Ajman University at SemEval-2026 Task 2: Overcoming Scale Collapse in Temporal Emotion Modeling via Residual Learning

Haseebullah Jumakhan, Soud Asaad Alhazba,
Seyed Abdullah, and Mahmoud Al-Ayyoub
Artificial Intelligence Research Center (AIRC)
College of Engineering and Information Technology, Ajman University
{202320029, 202420468, 202412068}@ajmanuni.ac.ae
m.alshboul@ajman.ac.ae

Abstract

The Ajman University Team developed specialized architectures for longitudinal affective forecasting for SemEval-2026 Task 2 (Soni et al., 2026). We established a reliable performance floor with a hybrid contextual-recurrent model in Subtask 1 (ranked 14th). In Subtask 2A (ranked 5th) and Subtask 2B (ranked 6th), our primary contribution is mitigating scale collapse. To achieve this, we introduce a novel "bifurcated leviathan" architecture that combines explicit residual learning with target scaling. We further counteract regression-to-the-mean by optimizing covariance via specialized objective functions (CCC and Huber) (Huber, 1964; Lin, 1989). Finally, we analyze our official leaderboard metrics to empirically demonstrate that while explicit residual learning excels at intra-subject memorization, it poses severe vulnerabilities to zero-shot cross-subject generalization. To support reproducibility, our code is publicly available at https://github.com/Soudk21/NLP_Project/tree/main.

Keywords: Valence-Arousal-Dominance (VAD), Longitudinal Modeling, Temporal Affect, Bifurcated Leviathan, Explicit Residual Learning, SemEval.

1 Introduction

Human emotion is widely represented mathematically via the Circumplex Model of Affect (Russell, 1980), which maps continuous emotional states across a 2-dimensional space utilizing Valence (Positive/Negative) and Arousal (High Energy/Low Energy). SemEval-2026 Task 2 challenges systems to predict these temporal dynamics from naturalistic longitudinal essays, representing weeks or months of a user's affective experience.

Temporal affective modeling in longitudinal text necessitates a delicate trade-off between intra-subject specificity and zero-shot cross-subject generality, particularly when constrained by strict hardware limits (i.e., 8 GB VRAM). Unlike traditional static sentiment classifiers, we focus on architectural optimizations for continuous, autoregressive emotion forecasting.

In this paper, we detail our baseline approach for Subtask 1 and our core methodological contributions for Subtasks 2A and 2B. Specifically, we address "scale collapse" in temporal forecasting by utilizing explicit residual learning, target scaling, and a "bifurcated leviathan" architecture to actively preserve predicted distributional variance across sequences.

2 Subtask 1: Evaluating the Variation in Valence and Arousal

We established a reliable performance floor using a robust hybrid architecture combining a pre-trained transformer backbone with user-specific embeddings.

2.1 Data Processing and Architecture

Our approach uses a `SingleTextDataset` to process the input texts sequentially. The first matrix row is explicitly reserved for unknown/unseen users to enable zero-shot inference, while all subsequent rows are mapped to corresponding training User IDs.

We utilize a combination of trainable user embeddings ($\mathbb{R}^{32}$) and 768-dimensional DistilBERT (Sanh et al., 2019) text embeddings to simultaneously model both the user's inherent baseline disposition and the semantic content of the current text. The concatenated embeddings are passed through a single-layer `BiLSTM` (Hochreiter and Schmidhuber, 1997) (hidden units = 128) to aggregate temporal context. Finally, two isolated Multi-Layer Perceptron (MLP) networks (utilizing ReLU activations and dropout $p = 0.3$) regress the continuous Valence and Arousal values. Specific error constraints regarding this continuous formulation are detailed

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 457–462

in Appendix B.

3 Subtask 2A: Forecasting Emotional State Changes

Unlike the continuous trajectory tracking in Subtask 1, Subtask 2A requires predicting immediate, step-wise affective shifts based on newly authored text. We frame the incoming essay as a semantic catalyst acting upon the user's current emotional baseline, necessitating a fusion strategy that balances high-dimensional text representations with low-dimensional scalar states.

