousal scores on the 1–9 scale. We ran two passes: the first rates each text on its own, and the second also instructs the model to compare the texts within the batch. The first pass and the z-normalized average of both passes enter the ensemble as separate vectors. Appendix C describes the sampling settings, the unarchived prompt, and a batch-effect check.

3.7 Leaderboard-Feedback Closed-Form Ensemble

Ensemble weights are normally fitted on a labeled validation set, which we did not have. The evaluation service, however, privately returns the four metrics for every submission; we call these returned scores readings, or leaderboard feedback. Except for two baseline readings that we recorded to four decimals, the readings are available at full

floating-point precision, and we use them to compute the PCC of any linear combination of previously submitted prediction vectors.

3.7.1 PCC Reconstruction

Let y denote the hidden gold vector and p_i the prediction vector of the i-th submission. The leaderboard reports $r_i = \mathrm{corr}(p_i, y) = \mathrm{cov}(p_i, y)/(\sigma(p_i)\sigma(y))$. Although $\sigma(y)$ is unknown, $\sigma(p_i)$ can be computed locally, so

$$u_i = \frac{\mathrm{cov}(p_i, y)}{\sigma(y)} = r_i \sigma(p_i). \tag{1}$$

For a linear ensemble $b = \sum_i w_i p_i$, bilinearity gives $\mathrm{cov}(b, y) = \sum_i w_i \,\mathrm{cov}(p_i, y)$, so

$$\mathrm{corr}(b, y) = \frac{\sum_i w_i u_i}{\sigma(b)}, \tag{2}$$

where $\sigma(b)$ is computed from the locally available prediction vectors: the unknown $\sigma(y)$ cancels, and the PCC of any linear mixture is obtained without access to the test labels.

With Σ the covariance matrix of the submitted prediction vectors and u the vector of the u_i, the PCC-maximizing unconstrained weight direction is

$$w \propto \Sigma^{-1} u. \tag{3}$$

Equivalently, w is the least-squares regression of the gold vector on the submitted predictions, fitted on the test set itself through the sufficient statistics Σ and u; it has as many free parameters as basis vectors (ten for Arousal), which is the source of the optimism measured in Section 5.5. We solve this system without regularization. A submission that is an affine transformation or a blend of vectors already in the basis is not added to it, since it would make Σ singular; nearly collinear vectors are kept (Appendix A).

3.7.2 MAE-Oriented Affine Scaling

Because PCC is invariant to positive affine transformations, the blend b can be replaced by $b' = cb + d$ with $c > 0$, which changes its mean and spread, and hence its MAE, without changing its PCC. The gold-test mean μ_y and standard deviation σ_y needed to choose c and d are estimated from the leaderboard MAE readings. We model the error of a submitted vector as normal,

$$p_i - y \sim \mathcal{N}\big(\bar{p}_i - \mu_y,\ \sigma(p_i)^2 + \sigma_y^2 - 2\sigma_y u_i\big)\,, \tag{4}$$

so that its expected absolute error is the mean of a folded normal distribution, and choose μ_y and σ_y by a grid search that minimizes the squared difference between these expected values and the MAE readings. The fit gives $\mu_y = 5.50$ and $\sigma_y = 1.12$ for Arousal (17 readings) and $\mu_y = 5.81$ and $\sigma_y = 1.54$ for Valence (6 readings). Under Eq. (4), the MAE-optimal shift d centers the error at zero, and the scale c is chosen by a sweep over $[0.2, 3.0]$ in steps of 0.02, which recovers the model's closed-form optimum $c = r\sigma_y/\sigma(b)$ up to the grid step.

When the correlation is weak, the MAE-optimal output spread, about $r\sigma_y$ under the normal model, is much narrower than the gold spread, so the affine step deliberately shrinks the Arousal output: the final submission has an Arousal standard deviation of 0.50, close to 0.444×1.12, while the submitted vectors of the supervised encoders have 0.64–0.83. The error model is slightly optimistic for the submission itself, predicting an A-MAE of 0.804 against the reading of 0.816.

Per-dimension solving. Both steps are carried out separately for each dimension, each with its own basis of submitted vectors; every submission carries a Valence and an Arousal column and therefore yields one reading per dimension. The final Arousal blend uses ten vectors and the final Valence blend four; Appendix A lists each vector with its standalone PCC and weight.

4 Experiments

We use the labeled ROCLING 2025 test set (1,541 texts) as a local proxy for model development, and the official evaluation on the 1,100 ROCLING 2026 test texts to select and refine the final ensemble. The proxy differs in domain from the target but shares its domain with the 994 ROCLING 2025 validation texts in our training data, which flatters the proxy scores. Table 1 reports the supervised models, their equal-weight ensembles, and the two synthetic-data variants on the proxy; the three-model ensemble is stronger than any of its components.

Before the test phase, three validation-phase submissions, on 200 texts whose labels were never released, chose a constant shift of $+0.5$ for the Arousal output of the RoBERTa regressor (from $+0.25$, $+0.5$, and $+1.0$), which changes only its Arousal MAE. All ROCLING 2026 test-phase submissions were made on 13–14 July 2026, before the test-phase deadline of 15 July 2026, 16:00 UTC;

System	V-MAE↓	V-PCC↑	A-MAE↓	A-PCC↑
RoBERTa	0.500	0.795	0.771	0.629
e5-SVR	0.524	0.761	0.795	0.566
MacBERT	0.513	0.794	0.768	0.637
RoBERTa + e5-SVR	0.469	0.808	0.738	0.638
Encoder ensemble	**0.466**	**0.812**	**0.729**	**0.650**
Synthetic, filtered (207)	0.510	0.795	0.777	0.633
Synthetic, full (300)	0.481	0.795	0.781	0.641

Table 1: Performance on the ROCLING 2025 test set used as a local proxy evaluation. Ensembles use equal weights; the encoder ensemble is RoBERTa + MacBERT + e5-SVR. Single runs with seed 42, the models used in the submitted system; three-seed means are given in Section 5.4.

