cclin at SemEval-2026 Task 2: SLM-Enhanced Lightweight Multi-BERT Ensemble for Longitudinal Affect Assessment

Jing-Jun Lin
Independent Researcher
Hsinchu, Taiwan
xup6vu84m3ul6@gmail.com

Abstract

This paper describes the system developed by team **cclin** for SemEval-2026 Task 2, Subtask 1: Longitudinal Affect Assessment (Soni et al., 2026). Our goal is to predict Valence and Arousal from ecological essays and feeling words over time. We propose an efficient hybrid framework that uses quantized 7B-scale language models as deterministic meta-feature extractors and combines them with an ensemble of DeBERTa, RoBERTa, and DistilBERT encoders. The system is designed to run on a single consumer-grade RTX 5060 Ti (16GB) GPU while remaining competitive. To bridge discrete supervision and continuous evaluation, we train the model as an ordinal classification problem and decode class probabilities into continuous scores through expected-value decoding. Our best system achieved an overall V&A average of 0.587, with per-dimension composite correlations of 0.647 for Valence and 0.527 for Arousal, ranking **3rd out of 31 teams**. The results show that lightweight SLM-derived priors and multi-encoder fusion provide a strong performance–efficiency trade-off, especially for Arousal, where contextual anchoring is crucial.

1 Introduction

Emotion intensity prediction from diaries, essays, and short affective expressions is a challenging problem in natural language processing because similar emotional states may be expressed through very different lexical and narrative patterns (Wiebe et al., 2005; Mohammad and Kiritchenko, 2018). This challenge becomes harder in longitudinal settings, where systems must estimate not only the polarity of an utterance but also how that signal varies across users and time.

SemEval-2026 Task 2 addresses this setting by modeling Valence and Arousal from temporally ordered ecological text sequences. In **Subtask 1**, the goal is to predict a Valence–Arousal score pair for each item in a chronological sequence. In **Subtask 2**, systems must forecast future affective change. In this work, we focus exclusively on **Subtask 1**.[1]

Our main objective was not to build the largest possible model, but to design a strong and reproducible system under strict hardware constraints. Instead of fine-tuning large generative models, we use two quantized 7B instruct-tuned language models—Qwen2.5-7B-Instruct and Mistral-Nemo-Instruct-2407—only for zero-/few-shot meta-feature extraction. These features are then fused with representations from three BERT-family encoders: DeBERTa-v3-base, RoBERTa-base, and DistilBERT-base.

The resulting system offers three practical advantages. First, it preserves the efficiency and training stability of encoder-based models. Second, it enriches the final predictor with low-cost semantic priors from SLMs. Third, it remains lightweight enough to support rapid development on a single consumer GPU. Empirically, our best submission ranked **3rd out of 31 teams** on the official leaderboard.

Our contributions are as follows:

- We propose a lightweight hybrid framework that combines deterministic SLM-based meta-features with a multi-BERT ensemble.
- We introduce a stable expected-value decoding scheme that maps ordinal class probabilities back to continuous Valence and Arousal scores.
- We show through ablation and subgroup analyses that SLM-derived features and user-profile statistics are particularly useful for longitudinal affect assessment, especially for Arousal.

[1] Valence is defined on a 5-point scale $\{-2, -1, 0, 1, 2\}$ and Arousal on a 3-point scale $\{0, 1, 2\}$ in the released labels.

