"AGI" Team at SemEval-2026 Task 2: Predicting Variation in Emotional Valence and Arousal over Time from Ecological Essays

Harsh Rathva
Sardar Vallabhbhai National Institute of Technology, Surat, India
u24ai036@aid.svnit.ac.in

Abstract

This paper describes our submission to SemEval-2026 Task 2: Predicting Variation in Emotional Valence and Arousal. We combine RoBERTa-Large text encoding with a unidirectional GRU for temporal modeling and gated user embeddings for personalization. A four-phase staged training curriculum employs ordinal regression for absolute affect prediction and a zero-inflated delta model for change detection. Our approach achieves competitive performance on Subtask 1 (longitudinal affect assessment) with composite correlation $r = 0.600$ for valence and $r = 0.452$ for arousal. However, we observe systematic degradation in Subtask 2A (state change detection) with negative correlations ($r = -0.167$ for valence, $r = -0.147$ for arousal), revealing a fundamental trade-off between stability-oriented representations and change sensitivity. We provide detailed empirical analysis of these failure modes, contributing insights into the challenges of modeling emotional dynamics in ecological data. Code and trained checkpoints are publicly available.[1,2]

1 Introduction

Modeling emotional states from naturalistic text requires capturing both individual differences in affective expression and temporal dynamics of emotional change. SemEval-2026 Task 2 (Soni et al., 2026) presents a challenging benchmark for longitudinal affect modeling using ecological essays and feeling words collected over multiple years.

The task comprises three interconnected subtasks: (1) predicting absolute valence and arousal for each text in a user's timeline, (2A) detecting short-term emotional state changes between consecutive texts, and (2B) predicting long-term dispositional changes in baseline emotional state.

We present a hierarchical temporal architecture combining transformer-based text encoding with sequential modeling and user personalization. Our approach preserves causal temporal structure through unidirectional processing, preventing future information leakage during prediction.

Our key contributions are:

- We provide one of the first empirical analyses showing how stability-oriented temporal modeling induces systematic delta underprediction in longitudinal affect tasks.
- We achieve competitive longitudinal prediction performance ($r = 0.600$ valence, $r = 0.452$ arousal on Subtask 1) while documenting failure modes in change prediction.
- We publicly release code and model weights to support reproducibility.

Rather than focusing solely on leaderboard optimization, our work emphasizes understanding when temporal models fail and why architectures effective for absolute prediction can systematically degrade change prediction performance.

2 Background

2.1 Task Description

SemEval-2026 Task 2 focuses on predicting emotional valence and arousal from longitudinal ecological texts. The data consists of ecological momentary assessments written by U.S. service-industry workers describing their emotional state through essays and curated feeling-word lists.

The complete dataset contains 5,285 longitudinal texts written by 182 users, collected between 2021 and 2024. The official training split includes

[1] https://github.com/ezylopx5/AffectDynamics-SemEval2026Task2
[2] https://huggingface.co/Haxxsh/AffectDynamics-SemEval2026Task2

2,764 texts from 137 users. Each instance includes timestamped emotional annotations with valence on an ordinal scale from -2 (highly negative) to $+2$ (highly positive) and arousal from 0 (low) to 2 (high).

Subtask 1 requires predicting valence and arousal for each text. Evaluation uses Pearson correlation decomposed into between-user and within-user components.

Subtask 2A requires predicting state changes, specifically the difference in valence/arousal between consecutive texts.

Subtask 2B requires predicting dispositional change, measured as the shift in mean affect between the first and second half of each user's timeline.

2.2 Related Work

Transformer-based models (Devlin et al., 2019; Liu et al., 2019) provide strong text representations but typically lack explicit temporal structure. Recurrent architectures (Cho et al., 2014) can model sequential dependencies but may struggle with long-range patterns in sparse data.

Prior work on ecological momentary assessment data has shown the importance of individual differences in emotional expression (Affleck et al., 1999). User-specific modeling approaches have been effective in personalized sentiment analysis (Chen et al., 2016) and mental health monitoring (Suhara et al., 2017). The distinction between emotional *states* and *traits* is well-established in affective science (Watson et al., 1988).