System	V-MAE↓	V-PCC↑	A-MAE↓	A-PCC↑
RoBERTa	0.6666	0.8570	0.8750	0.3499
Encoder ensemble	0.6135	0.8756	0.8655	0.3518
+ Arousal signals	0.6071	0.8757	0.8290	0.4156
+ Claude (final)	**0.5555**	**0.8951**	**0.8159**	**0.4438**

Table 2: Progressive performance on the official ROCLING 2026 test set. Every row is an official reading of a submission made before the test-phase deadline; the last row is our final test-phase submission. The third row is an 8-vector blend that adds MacBERT, the synthetic specialist, the 2025-only RoBERTa, the CVAW lexicon, and the two Qwen scorers to the first two rows. In the third row, Valence is a fixed blend of the first two rows' Valence columns and an equal-weight Valence average that also includes the synthetic specialist, first submitted with an earlier intermediate blend. Table 7 lists all submissions.

the last of them (evaluation-service submission 849542, 14 July) is our final test-phase submission.

On the test set we first submit the RoBERTa regressor and the equal-weight ensemble of RoBERTa, MacBERT, and e5-SVR, hereafter the encoder ensemble, both with Arousal shifted by $+0.5$; the RoBERTa Valence is passed through a linear calibrator, a least-squares map from the model's two outputs to the gold labels of the ROCLING 2025 validation data (994 texts, which are also part of the model's training data). We then add further Arousal signals, including the lexicon and the two Qwen scorers, combined with the closed-form ensemble of Section 3.7, and finally the Claude scorer in both dimensions (Table 2).

5 Results

5.1 Overall Performance

Each stage in Table 2 improves on the previous one. The encoder ensemble improves on RoBERTa alone mainly on Valence. The closed-form blend of the heterogeneous Arousal signals raises A-PCC from 0.3518 to 0.4156 and lowers A-MAE from 0.8655 to 0.8290; its small Valence change comes from a separate fixed Valence blend (see the caption), not from the Arousal signals. Adding Claude direct scoring improves all four metrics again.

The largest relative gain is in Arousal PCC: compared with the RoBERTa regressor, the final submission increases A-PCC by 0.0939 (27%) and reduces A-MAE by 0.0591. Part of this gain is optimism from fitting to test-set feedback, which Section 5.5 estimates at about 0.016 PCC and 0.008 MAE at the 2026 signal strength, well below the gain itself. This supports our hypothesis that heterogeneous prediction sources can improve the weak Arousal ranking of the supervised models.

The A-MAE gain has two sources. Under the error model of Section 3.7.2, rescaling the RoBERTa regressor alone would lower its predicted A-MAE from 0.876 to 0.840, a gain of 0.036; since the model is optimistic by about 0.012 at the shrinkage of the final submission, and possibly more at the stronger shrinkage this would need ($c \approx 0.56$), we attribute roughly 0.02–0.035 of the 0.059 A-MAE gain to rescaling and the remainder to the higher PCC. Rescaling is cheap and model-agnostic, and any system with a weak Arousal correlation could benefit from it, provided the gold mean and spread can be estimated, here from leaderboard MAE readings.

5.2 Comparison with the Official Ranking

Table 3 reproduces the organizers' official ranking, which ranks each metric separately and orders systems by mean rank (Lin et al., 2026). The evaluation service, hosted on Codabench, scores every submission privately, and each participant chooses which scored submission to show on the public leaderboard; the official ranking lists the entries that were public when it was compiled. Our final submission (participant `yu_mkt`, submission 849542, 14 July 2026) was scored during the test phase but, like all our other submissions, had not yet been made public when the ranking was compiled, so we are not part of the official ranking. We add it to the table for comparison only and do not claim a rank. Relative to the listed systems, it would rank first in A-MAE and A-PCC and second in V-MAE and V-PCC, a mean rank of 1.5; CYUT would move to second place in both Arousal metrics, also a mean rank of 1.5.

On Valence, our system trails CYUT by 0.0189 in MAE and 0.0020 in PCC. On Arousal, our

Team	V-MAE	V-PCC	A-MAE	A-PCC	Rank
CYUT	**0.5366**	**0.8971**	0.8658	0.4376	1
Ours†	0.5555	0.8951	**0.8159**	**0.4438**	—
SCUNLP	0.5793	0.8815	0.8951	0.4050	2
aoff123	0.6018	0.8688	0.9403	0.4052	3
DonkiDonki	0.6200	0.8663	0.9259	0.3566	4
BEDO	0.6418	0.8619	0.8992	0.3204	5
wuhongyi	0.6495	0.8586	0.9294	0.3366	6
845310	0.6684	0.8542	0.9341	0.3581	7

Table 3: Official ROCLING 2026 ranking (by mean per-metric rank). †Our final test-phase submission, scored by the same evaluation service on the same test set but not part of the official ranking, because it was not yet public on the leaderboard when the ranking was compiled. Team names as listed by the organizers.

Signal	A-PCC ↑	Blend without it
RoBERTa	0.3499	0.4435
MacBERT	0.3347	0.4438
e5 + SVR	0.2832‡	—
Encoder ensemble	0.3518	0.4436
Synthetic specialist	0.3623	0.4432
CVAW lexicon	0.1522	0.4430
RoBERTa, 2025 data only	0.2076	0.4360
Qwen3-4B	0.3628	0.4416
Qwen3-8B, 2 prompts	0.3013	0.4423
Claude, first pass	0.4203	0.4429
Claude, two-pass average	**0.4217**	0.4436

Table 4: Individual Arousal prediction signals evaluated on the official ROCLING 2026 test set. The last column is the A-PCC of the closed-form blend of the other nine vectors (full blend: 0.4438), computed offline from the official readings with Eq. (2). ‡Not submitted alone; recovered exactly from the reading of the encoder ensemble via Eq. (1); it enters the blend only through the encoder ensemble.