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 200–205

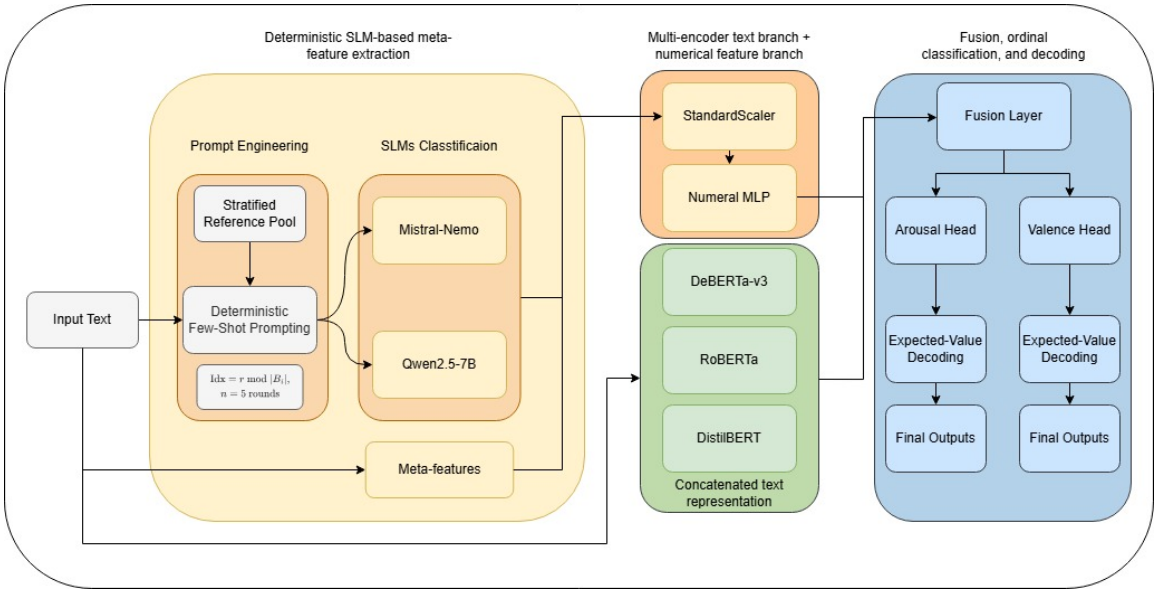

Figure 1: Overview of the proposed hybrid pipeline. The architecture combines LLM-based feature extraction with a fusion MLP for affective state classification.

2 Related Work

Transformer-based models consistently outperform earlier neural approaches in emotion and intensity prediction tasks (Lee et al., 2022). Prior work has shown that RoBERTa-style encoders are effective for modeling emotion and empathy intensity (Lin et al., 2023), while fusion architectures can improve performance by combining information across linguistic granularities (Deng et al., 2023).

BERT-family encoders remain attractive for affective computing because they provide strong representations at relatively low computational cost. RoBERTa improves robustness through larger-scale pre-training and dynamic masking (Liu et al., 2019); DeBERTa introduces disentangled attention to better model content and position (He et al., 2021); and DistilBERT offers a smaller, faster alternative that retains much of BERT's performance (Sanh et al., 2020; Devlin et al., 2019).

Recent large language models have demonstrated strong zero-shot reasoning, but their training and inference costs are often impractical for lightweight shared-task settings. Our approach differs from full LLM fine-tuning in that we use quantized 7B instruct-tuned models only as deterministic meta-feature generators. Compared with pure PLM ensembles, this hybrid design injects low-cost semantic priors into the final predictor while keeping computation manageable.

3 System Description

Figure 1 summarizes the overall architecture. The system consists of two stages: (1) deterministic SLM-based meta-feature extraction and (2) multi-transformer fusion for final prediction.

3.1 SLM-based Meta-Feature Extraction

The first stage uses Qwen2.5-7B-Instruct and Mistral-Nemo-Instruct-2407, both loaded in 4-bit precision, as lightweight meta-feature generators rather than trainable predictors. Our goal is not to rely on their raw outputs directly, but to extract stable affective priors that complement encoder-based representations.

To reduce the variance of generative outputs, we use **deterministic few-shot prompting**. We first build a stratified reference pool from the deduplicated training set. For Valence, the pool is partitioned into five buckets corresponding to $\{-2, -1, 0, 1, 2\}$. For Arousal, the pool is partitioned into three buckets corresponding to $\{0, 1, 2\}$. During inference iteration $r \in \{1, \ldots, n\}$, one example is selected from each bucket using a deterministic index:

$$\text{Idx} = r \bmod |B_i|.$$

This strategy ensures that each prompt contains balanced label anchors while avoiding the randomness of conventional few-shot sampling.