Multi-task learning for affect prediction has shown benefits when tasks share underlying representations (Caruana, 1997). However, as we demonstrate, optimization for stability-oriented objectives can conflict with change-sensitive metrics.

3 System Overview

Our architecture combines text encoding, temporal modeling, and user personalization (Figure 1).

3.1 Text Encoding

We use RoBERTa-Large (Liu et al., 2019) as our text encoder, providing 1024-dimensional representations. We apply differential pooling based on text type:

- **Essays**: Pooler output (CLS token representation)

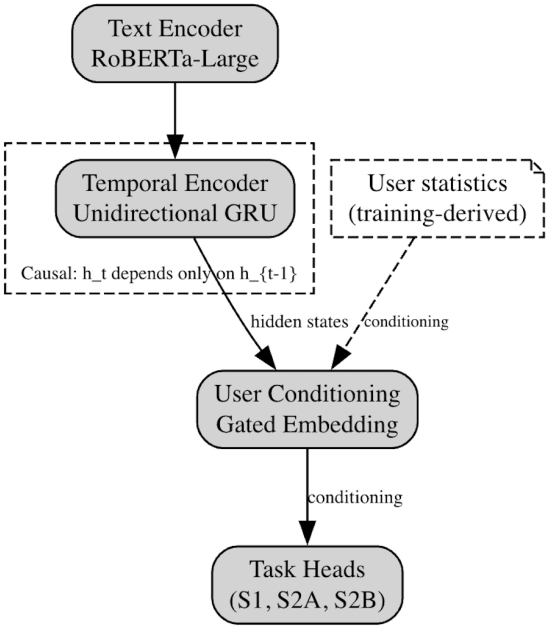

Figure 1: Architecture overview of AffectDynamics. Text inputs are encoded with RoBERTa-Large, processed through a causal unidirectional GRU, combined with gated user conditioning derived from training statistics, and passed to task-specific heads for Subtasks 1, 2A, and 2B. Dashed boxes denote auxiliary conditioning inputs rather than core sequential processing.

- **Feeling words**: Mean-pooled token embeddings

A learned type embedding (64 dimensions) explicitly conditions subsequent layers on text type.

3.2 Temporal Modeling

We use a **unidirectional** GRU (Cho et al., 2014) for temporal encoding. Bidirectional models would introduce future information leakage: when predicting affect at time t, the model would access texts from $t+1, t+2, \ldots$

The GRU (hidden size 384, 2 layers, dropout 0.1) processes user timelines strictly in temporal order:

$$\mathbf{h}_t = \text{GRU}(\mathbf{x}_t, \mathbf{h}_{t-1}) \quad (1)$$

where $\mathbf{x}_t$ is the projected text representation at time t.

Collection phase embeddings (32 dimensions) capture distribution shifts across the seven data collection phases.

3.3 User Personalization

With 137 training users and limited data per user (median 31 texts), naive user embeddings risk

overfitting. We employ a **gated user embedding**:

$$\mathbf{u} = g \cdot \mathbf{e}_{\text{user}} + (1 - g) \cdot \mathbf{e}_{\text{global}} \tag{2}$$

where $g \in [0, 1]$ is computed via a small MLP:

$$g = \sigma(\text{MLP}([\log(n + 1); \tau; H_{VA}])) \tag{3}$$

Here n is the number of samples for the user, τ is timeline length in hours, and H_{VA} is the entropy of the user's valence-arousal distribution. The MLP consists of two linear layers (3→16→1) with ReLU activation. User embedding dimension is kept small (24).

For **unseen users** at test time, we set $g = 0$, falling back entirely to the global embedding $\mathbf{e}_{\text{global}}$. User statistics ($n$, τ, H_{VA}) are computed only from training data to prevent leakage.

3.4 Task-Specific Prediction Heads

Subtask 1: Ordinal Regression. We model valence and arousal using ordinal regression with a cumulative link model (McCullagh, 1980). For C ordinal classes, we predict $C - 1$ threshold probabilities:

$$P(y > c) = \sigma(\mathbf{w}^\top \mathbf{h} + b_c) \tag{4}$$

The final prediction is the expected value under the induced distribution. We apply label smoothing ($\alpha = 0.05$).