A-MAE is 0.050 below CYUT's, far more than any of our estimates of the MAE optimism (Section 5.5); since the two A-PCCs are nearly equal, this margin reflects calibration, which the affine step targets, rather than better ranking. Our A-PCC exceeds CYUT's by only 0.0062, less than the PCC optimism we estimate at the 2026 signal strength (0.016), so CYUT's Arousal ranking is probably slightly better than ours, and the tie in mean rank rests on this margin: with an A-PCC below CYUT's, our mean rank would be 1.75 against CYUT's 1.25. Other teams may also have selected among several submissions, and with 1,100 texts the sampling standard error of a single A-PCC near 0.44 is about 0.024, so small differences between the top systems should not be over-interpreted.

5.3 Individual Arousal Signals

Table 4 reports the official A-PCC of each Arousal signal in the final system. These correlations are far below those of the same models on the 2025 proxy (Table 1): the RoBERTa regressor drops from an A-PCC of 0.629 to 0.350. The two corpora differ in annotation pools and Arousal distributions, and the proxy shares its domain with 994 of our training texts, so the gap is not a precise measure of domain transfer; it does indicate that much of the supervised models' apparent Arousal competence does not carry over to the new-immigrant domain, and it is the main reason we treat the proxy as a development tool only.

Claude direct scoring provides the strongest individual Arousal signal, followed by Qwen3-4B and the synthetic specialist. Because Eq. (2) is exact, the effect of removing any vector or group of vectors from the blend can be computed offline from the existing readings; these ablations are post-hoc analyses, not submissions. Each reduced blend is re-solved on the same readings, so the differences are in-sample: removing even a pure-noise vector lowers the blend's in-sample PCC by about $(1 - r^2)/(2rn) \approx 0.0008$ in expectation ($r = 0.444$, $n = 1{,}100$), and contributions of that size cannot be distinguished from zero. Removing all four LLM vectors lowers the blend's A-PCC from 0.4438 to 0.3690; removing only the two Claude vectors gives 0.4156, the reading of the eight-vector blend in Table 2, and removing only the two Qwen vectors gives 0.4376. The group effect is larger than the sum of the two family effects (0.075 versus 0.028 for Claude plus 0.006 for Qwen) because the Claude and Qwen scorers partly substitute for each other. On Valence, removing the Claude vector lowers the blend's V-PCC from 0.8951 to 0.8758, while removing any other Valence vector costs less than 0.002. LLM direct scoring therefore supplies most of the gain over the supervised encoders in both dimensions.

Individual PCC does not determine a vector's value in the blend. The 2025-only RoBERTa is the second-weakest signal on its own and receives a negative weight that, once weights are standardized, is second in magnitude only to that of the first Claude pass (Appendix A), yet it is the single most valuable vector to keep: removing it costs 0.0078 A-PCC, more than any LLM vector alone. We read it as a correction term that removes a 2025-specific component from the other encoders, but a weak vector with a large negative weight is also the pattern most exposed to overfitting, which our 2025 replay, where this model is in-domain, may not capture (Appendix E). The CVAW lexicon, the

weakest signal, contributes 0.0008, at the noise level, and receives a small negative weight. The final blend exceeds the strongest individual signal by 0.0221, and no vector other than the 2025-only RoBERTa accounts for more than 0.0022 of its A-PCC, so the pool is substantially redundant.

5.4 Synthetic Data Analysis

The synthetic specialist is the RoBERTa recipe plus the 300 synthetic texts. On the ROCLING 2026 test set, in the submitted single run, it raises A-PCC from 0.3499 to 0.3623 and its Arousal predictions are wider than the RoBERTa regressor's (standard deviation 0.83 versus 0.71). Its submitted A-MAE, after a constant shift of $+0.663$ that gives it the same Arousal mean as the encoder ensemble (5.40), is worse (0.8951 versus 0.8750), as Section 3.7.2 predicts: with a weak correlation, a wider output raises MAE. Its contribution to the final Arousal blend (0.0006) is below the noise level of Section 5.3, and its Valence enters the final Valence blend with a small negative effective weight (Appendix A).

To separate the effect of the synthetic data from run-to-run variation, we retrained the RoBERTa regressor without synthetic data, with the filtered 207 texts, and with the full 300 texts under two further random seeds (43 and 44, besides the seed 42 of the submitted models). The seed changes the initialization of the regression head, the order of the training batches, dropout, and the 10% held-out split used for checkpoint selection; data and hyperparameters are unchanged. The widening seen in the submitted run does not survive retraining: averaged over the three seeds, the standard deviation of the Arousal predictions on the 2026 test texts is 0.76 with the full synthetic set and 0.77 without synthetic data. On the ROCLING 2025 proxy, the full set gives a small but consistent gain: A-PCC is 0.634 ± 0.007 (mean $\pm$ standard deviation over seeds), against 0.626 ± 0.007 for the filtered set and 0.623 ± 0.009 without synthetic data, and the full set is ahead of both in every seed. Its A-MAE is also the lowest (0.766 ± 0.013, against 0.777 ± 0.001 and 0.782 ± 0.012); its slightly worse A-MAE in Table 1 is specific to seed 42. Valence is unaffected (V-PCC 0.794–0.797 for all three).

We therefore find no benefit from teacher-consistency filtering and keep all 300 synthetic texts. The gain from the synthetic data is about one seed standard deviation, a modest effect that improves the ranking of Arousal rather than widening its scale.

5.5 Use of Test-Set Feedback and Overfitting

The final submission was shaped by the 19 test-phase submissions that preceded it (Table 7). Their feedback entered the final system through three direct paths, detailed in Appendix B: PCC readings set the blend weights, MAE readings set the affine parameters, and, for Valence only, readings informed which vectors were used. The readings also influenced which signals we developed and submitted next, an adaptivity that the replay below does not capture.