We perform $n = 5$ deterministic prompting rounds and average the outputs. Sampling is dis-

Dimension	r_{comp}	$r_{between}$	r_{within}	MAE_{comp}	$MAE_{between}$	MAE_{within}
Valence (V)	**0.647**	0.695	0.593	0.653	0.453	0.790
Arousal (A)	**0.527**	0.611	0.430	0.365	0.226	0.489

Table 1: Official test-set performance on SemEval-2026 Task 2, Subtask 1.

abled (`do_sample=False`) so that the SLM behaves as a reproducible scoring function conditioned on a fixed set of references. We also use dimension-specific prompting instructions: Valence prompts emphasize polarity on the range $[-2, 2]$, while Arousal prompts explicitly describe physiological energy levels from calm/tired to anxious/energetic.

From the two SLMs, we construct a compact numerical meta-feature vector including:

- the predicted Valence and Arousal scores from each SLM,
- signed differences between the two models,
- absolute differences as a simple uncertainty proxy,
- user-level historical mean Valence and Arousal statistics.

For unseen users, the profile statistics are imputed using the global training mean. These features help the downstream model calibrate both inter-user baselines and model agreement.

3.2 Multi-Transformer Fusion and Continuous Score Decoding

In the second stage, the task is formulated as an **ordinal classification problem** rather than direct regression. Valence is modeled as a 5-way classification task over $\{-2, -1, 0, 1, 2\}$, while Arousal is modeled as a 3-way classification task over $\{0, 1, 2\}$. We adopted this design because the released labels are discrete integers with fixed valid ranges.

This formulation has practical advantages. First, it aligns the training objective with the observed label space. Second, it avoids out-of-range predictions at inference time, which are common in unconstrained regression and would otherwise require clipping. Third, it provides a stable development signal: in our experiments, validation trends under the classification setup closely tracked final submission performance, which made iterative model selection more reliable.

To recover continuous values for official evaluation, we decode the class probability distribution using the expected value over ordinal labels:

$$\hat{s} = \sum_{i=1}^{K} p_i \cdot c_i,$$

where p_i is the softmax probability of class i, and c_i is the corresponding class value. For Valence, $c_i \in \{-2, -1, 0, 1, 2\}$; for Arousal, $c_i \in \{0, 1, 2\}$. This expected-value decoding preserves ordinal information, avoids the discontinuity of hard `argmax` decoding, and guarantees that the prediction remains inside the valid score range.

We did **not** apply additional post-hoc probability calibration methods, such as temperature scaling or isotonic regression, to adjust predictions near class boundaries. Instead, probability stability was promoted through label smoothing, deterministic few-shot meta-feature extraction, and ensemble fusion.

For text encoding, the input is processed by three encoder models in parallel: DeBERTa-v3-base, RoBERTa-base, and DistilBERT-base. For each encoder, we compute mean-pooled sentence representations using the attention mask. The resulting vectors are concatenated into a unified textual representation.

In parallel, the numerical meta-features are normalized with a standard scaler and passed through a small multilayer perceptron with Batch Normalization and GELU activation (Hendrycks and Gimpel, 2023). The textual and meta-feature representations are then concatenated and fed into a fusion layer with Dropout (0.3), followed by independent classification heads for Valence and Arousal.

We also considered whether more complex fusion mechanisms, such as attention-based fusion, might further improve performance. However, the goal of this work was to maintain a strong balance between efficiency, reproducibility, and predictive accuracy on a single consumer GPU. Under that constraint, simple concatenation offered a favorable trade-off between implementation cost and empirical performance.