3.1 Preprocessing and Feature Engineering

Data was chronologically ordered by `user_id` and `timestamp`. To strictly prevent longitudinal data leakage, we utilized `GroupShuffleSplit` from the *scikit-learn* toolkit (Pedregosa et al., 2011) to isolate 10% of users as an independent hold-out validation set. Furthermore, the current state inputs (V_t, A_t) were Z-score normalized (μ, σ derived strictly from the training split) prior to network ingestion to prevent scale disparity.

3.2 Model Architecture

We propose a Hybrid Fusion architecture (illustrated in Figure 1) designed to bridge the dimensional gap between high-capacity contextualized textual representations and low-dimensional scalar state inputs.

3.2.1 The Drowning Problem

Initially, we observed that simply concatenating the 768-dimensional textual representation with the 2-dimensional state representation caused the numeric signal to be "overwhelmed" during backpropagation; the network viewed the scalars as negligible noise. To alleviate this, we engineered a **Feature Projection** module to synthetically expand the numeric footprint.

3.2.2 Components of the Model

1. **Text Encoder:** A `microsoft/deberta-v3-base` encoder maps the essay x_t into a dense vector $h_{text} \in \mathbb{R}^{768}$ utilizing the [CLS] token (He et al., 2021).

2. **State Projection and Fusion:** The normalized current state is projected onto a 64-dimensional space via an MLP (Linear $\rightarrow$ GELU $\rightarrow$ Dropout). This projected numeric representation h_{num} is concatenated with h_{text}. The 832-dimensional fused representation is regulated via LayerNorm to stabilize variance before being routed through a final regression head to yield $[\Delta\hat{V}, \Delta\hat{A}]$.

3.3 Experimental Design and Ablation Study

3.3.1 Phase 1: Stabilization

Our initial model employed a Weighted MSE loss. However, it suffered from **Gradient Explosions**. We transitioned to **Huber Loss** (SmoothL1) and applied gradient norm clipping. While training stabilized, the model lazily predicted near-zero changes for all inputs, yielding poor correlation ($r \approx 0.31$).

3.3.2 Phase 2: Feature Engineering

We introduced the 64-dimensional projection MLP (Section 3.2) and applied LayerNorm. Correlation increased to $r \approx 0.39$, allowing the model to utilize numeric context, though predictions remained overly conservative.

3.3.3 Phase 3: Optimization via CCC Loss

Euclidean losses inherently cause regression to the mean; minimizing MSE incentivizes predicting the dataset average (zero change). We replaced the loss function with **Concordance Correlation Coefficient (CCC) Loss**, forcing the model to optimize for covariance matching:

$$\mathcal{L}_{CCC} = 1 - \frac{2\sigma_{y\hat{y}}}{\sigma_y^2 + \sigma_{\hat{y}}^2 + (\mu_y - \mu_{\hat{y}})^2}$$

where $\sigma_{y\hat{y}}$ is the covariance, σ_y^2 and $\sigma_{\hat{y}}^2$ are the respective variances, and μ_y and $\mu_{\hat{y}}$ are the means. This transition, combined with differential learning rates, culminated in a massive internal performance leap to an average correlation of $r = 0.642$.

3.4 Qualitative Error Analysis

Manual inspection of specific textual failure cases reveals that while the projected architecture tracks overt emotional shifts well, it struggles with complex pragmatics. For instance, when a user vented using ostensibly positive vocabulary in a sarcastic tone (*"Oh great, another perfectly ruined weekend"*), the model naively interpreted the standalone token embeddings as indicators of a positive valence shift, failing to capture the underlying linguistic irony.