Replay on labeled data. To measure how much this procedure overstates performance, we replay it after the competition on the ROCLING 2025 test set, whose 1,541 gold labels are known. In each of 1,000 repetitions, the set is split at random into a 1,100-text "board", the size of the 2026 test set, and a 441-text held-out part. Only the board PCC and MAE readings of the basis vectors feed the same procedure that produced the final submission: closed-form weights, the folded-normal MAE model (fitted to these readings only, without intermediate blends), and the c/d sweep. Optimism is the gain over a fixed single model (RoBERTa) on the board minus the same gain on the held-out part, paired per split; the fixed model itself serves as a null control.

The main replay rebuilds the basis of the final submission on 2025 data with the same recipes: the ten Arousal and four Valence vectors of the submitted blends (Table 6), with the Qwen and CVAW scorers rerun unchanged. The Claude vectors are an approximation: Claude Opus 4.8 could no longer be selected, so we re-scored the unlabeled 2025 texts with Claude Opus 5.5 (`claude-opus-5-5`), using rewritten prompts that follow Section 3.6 (Appendix D). A second replay uses eight distinct Arousal vectors from our earlier development pool, without LLM or lexicon signals. Because the 2025 signals are much stronger than the 2026 ones (full-set blend A-PCC 0.72 against 0.44), a third replay weakens them: in each repetition the 2025 Arousal labels are shrunk toward their mean and Gaussian noise is added, keeping their standard deviation, so that the blend's board PCC falls to 0.45, close to the 2026 reading of 0.444.

Table 5 shows that, on the final-submission basis at the 2025 signal strength, the Arousal blend keeps about 94% of its board gain on the held-out part:

Procedure	MAE optimism	PCC optimism
Final-submission basis, Arousal (10 vectors)		
Blend, as submitted	+0.0071	+0.0055
Blend, ridge $\lambda = 0.1$	+0.0045	+0.0028
Blend, signal weakened to 2026 level	+0.0077	+0.0157
Final-submission basis, Valence (4 vectors)		
Blend, as submitted	+0.0008	+0.0003
Development pool, Arousal (8 vectors)		
Blend, as submitted	+0.0066	+0.0049
Best single model on board	+0.0113	+0.0105
Fixed model (control)	−0.0003	−0.0009

Table 5: Post-competition replay on ROCLING 2025: optimism of each procedure, averaged over 1,000 random 1,100/441 splits (Monte Carlo standard error $\leq$0.001, excluding the sampling uncertainty of the 2025 test set itself). Rows other than the control give the board gain over the fixed RoBERTa model minus the held-out gain; the control row gives the fixed model's own board-minus-held-out difference (Arousal). The Claude vectors in the final-submission basis were rescored with Claude Opus 5.5 and only approximate the submitted ones. Ridge adds $\lambda\,\mathrm{tr}(\Sigma)/k$ to the diagonal of Σ for k vectors. In the weak-signal row, the label noise is redrawn for every split.

relative to the control, its MAE gain shrinks from 0.1090 to 0.1020 and its PCC gain from 0.0886 to 0.0831. Adding the LLM and lexicon signals changes the optimism only slightly (0.0071 versus 0.0066 MAE and 0.0055 versus 0.0049 PCC for the development pool), but weakening the signal to the 2026 level nearly triples the PCC optimism, to 0.0157, while the MAE optimism rises only to 0.0077. This is expected of a fitted linear blend, whose PCC optimism grows roughly as $k(1-r^2)/(nr)$ for k fitted weights, n texts, and correlation r, as we derive from the covariance-penalty estimate of the optimism of least squares (Hastie et al., 2009); it is twice the in-sample inflation of Section 5.3 because the held-out PCC falls by about as much as the board PCC rises. The final submission is in the weak-signal regime, so we take these values as our estimate of its optimism. The four-vector Valence blend is essentially unaffected, and its signal is no weaker in 2026 than in 2025 (V-PCC 0.895 against 0.858 on the board). In the development pool, choosing the single model with the best board score, the usual alternative, is about twice as optimistic as the blend; on the final-submission basis this comparison is uninformative, because the Claude two-pass average has the best board score in every split. At the 2025 signal strength, ridge regularization of Σ, which we did not use, would have halved the PCC optimism at no cost in held-out accuracy. Correcting the final submission by the weak-signal estimates gives roughly A-MAE 0.824 and A-PCC 0.428: still well below CYUT's A-MAE of 0.8658, but about 0.01 below its A-PCC of 0.4376.

6 Conclusion

We presented a heterogeneous ensemble for Chinese Valence–Arousal regression that combines fine-tuned encoders, frozen embedding regression, synthetic data, lexicon scoring, and LLM direct scoring. Our main contribution is a closed-form ensemble that derives blend weights from leaderboard PCC feedback, followed by a positive affine rescaling that improves MAE without changing PCC. The final system substantially improves Arousal over the supervised encoder baseline, with LLM direct scoring providing the strongest complementary signal.

Our final test-phase submission achieves V-MAE 0.5555, V-PCC 0.8951, A-MAE 0.8159, and A-PCC 0.4438. Although it was not included in the official ranking, its scores are competitive with the top-ranked system. A replay on labeled ROCLING 2025 data estimates that leaderboard adaptation inflates the Arousal gain by about 0.016 PCC and 0.008 MAE at the 2026 signal strength, substantially less than the observed improvement but large enough that the small A-PCC advantage over the top-ranked system should not be over-interpreted. Without leaderboard feedback, the same ensemble principles would require a held-out labeled validation set.