Setting	Approach	System	Valence		Arousal	
			r_{comp}	$\mathbf{MAE}_{comp}$	r_{comp}	$\mathbf{MAE}_{comp}$
Analysis	User split	Seen users	0.664	0.652	0.457	0.491
	User split	Unseen users	0.629	0.653	0.614	0.486
	Text type	Words only	0.662	0.621	0.620	0.453
	Text type	Essay only	0.632	0.650	0.422	0.518
Ablation	Hybrid model	Full system	**0.647**	0.653	**0.527**	0.365
	Hybrid model	w/o SLM meta-features	0.591	0.572	0.448	0.496
	Hybrid model	w/o user profile	0.612	0.644	0.485	0.426
	Encoder choice	Single PLM (DeBERTa-v3)	0.624	0.637	0.491	0.418
Prompting	SLM prompting	Zero-shot	0.605	0.685	0.382	0.427
	SLM prompting	Deterministic few-shot	**0.647**	0.653	**0.527**	0.365

Table 2: Unified comparison table for subgroup analysis, ablation study, and prompting strategy. The updated table includes r_{comp} and MAE_{comp} for both Valence and Arousal dimensions.

3.3 Training Strategy

To reduce catastrophic forgetting while allowing the fusion layers to adapt, we use a layer-wise learning-rate strategy. The PLM encoders are optimized with a learning rate of 1×10^{-5}, while the MLP and fusion layers use 1×10^{-4}. The model is trained with AdamW (Loshchilov and Hutter, 2019), cosine scheduling with warmup, and cross-entropy loss with label smoothing of 0.1 (Szegedy et al., 2015).

This training setup is intentionally simple. Our design philosophy is that, in low-resource shared-task settings, methodological stability and reproducibility may be more valuable than adding many high-variance architectural components.

4 Experimental Results

4.1 Main Results

Table 1 reports the official results of our best system on Subtask 1. The system achieved per-dimension composite correlations of 0.647 on Valence and 0.527 on Arousal, corresponding to an overall V&A average of 0.587. These results placed our system **3rd out of 31 teams** on the official leaderboard.

The relatively strong $r_{between}$ scores suggest that the system effectively captures user-level affective baselines, which is consistent with the contribution of the profile features. At the same time, the gap between Valence and Arousal confirms that physiological activation remains harder to estimate than polarity, even with SLM-based contextual priors.

4.2 Consolidated Comparison of System Behavior

To improve readability, we merge the subgroup analysis, ablation study, and prompting comparison into a single structured table, following the presentation style commonly used in shared-task system papers. Instead of separating the results into three small tables, Table 2 organizes them by *setting type*, *approach*, and *system variant*, while keeping Valence and Arousal under grouped headers.

Three patterns are immediately visible. First, the **Words Only** setting is substantially easier than **Essay Only**, especially for Arousal, which confirms that isolated affective words provide clearer physiological cues than longer narratives. Second, among all ablations, removing **SLM meta-features** causes the largest overall drop, indicating that the proposed hybrid design is most effective when encoder representations are complemented by deterministic SLM-derived priors. Third, deterministic few-shot prompting yields a large improvement over zero-shot prompting, with the gain being much larger on Arousal than on Valence. This again supports our view that physiological activation requires stronger contextual anchoring.

4.3 Discussion

Two additional observations are worth highlighting.

Sensitivity to SLM choice. In our experiments, the system was not highly sensitive to the exact pair of 7B-scale SLMs once deterministic few-shot prompting was applied. Qwen2.5-7B-Instruct and Mistral-Nemo-Instruct-2407 produced very similar score estimates, and the main advantage of us-

Team	Valence	Arousal	V&A Avg.
UKP_Psycontrol	0.667	0.554	**0.611**
YNU	0.677	0.528	0.603
cclin	**0.647**	**0.527**	**0.587**
AFourP	0.679	0.466	0.573
lamanhnguyen	0.687	0.458	0.573

Table 3: Top-5 teams on the official leaderboard for Subtask 1.

ing both models came from improved robustness through agreement/disagreement features rather than strong complementarity. In other words, the framework appears to benefit more from stable prompting and lightweight ensembling than from any specific proprietary property of one SLM.