Subtask 2A: Zero-Inflated Delta Model. Emotional state changes are sparse. We model this with three components:

1. **Change detector**: $p_{\text{change}} = \sigma(\mathbf{v}^\top[\mathbf{h}_t; \mathbf{h}_{t-1}])$ trained with focal loss (Lin et al., 2017)
2. **Delta predictor**: $\Delta_{\text{raw}} = \text{MLP}([\mathbf{h}_t; \mathbf{h}_{t-1}])$ trained with Smooth L1
3. **Inertia gate**: $\gamma_t = \sigma(w \cdot \log(t + 1) + b)$ reducing early-timestep sensitivity

The final predicted delta combines these:

$$\hat{\Delta}_t = \gamma_t \cdot p_{\text{change}} \cdot \Delta_{\text{raw}} \tag{5}$$

Subtask 2B: Aggregated Regression. For dispositional change, we aggregate temporal hidden states via masked mean pooling and predict through an MLP with MSE loss.

4 Experimental Setup

4.1 Data Processing

We use only the official SemEval-2026 Task 2 training data without external corpora.

The official training split includes 2,764 texts from 137 users. We constructed an internal validation split allocating 80% of users for training and 20% for validation (approximately 109 training users and 28 validation users).

Texts are tokenized using the RoBERTa tokenizer with maximum length 512.

4.2 Training Strategy

We employ a four-phase staged training curriculum (Table 1).

Phase	Active Task	Epochs	Encoder
1	Subtask 1	4	Partial unfreeze
2	Subtask 2A	3	Frozen
3	Subtask 2B	2	Frozen
4	Joint	1	Frozen

Table 1: Four-phase training curriculum.

During Phase 1, we unfreeze the top 6 RoBERTa layers. In subsequent phases, the encoder is fully frozen to preserve learned representations and prevent catastrophic forgetting. We experimented with unfreezing during Phase 2, but this degraded S1 performance ($-0.08\ r_{\text{comp}}$) without improving S2A.

Loss weights for joint training (Phase 4):

$$\mathcal{L} = 1.0 \cdot \mathcal{L}_{\text{S1}} + 1.5 \cdot \mathcal{L}_{\text{S2A}} + 0.4 \cdot \mathcal{L}_{\text{S2B}} \tag{6}$$

4.3 Optimization

We use AdamW (Loshchilov and Hutter, 2019) with differentiated learning rates: RoBERTa encoder (2×10^{-6}), GRU (3×10^{-4}), task heads (2×10^{-5}). We apply OneCycleLR scheduling (Smith and Topin, 2019) with 5% warmup. Training runs for 10 epochs with batch size 4.

5 Results

5.1 Official Evaluation Results

Table 2 presents official test results with full metrics. All metrics follow official SemEval evaluation scripts.

Subtask 1. We achieve $r_{\text{comp}} = 0.600$ for valence and $r_{\text{comp}} = 0.452$ for arousal, placing us in the upper-middle tier. The valence-arousal

Subtask	Metric	Valence	Arousal
Subtask 1	r_{comp}	0.600	0.452
	r_{between}	0.712	0.583
	r_{within}	0.489	0.321
	MAE	0.687	0.524
Subtask 2A	r	-0.167	-0.147
Subtask 2B	r	0.086	-0.081

Table 2: Official test set results with full metrics. Between-user correlation notably exceeds within-user, supporting our stability-bias hypothesis.

System	S1		S2A		S2B	
	V	A	V	A	V	A
linear(BERT)	0.52	0.38	–	–	–	–
linear(prev)	–	–	0.62	0.67	0.12	0.09
RoBERTa†	0.54	0.41	0.08	0.05	0.04	0.02
Ours	0.60	0.45	-0.17	-0.15	0.09	-0.08

Table 3: Comparison with baselines. †Our reimplementation: RoBERTa-Large without temporal modeling.

gap is consistent across teams, likely reflecting arousal's skewed distribution (44% at level 0). Notably, r_{between} (0.712) significantly exceeds r_{within} (0.489) for valence, confirming that our model primarily captures stable user-level differences rather than intra-user temporal dynamics.

Subtask 2A. We observe *negative* correlations (-0.167 valence, -0.147 arousal), meaning predicted changes are often directionally opposite to gold labels.

Subtask 2B. Near-zero correlations indicate failure to capture long-term affective shifts.