7 Acknowledgment

This work was supported by the NSTC, Taiwan (Grant NSTC 115-2221-E-011-151), and the Empower Vocational Education Research Center at NTUST under the MOE Higher Education Sprout Project. The authors thank the National Center for High-Performance Computing, Taiwan, for computational and storage resources.

References

Anthropic. 2026. Models overview. Claude Platform documentation.

Oscar Araque, Ignacio Corcuera-Platas, J. Fernando Sánchez-Rada, and Carlos A. Iglesias. 2017. Enhancing deep learning sentiment analysis with ensemble techniques in social applications. *Expert Systems with Applications*, 77:236–246.

Avrim Blum and Moritz Hardt. 2015. The Ladder: A reliable leaderboard for machine learning competitions.

In *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, pages 1006–1014.

Margaret M. Bradley and Peter J. Lang. 1994. Measuring emotion: The self-assessment manikin and the semantic differential. *Journal of Behavior Therapy and Experimental Psychiatry*, 25(1):49–59.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33*, pages 1877–1901.

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 the Association for Computational Linguistics: EMNLP 2020*, pages 657–668.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, and Ziqing Yang. 2021. Pre-training with whole word masking for Chinese BERT. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3504–3514.

Yu-Chih Deng, Cheng-Yu Tsai, Yih-Ru Wang, Sin-Horng Chen, and Lung-Hao Lee. 2022. Predicting Chinese phrase-level sentiment intensity in Valence-Arousal dimensions with linguistic dependency features. *IEEE Access*, 10:126612–126620.

Yu-Chih Deng, Yih-Ru Wang, Sin-Horng Chen, and Lung-Hao Lee. 2023. Toward transformer fusions for Chinese sentiment intensity prediction in Valence-Arousal dimensions. *IEEE Access*, 11:109974–109982.

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

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Harris Drucker, Christopher J. C. Burges, Linda Kaufman, Alex Smola, and Vladimir Vapnik. 1997. Support vector regression machines. In *Advances in Neural Information Processing Systems 9*, pages 155–161.

Cynthia Dwork, Vitaly Feldman, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Aaron Roth. 2015. The reusable holdout: Preserving validity in adaptive data analysis. *Science*, 349(6248):636–638.

Elisabetta Fersini, Enza Messina, and Federico Alberto Pozzi. 2014. Sentiment analysis: Bayesian ensemble learning. *Decision Support Systems*, 68:26–38.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, second edition. Springer.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531.

Sieh-Chuen Huang and Hsuan-Lei Shao. 2025. NTU-LAW 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.

Man-Chen Hung, Chao-Yi Chen, Pin-Jung Chen, and Lung-Hao Lee. 2021. NCU-NLP at ROCLING-2021 Shared Task: Using MacBERT transformers for dimensional sentiment analysis. In *Proceedings of the 33rd Conference on Computational Linguistics and Speech Processing (ROCLING 2021)*, pages 380–384.

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.

Tom Kocmi and Christian Federmann. 2023. Large language models are state-of-the-art evaluators of translation quality. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation (EAMT)*, pages 193–203.

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):65:1–65:18.

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.

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.

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 (ROCLING 2026)*.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-Eval: NLG evaluation using GPT-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Qwen Team. 2025. Qwen3 technical report. arXiv preprint arXiv:2505.09388.

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

Andreas Töscher, Michael Jahrer, and Robert M. Bell. 2009. The BigChaos solution to the Netflix Grand Prize. Technical report, commendo research & consulting.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024. Multilingual E5 text embeddings: A technical report. arXiv preprint arXiv:2402.05672.

Jacob Whitehill. 2018. Climbing the Kaggle leaderboard by exploiting the log-loss oracle. In *The Workshops of the Thirty-Second AAAI Conference on Artificial Intelligence*, pages 392–399.

David H. Wolpert. 1992. Stacked generalization. *Neural Networks*, 5(2):241–259.

Liang-Chih Yu, Lung-Hao Lee, Shuai Hao, Jin Wang, Yunchao He, Jun Hu, K. Robert Lai, and Xuejie Zhang. 2016a. Building Chinese affective resources in Valence-Arousal dimensions. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 540–545.

Liang-Chih Yu, Lung-Hao Lee, Jin Wang, and Kam-Fai Wong. 2017. IJCNLP-2017 Task 2: Dimensional sentiment analysis for Chinese phrases. In *Proceedings of the IJCNLP 2017, Shared Tasks*, pages 9–16.

Liang-Chih Yu, Lung-Hao Lee, and Kam-Fai Wong. 2016b. Overview of the IALP 2016 shared task on dimensional sentiment analysis for Chinese words. In *Proceedings of the 2016 International Conference on Asian Language Processing (IALP)*, pages 156–160.

Liang-Chih Yu, Jin Wang, Bo Peng, and Chu-Ren Huang. 2021. ROCLING-2021 Shared Task: Dimensional sentiment analysis for educational texts. In *Proceedings of the 33rd Conference on Computational Linguistics and Speech Processing (ROCLING 2021)*, pages 385–388.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems 36: Datasets and Benchmarks Track*, pages 46595–46623.

A Final Blend and Submission Log

Table 6 lists every vector in the final submission. Each dimension is a closed-form max-PCC blend followed by a positive scale c and shift d. The weights shown already include c, so each output is fully given by its weights and its shift: $d = +2.894$ for Arousal and $d = -0.448$ for Valence. The weights do not sum to one. The Qwen and Claude vectors were submitted as rescaled z-scores and the others on their own scales, so raw weights are not comparable across vectors; the last column multiplies each weight by the vector's standard deviation, giving its contribution in output units. A negative weight does not mean the signal points the wrong way: the vector overlaps strongly with others, and the solver uses it to cancel shared noise. In the Valence blend, the encoder ensemble and the four-model Valence average are nearly collinear ($r = 0.998$) and differ only by the synthetic specialist, so their weights of $+1.259$ and -0.872 amount to $+0.605$ on the encoder ensemble and -0.218 on the synthetic specialist's Valence. The final submission can be rebuilt byte for byte from the submitted vectors and readings with `scripts/final_mix.py`.