Temporal limitation. Although the task is longitudinal, our approach is only weakly temporal. It uses user-level historical averages as a compact longitudinal signal, but it does not explicitly model trajectories, transitions, or sequence dynamics. We deliberately accepted this limitation in order to keep the system efficient and reproducible. Future work could extend the current architecture with temporal encoders or sequence-aware fusion, particularly for Subtask 2.

5 Conclusion

This paper presented a lightweight hybrid system for SemEval-2026 Task 2, Subtask 1: Longitudinal Affect Assessment. The core idea is simple: use quantized 7B-scale instruct-tuned language models as stable affective prior generators, then combine these priors with a compact multi-BERT ensemble. This design avoids expensive LLM fine-tuning while retaining enough semantic richness to remain competitive in a challenging shared-task setting.

Our results lead to three main conclusions. First, **lightweight SLMs are useful even without fine-tuning** when they are used as deterministic meta-feature extractors rather than direct predictors. Second, **classification with expected-value decoding is a practical bridge between discrete supervision and continuous evaluation**. In our setting, this formulation gave bounded predictions and stable validation behavior without additional calibration. Third, **Arousal requires stronger contextual anchoring than Valence**. The large gain from deterministic few-shot prompting and the large gap between word-level and essay-level performance both support this observation.

More broadly, this work reflects a practical perspective on shared-task modeling: stronger performance does not always require larger models. Under realistic hardware constraints, carefully combining modest components—stable prompting, lightweight ensembling, bounded decoding, and simple user calibration—can produce a robust system with competitive results.

At the same time, the current work leaves several promising directions open. A fuller temporal treatment of emotion trajectories, attention-based or sequence-aware fusion, and systematic comparisons between regression, ordinal regression, and classification-based decoding would all be valuable next steps. Finally, we suspect that there is an important mismatch between the continuity of human affect and the discreteness of affective language. The present classification-plus-expectation design is one practical way to handle that mismatch, but we believe this question deserves deeper study in future work.

References

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.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *Preprint*, arXiv:1810.04805.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: Decoding-enhanced bert with disentangled attention. *Preprint*, arXiv:2006.03654.

Dan Hendrycks and Kevin Gimpel. 2023. Gaussian error linear units (gelus). *Preprint*, arXiv:1606.08415.

Lung-Hao Lee, Jian-Hong Li, and Liang-Chih Yu. 2022. Chinese emobank: Building valence-arousal resources for dimensional sentiment analysis. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 21(4).

Tzu-Mi Lin, Jung-Ying Chang, and Lung-Hao Lee. 2023. NCUEE-NLP at WASSA 2023 shared task 1: Empathy and emotion prediction using sentiment-enhanced RoBERTa transformers. In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 548–552.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019.

Roberta: A robustly optimized bert pretraining approach. *Preprint*, arXiv:1907.11692.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. *Preprint*, arXiv:1711.05101.

Saif Mohammad and Svetlana Kiritchenko. 2018. Understanding emotions: A dataset of tweets to study interactions between affect categories. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2020. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *Preprint*, arXiv:1910.01108.

Nikita Soni, H. Andrew Schwartz, Ryan L. Boyd, Phi Long Bui, Syeda Mahwish, August Håkan Nilsson, Adithya V Ganesan, Lyle Ungar, Niranjan Balasubramanian, and Saif M. Mohammad. 2026. SemEval-2026 task 2: Predicting variation in emotional valence and arousal over time from ecological essays. In *Proceedings of the 20th International Workshop on Semantic Evaluation (SemEval-2026)*. Association for Computational Linguistics.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. 2015. Rethinking the inception architecture for computer vision. *Preprint*, arXiv:1512.00567.

Janyce Wiebe, Theresa Wilson, and Claire Cardie. 2005. Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation*, 39:165–210.