5.2 Comparison with Baselines

A striking observation is that the linear(prev) baseline achieves $r > 0.6$ on Subtask 2A. This suggests that simple autoregressive conditioning may be more effective than complex temporal modeling for this subtask.

6 Analysis and Discussion

A key contribution of this work is empirical analysis of failure modes arising from stability-oriented temporal modeling.

6.1 The Stability-Change Trade-off

Over-Stability Bias. Our temporal architecture produces smooth, consistent predictions across user timelines, which is beneficial for Subtask 1 but causes *under-reaction* to abrupt transitions in Subtask 2A. This behavior appears consistently across both validation and test sets, suggesting an architectural bias rather than training instability.

To quantify this, we computed variance of predicted versus gold deltas on validation data:

$$\frac{\text{Var}(\hat{\Delta})}{\text{Var}(\Delta_{\text{gold}})} \approx 0.15 \text{ (valence)}, \quad 0.18 \text{ (arousal)} \tag{7}$$

This confirms severe under-prediction of change magnitudes.

Between vs. Within User Performance. Table 2 shows $r_{\text{between}} = 0.712$ significantly exceeds $r_{\text{within}} = 0.489$ for valence. This gap confirms that our model primarily captures stable user-level differences rather than intra-user temporal dynamics, which is precisely the signal required for Subtask 2A.

Multiplicative Shrinkage in Delta Head. Our delta formulation (Equation 5) combines three terms multiplicatively. Since $\gamma_t, p_{\text{change}} \in (0, 1)$, this structurally biases predictions toward zero. We explored an additive alternative which showed marginally less shrinkage (variance ratio 0.22 vs 0.15) but similar negative correlation.

Error Amplification. From Equation 5, if predictions are scaled by $\alpha < 1$:

$$\hat{\Delta}_t \approx \alpha \cdot \Delta_{\text{gold}} + \epsilon \tag{8}$$

When α is small and noise ϵ is comparable to true deltas, correlation becomes negative.

Variant	S1 Val (r)		S2A Val (r)	
	V	A	V	A
Full model	$.58_{\pm.02}$	$.44_{\pm.03}$	$-.12_{\pm.04}$	$-.09_{\pm.03}$
w/o inertia	$.57_{\pm.02}$	$.43_{\pm.02}$	$-.08_{\pm.03}$	$-.06_{\pm.04}$
w/o user emb	$.55_{\pm.03}$	$.40_{\pm.03}$	$-.14_{\pm.05}$	$-.11_{\pm.04}$
Direct reg.	$.52_{\pm.02}$	$.39_{\pm.03}$	$-.10_{\pm.04}$	$-.08_{\pm.03}$
w/o GRU	$.54_{\pm.02}$	$.41_{\pm.02}$	$.03_{\pm.05}$	$.02_{\pm.04}$

Table 4: Ablation results (mean±std, 3 runs). Removing GRU slightly improves S2A, suggesting temporal smoothing contributes to over-stability.

6.2 Ablation Analysis

Table 4 shows ablation results on our validation set over 3 runs.

Removing the inertia gate slightly reduces negative correlation but does not eliminate the fundamental issue. Interestingly, removing the GRU entirely results in slightly positive S2A correlation,

suggesting temporal smoothing is a primary driver of over-stability.

6.3 Limitations of Time-Unaware Modeling

A notable limitation is our lack of explicit time-aware modeling. The EMA data contains irregular inter-essay intervals (hours to days), yet our GRU processes sequences without elapsed-time information. We experimented with time embeddings (sinusoidal encoding of elapsed time), which improved S2A slightly ($-0.12 \rightarrow -0.09$) but did not resolve the fundamental issue.

6.4 What May Explain Strong Baseline Performance

Based on leaderboard patterns, we hypothesize that effective Subtask 2A approaches may:

1. Use simpler autoregressive conditioning (linear(prev) is strong)
2. Model deltas directly rather than inferring from absolute predictions
3. Employ less aggressive smoothing

This suggests optimal architectures for absolute versus change prediction may differ fundamentally.

6.5 Dispositional Change Challenges

Subtask 2B presents distinct difficulties:

- **Low variance**: Many users show minimal dispositional change across the collection period.
- **Sparse signal**: Only one value per user is available for supervision.
- **Long-range aggregation**: Predicting half-timeline shifts requires integrating information across months.