B Use of Test-Set Feedback

This section details the three direct paths by which the readings of Table 7 entered the final system (Section 5.5).

PCC readings set the blend weights. Each reading r_i and the locally computed $\sigma(p_i)$ give u_i (Eq. (1)), and the weights follow from Eq. (3); this is the only way test-set feedback affects the weights.

MAE readings set the affine parameters. All 17 Arousal MAE readings obtained before the final submission, and for Valence the readings of the six distinct Valence vectors, are used to fit the folded-normal error model that estimates the gold mean and standard deviation; c and d are then chosen by a sweep that minimizes the predicted MAE.

Vector	Standalone PCC	Weight	Weight $\times\sigma$
Arousal (10 vectors)			
RoBERTa regressor, Arousal shifted by $+0.5$	0.3499	$+0.082$	$+0.058$
RoBERTa + MacBERT + e5-SVR, equal weights	0.3518	$+0.083$	$+0.053$
Synthetic specialist (300 texts)	0.3623	$+0.088$	$+0.073$
MacBERT regressor	0.3347	-0.020	-0.014
RoBERTa trained on ROCLING 2025 validation data only	0.2076	-0.195	-0.137
`Qwen3-4B-Instruct-2507`, zero-shot digit expectation	0.3628	$+0.090$	$+0.094$
CVAW lexicon, top-3 matches	0.1522	-0.032	-0.034
`Qwen3-8B`, two prompts, z-averaged	0.3013	$+0.071$	$+0.071$
Claude direct scoring, first pass	0.4203	$+0.207$	$+0.217$
Claude direct scoring, two-pass z-average	0.4217	$+0.113$	$+0.118$
Valence (4 vectors)			
RoBERTa regressor with linear VA calibrator	0.8570	$+0.176$	$+0.213$
RoBERTa + MacBERT + e5-SVR, equal weights	0.8756	$+1.259$	$+1.546$
Equal-weight Valence average of RoBERTa, MacBERT, e5-SVR, and the synthetic specialist	0.8742	-0.872	-1.093
Claude direct scoring, two-pass z-average	0.8763	$+0.534$	$+0.747$

Table 6: Components of the final submission. Standalone PCC is the official reading of that vector submitted on its own; weight $\times\sigma$ is the weight times the vector's standard deviation on the test texts.

Selection of vectors. For Arousal, every signal that received a reading entered the blend regardless of its score, including the weakest (CVAW, 0.152); two further probes were abandoned before submission on the basis of an offline redundancy analysis, without leaderboard readings. For Valence, four of the six distinct vectors were used: one of the two left out is itself a blend of earlier vectors, and the other (the first Claude pass) is nearly collinear with the two-pass average and would raise the predicted V-PCC by only 0.00004. This choice did look at leaderboard readings and is therefore a mild form of selection.

Intermediate blends. The four intermediate blends in Table 7 checked the reconstruction: each predicted PCC matched the returned reading to within 10^{-5}, and each MAE reading was at most about 0.012 above the error model's prediction, the bias also seen for the final submission. Their readings were used only for the MAE model.

C Implementation Details

Fine-tuned encoders. Of the 6,531 training examples, a random 10% (654) is held out, and after each epoch the checkpoint with the highest Valence PCC on this split, the default of our training script, is kept; the other 5,877 examples are used for training. Training uses four epochs, a batch size of 16, a maximum sequence length of 256, a learning rate of 2×10^{-5} with 10% warm-up and weight decay 0.01, and random seed 42, which also fixes the held-out split. The 2025-only RoBERTa selects its checkpoint by Arousal PCC, and the synthetic specialist draws its 10% held-out split from the combined pool.

Frozen embedding regression. Texts are encoded without an instruction prefix and truncated to 512 tokens, with masked mean pooling and L_2 normalization. The RBF support vector regressors use the raw 1–9 targets, $C = 10$, and $\epsilon = 0.2$ as in Li and Lin (2025), implemented with scikit-learn (Pedregosa et al., 2011).

Synthetic-data generation. The 300 texts were generated by two parallel subagents of Claude Code orchestrated from a Claude Fable 5 session; according to our records of the period the subagents ran Claude Opus 4.8, but the session logs that would confirm this were not retained. One subagent produced 150 high-Arousal texts and the other 150 low- to mid-Arousal texts, each against a stratified quota of target VA cells (high-Arousal positive and negative; calm positive, negative, and neutral). Topics (marriage, workplace experiences, homesickness, children's education, and identity documents) were anchored on the unlabeled ROCLING 2026 validation texts, and none of the 300 texts duplicates, or shares a 12-gram with, any ROCLING 2026 validation or test text. The original generation prompts were not archived; Appendix D gives a reconstruction checked in a small pilot. Because the texts are fictional rather than derived from real personal narratives, they raise no privacy concerns for actual new-immigrant informants.

Teacher-consistency filtering. The filter removes a synthetic text when its generation target and the teacher mean differ by more than 1.5 on either dimension; 207 of the 300 texts survive. The removed texts are concentrated at the Arousal extremes (72 of the 93 have a generation target Arousal of at most 3 or above 7, against 100 of the

No.	Submitted vector	V-MAE / V-PCC / A-MAE / A-PCC	Effect on the final system
1	RoBERTa regressor (baseline)	0.6666 / 0.8570 / 0.8750 / 0.3499	Both dimensions enter the blends
2	Encoder ensemble: RoBERTa + MacBERT + e5-SVR, equal weights (baseline)	0.6135 / 0.8756 / 0.8655 / 0.3518	Both dimensions enter the blends
3	Synthetic specialist (baseline); Valence column copied from no. 1	0.6666 / 0.8570 / 0.8951 / 0.3623	Arousal enters the blend
4	Probe: equal-weight Valence average of RoBERTa, MacBERT, e5-SVR, and the synthetic specialist; Arousal as no. 2	0.6079 / 0.8742 / 0.8655 / 0.3518	Valence enters the blend
5–6	Check: two known blends of nos. 2 and 3	A-PCC 0.3656, 0.3616	Verified the PCC reconstruction; MAE model only
7–8	Void: rows reordered by ID by mistake	PCC ≈ 0	Not used
9	Intermediate: first rescaled Arousal blend; first fixed Valence blend of nos. 1, 2, and 4	0.6071 / 0.8757 / 0.8439 / 0.3656	First test of scale c; MAE model only
10	Probe: MacBERT regressor	A-PCC 0.3347	Arousal enters the blend
11	Intermediate: 4-vector Arousal blend	0.6071 / 0.8757 / 0.8397 / 0.3671	MAE model only
12–14	Probes: RoBERTa trained on ROCLING 2025 data only; Qwen3-4B; CVAW lexicon	A-PCC 0.2076, 0.3628, 0.1522	Arousal enters the blend
15	Intermediate: 7-vector Arousal blend	0.6071 / 0.8757 / 0.8327 / 0.4066	MAE model only
16	Probe: Qwen3-8B	A-PCC 0.3013	Arousal enters the blend
17	Intermediate: 8-vector Arousal blend, without Claude	0.6071 / 0.8757 / 0.8290 / 0.4156	MAE model only
18	Probe: Claude, first pass	0.5887 / 0.8747 / 0.9450 / 0.4203	Arousal enters the blend
19	Probe: Claude, two-pass average	0.5832 / 0.8763 / 0.9469 / 0.4217	Both dimensions enter the blends
20	**Final submission**	**0.5555 / 0.8951 / 0.8159 / 0.4438**	Final system, submitted 14 July 2026

Table 7: All 20 ROCLING 2026 test-phase submissions, in chronological order. A probe submits one new vector to obtain its reading; check and intermediate submissions are blends of vectors already read. Two further probes were prepared but abandoned without submission.

207 retained), which follows almost by construction from the teachers' compressed Arousal scale, so the disagreement need not indicate a bad label. Over three training seeds the full set is at least as good as the filtered one on the ROCLING 2025 proxy (Section 5.4).

Qwen scorers. Both Qwen models are loaded with 4-bit NF4 quantization (Dettmers et al., 2023) and score each text in a single forward pass, so no sampling is involved. For `Qwen3-4B-Instruct-2507`, on 300 random ROCLING 2025 test texts, zero-shot scoring correlated better with the gold labels than a nine-shot variant (0.515 versus 0.462). The Valence gain from Claude (Section 5.3) suggests that using the Qwen scorers for Arousal only left room for improvement on Valence.

Claude scorer. As for the synthetic data, our records of the period give the model as Claude Opus 4.8, but the session logs were not retained. Claude Code exposes no sampling parameters, so we could fix neither the temperature nor a random seed, and the scores are not deterministic. Entering the first pass and the two-pass average as separate vectors amounts, since the average contains the first pass, to giving each pass its own weight; the passes agree closely (Arousal $r = 0.953$), so averaging adds little. The original prompt was not archived; Appendix D describes it and gives the rewritten prompt used for the post-competition replay. No ROCLING 2026 labels were ever available to us, so no gold information could reach the scorer, but without the prompt we cannot document which other files in the working directory the agents could read, or whether they were barred from scoring with code or keyword rules; the replay prompt forbids both. Batch scoring could introduce batch-level calibration differences, but the residuals of the Claude scores against the averaged supervised encoders do not differ between consecutive 50-text blocks of the official row order ($F \leq 1.24$ for both

passes and dimensions, against a 5% critical value of 1.58), although our records do not confirm that the batches followed that order.

D Prompts

The Qwen prompts and the replay rubric were written in Traditional Chinese, and we give English translations; the synthetic-data reconstruction was written in English. The originals are in our repository (https://github.com/Makoto1018/Rocling2026-data).

Qwen Arousal, scale definition (Qwen3-4B; first prompt of Qwen3-8B). System: *You are an expert in rating the emotional arousal of Chinese texts. Arousal measures how agitated or excited the expressed emotion is, regardless of whether it is positive or negative. 1 = very calm: flat tone, objective statement, no emotional ripple. 5 = medium: clear emotion but not intense. 9 = very agitated: extremely strong emotion, such as ecstasy, rage, extreme fear or grief, or shouting with excitement. Note: extreme sadness and extreme excitement both count as high arousal; relaxed, serene, and detached states count as low arousal. Output a single integer from 1 to 9 and nothing else.* User: *Rate the following text (output a single integer from 1 to 9). Text: {text} Arousal:*

Qwen Arousal, reversed scale (second prompt of Qwen3-8B; score flipped as $10 - \hat{y}$). System: *You are an expert in rating the calmness of Chinese texts. Rate how calm the emotion in the text is. 1 = very agitated: extremely strong emotion, such as ecstasy, rage, extreme fear or grief. 5 = medium. 9 = very calm: flat, serene tone, objective statement, no emotional ripple. Note: extreme sadness and extreme excitement both count as agitated (low score); relaxed and detached states count as calm (high score). Output a single integer from 1 to 9 and nothing else.* User: *Text: {text} Calmness:*

Claude direct scoring (both passes). The exact prompt was not archived, and no reconstruction has been validated against it. Each pass gave the model a rubric defining Valence and Arousal on the 1–9 scale after the Self-Assessment Manikin, followed by a batch of 50 numbered texts, and asked for one Valence and one Arousal score per text. The second pass added an instruction to compare the texts within the batch before scoring them.