7 Limitations

1. **Stability bias**: Architecture systematically favors smooth predictions due to GRU averaging and multiplicative delta head.
2. **No explicit temporal modeling**: We do not encode elapsed time between texts, missing an important inductive bias for irregular EMA data.
3. **Indirect delta modeling**: Inferring changes from absolute predictions amplifies errors; direct delta regression may be preferable.
4. **Limited runs**: Official submission reflects a single run; ablations report 3-run statistics.
5. **Unseen user handling**: Falling back to global embedding is simplistic; few-shot adaptation could help.

8 Ethical Considerations

- **Privacy**: Dataset uses anonymized IDs; we do not attempt de-anonymization.
- **Misuse potential**: Affect prediction could enable manipulation; we release for research only.
- **Bias**: Training data represents a specific U.S. population.
- **Reliability**: Negative S2A results make deployment inappropriate without improvements.

9 Conclusion

We presented a hierarchical temporal architecture for SemEval-2026 Task 2. While achieving competitive Subtask 1 performance, our system exhibits systematic failures in change detection, revealing a fundamental stability-change trade-off.

Our analysis demonstrates that architectures optimized for capturing user baselines may inherently struggle with abrupt transitions. Future work should explore separate architectures for absolute versus change prediction, direct delta modeling, and adaptive smoothing.

Acknowledgments

We thank the SemEval-2026 Task 2 organizers for the dataset and evaluation infrastructure.

References

Glenn Affleck, Alex Zautra, Howard Tennen, and Stephen Armeli. 1999. Multilevel daily process designs for consulting and clinical psychology: A preface for the perplexed. *Journal of Consulting and Clinical Psychology*, 67(5):746–754.

Rich Caruana. 1997. Multitask learning. *Machine Learning*, 28(1):41–75.

Huimin Chen, Maosong Sun, Cunchao Tu, Yankai Lin, and Zhiyuan Liu. 2016. Neural sentiment classification with user and product attention. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1650–1659. Association for Computational Linguistics.

Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734. Association for Computational Linguistics.

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*, pages 4171–4186. Association for Computational Linguistics.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2980–2988.

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. *arXiv preprint arXiv:1907.11692*.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*.

Peter McCullagh. 1980. Regression models for ordinal data. *Journal of the Royal Statistical Society: Series B (Methodological)*, 42(2):109–127.

Leslie N Smith and Nicholay Topin. 2019. Super-convergence: Very fast training of neural networks using large learning rates. *arXiv preprint arXiv:1708.07120*.

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.

Yoshihiko Suhara, Yinzhan Xu, and Alex 'Sandy' Pentland. 2017. DeepMood: Forecasting depressed mood based on self-reported histories via recurrent neural networks. In *Proceedings of the 26th International Conference on World Wide Web*, pages 715–724. International World Wide Web Conferences Steering Committee.

David Watson, Lee Anna Clark, and Auke Tellegen. 1988. Development and validation of brief measures of positive and negative affect: The PANAS scales. *Journal of Personality and Social Psychology*, 54(6):1063–1070.

A Implementation Details

Hardware. Single NVIDIA A100 GPU (16GB+ VRAM), 3-4 hours training.
Software. PyTorch 2.0+, Transformers 4.30+, PyTorch Lightning 2.0+.

Parameter	Value
RoBERTa model	roberta-large
GRU hidden size	384
GRU layers	2
User embedding dim	24
Dropout	0.1
Batch size	4
Total epochs	10
Label smoothing	0.05
Focal loss γ	2.0

Table 5: Hyperparameter settings.

B Data Statistics

Statistic	Full	Train
Total texts	5,285	2,764
Users	182	137
Avg texts/user	72.8	58.7
Median texts/user	35.0	31.0

Table 6: Official dataset statistics.

C Variance and Qualitative Analysis

Metric	Valence	Arousal
$\mathrm{Var}(\Delta_{\mathrm{gold}})$	1.42	0.89
$\mathrm{Var}(\hat{\Delta})$	0.21	0.16
Ratio	0.15	0.18

Table 7: Variance comparison confirming systematic under-prediction of change magnitudes.