Claude direct scoring for the 2025 replay (Section 5.5). Rewritten after the competition from the description above and run with Claude Opus 5.5 subagents, one per batch of 50 unlabeled ROCLING 2025 test texts; it approximates, and is not a copy of, the lost original prompt. Rubric (both passes): *You are an annotator of emotion in Chinese texts. Following the Self-Assessment Manikin, rate each text on two dimensions from 1 to 9 (one decimal allowed). Valence: 1 = very negative (pain, despair, anger, sadness), 3 = somewhat negative, 5 = neutral or mixed, 7 = somewhat positive, 9 = very positive (happiness, gratitude, contentment, joyful excitement). Arousal, regardless of polarity: 1 = very calm (flat tone, objective statement, relaxed, detached), 3 = somewhat calm, 5 = medium (clear but not intense emotion), 7 = somewhat agitated, 9 = very agitated (ecstasy, rage, extreme fear or grief, panic). Extreme sadness and extreme excitement both count as high Arousal; relaxed and serene states count as low Arousal. Rate the emotion expressed by the author, not your view of the events.* Pass A: *Rate each text independently, looking only at that text.* Pass B: *First read the whole batch and compare the texts' polarity and agitation so that the relative order of scores is consistent within the batch (more agitated texts get higher Arousal, more negative texts lower Valence), then rate each text.* Both passes asked for a CSV with one row per text and forbade reading any other file or scoring with code or keyword rules.

Synthetic-data generation (reconstruction). The original prompt was not archived. The following reconstruction is written from our design notes. In a 30-text pilot by a subagent that had not seen the original set, the target VA values it produced were distributed similarly to those of the original set (we did not test this formally), although its texts were shorter (48 versus 62 characters on average). *Generate fictional Traditional Chinese first-person texts by new immigrants in Taiwan (e.g., spouses from Vietnam, Indonesia, Thailand, the Philippines, or mainland China) about marriage and in-laws, work, homesickness, children's education, identity documents and residence permits, language learning, or daily life. Imitate the register, length (about 45–85 characters), and voice of the ROCLING 2026 validation texts without copying or paraphrasing any of them. All people and events must be fictional; avoid stereotypes. Valence ranges*

from 1 (very negative) to 9 (very positive) and Arousal from 1 (very calm) to 9 (very agitated). Batch A (target Arousal 5.5–8.0) splits between high-Arousal negative (anger, panic, fear; Valence 2–3.5) and high-Arousal positive (excitement, joy, pride; Valence 7–8.5), plus a few mixed texts. Batch B (target Arousal 2–5.5) splits between calm negative (loneliness, resignation; Valence 3–4), calm neutral (routine; Valence 4.5–5.5), and calm positive (contentment, gratitude; Valence 6–7.5). Give each text its intended Valence and Arousal to one decimal, as a CSV with columns ID, Text, gen_valence, gen_arousal.

E Limitations

The method requires leaderboard feedback and is adaptive by construction. Our central contribution only applies when an evaluation service returns a correlation for submitted predictions. It is therefore not a general-purpose ensembling method: in a deployment setting, or under a single-submission protocol, the weights would have to be fitted on a held-out labeled set instead. More importantly, fitting weights against repeated leaderboard readings is exactly the adaptive-data-analysis regime that Dwork et al. (2015) and Blum and Hardt (2015) warn about. Each reading leaks information about the hidden labels, so the reported test scores are optimistic. Our replay on ROCLING 2025 data (Section 5.5) puts this optimism at about 0.016 A-PCC and 0.008 A-MAE at the 2026 signal strength, and below 0.001 for Valence, but the estimate assumes that 2025 labels weakened with Gaussian noise mimic the 2026 setting and that our models behave on 2025 data as they do on 2026 data. Its Claude vectors were re-scored with a newer model (Claude Opus 5.5) and rewritten prompts, so they approximate rather than reproduce the submitted Claude signals. Claude Opus 5.5 may also have seen the ROCLING 2025 texts in training, which would make the replayed Claude signal stronger than the original, and the Qwen prompts were selected on the same 2025 test set. The replay also cannot reproduce the role of the RoBERTa model trained only on ROCLING 2025 data: on the 2025 test set it is an in-domain model (A-PCC 0.624), whereas on the 2026 test set it is a weak vector (0.208) that receives a large negative weight, which is the pattern most exposed to overfitting.

No target-domain labeled data and no significance testing. No annotated new-immigrant VA data was available to us, so every component is trained on out-of-domain supervision (Chinese EmoBank and ROCLING 2025) or on synthetic text. Apart from the three-seed synthetic-data comparison (Section 5.4), we report single-run numbers with a fixed seed (42) and no confidence intervals. The seed standard deviation of A-PCC on the proxy is 0.007–0.009, so differences below about 0.01 between systems are within run-to-run variation and should not be over-interpreted; our Valence MAE gap to the top-ranked system (0.019) is larger than that.

Affine rescaling relies on a distributional assumption. The MAE-oriented rescaling of Section 3.7.2 estimates the hidden gold mean and standard deviation with a folded-normal error model. If the true Arousal error distribution is skewed or heavy-tailed, the recovered parameters are biased and the MAE gain will not fully materialize. PCC is unaffected either way, since the transformation is positive affine.

LLM components are neither reproducible nor free. The Qwen scorers run on open weights and are reproducible, but the Claude scorer was accessed through a service that updates its models over time. The Claude scoring and the synthetic-data generation ran through Claude Code subagents whose session logs and prompts were not retained (Appendix C), so the Claude signals cannot be regenerated exactly. This is a reproducibility gap on our side rather than a property of the method, and future work using this pipeline should pin and report model versions for every LLM component. Being LLM-generated, the synthetic training data also risks importing the generator's own affective stereotypes about new-immigrant experience into a system that is then evaluated on texts about real people. We did not audit the 300 generated texts for such bias, and we would not recommend deploying this system for any consequential judgement about individuals without that audit.