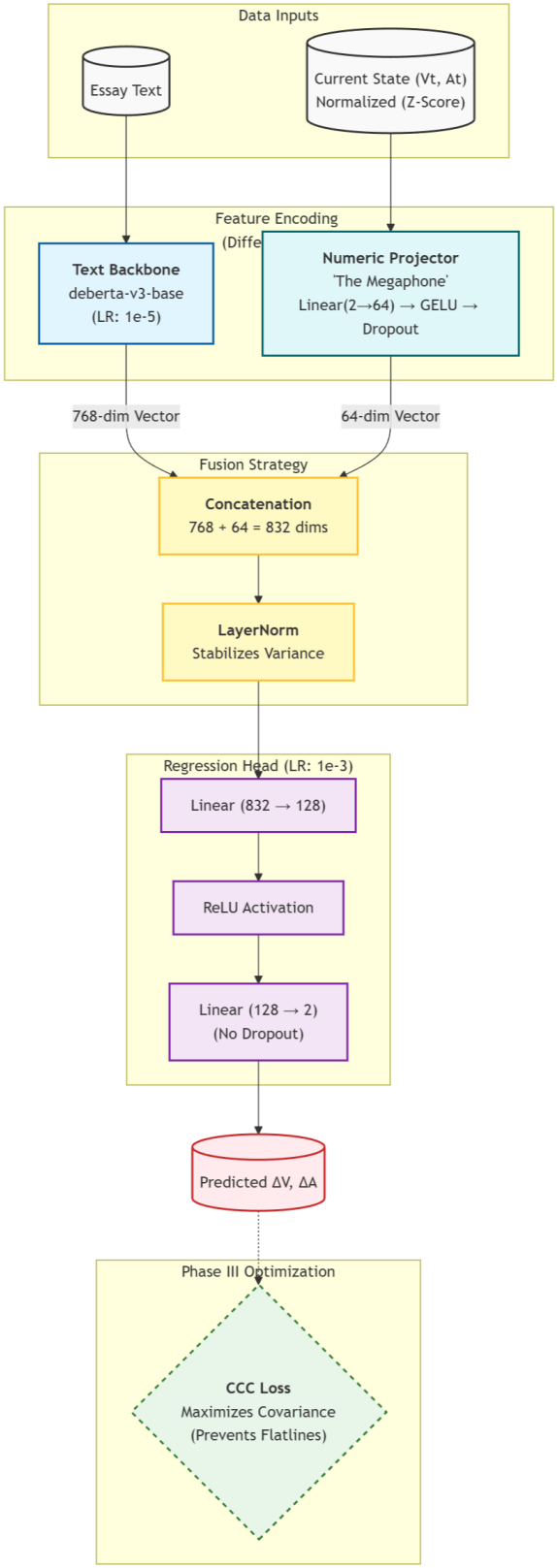

Figure 1: Proposed Framework for Subtask 2A showing the Feature Projection path to solve the "Drowning" problem.

4 Subtask 2B: Longitudinal Trajectory

Subtask 2B extends the forecasting horizon from immediate state changes to long-term dispositional drifts across a user's entire available history. Because mapping macro-level trajectorie

4.1 Preprocessing and the "Leviathan" Protocol

Siamese Sampling We chronologically ordered each user's history, extracting the **Head** (initial k essays) and **Tail** (final k essays). This frames the problem identically to a Siamese network (Bromley et al., 1993), forcing the evaluation of the delta between two temporal poles.

Explicit Residual Learning via Naive Injection To combat Scale Collapse, we calculated the "Naive Statistical Change" directly from the Head and Tail segments: $\Delta_{naive} = \mu(Y_{tail}) - \mu(Y_{head})$. We explicitly inject this scalar directly into the final regression head. Mathematically, this shifts the architecture to **Residual Learning** (He et al., 2016). The deep network merely learns a contextual refinement function $F(X)$ such that the prediction is $\Delta_{naive} + F(X)$.

Target Scaling We mapped the ground truth targets to a standard normal distribution during training: $Y_{scaled} = (Y - \mu_Y)/(\sigma_Y + 1e^{-8})$. While real-world emotional shifts may exhibit skewed or heavy-tailed distributions rather than a perfect Gaussian curve, empirical results proved that Z-score scaling acts as a critical variance-stabilizing regularizer. The benefit of preventing the optimizer from collapsing to a zero-mean vastly outweighed the risk of minor distributional mismatches. Predictions are inverse-transformed during inference.

4.2 Model Architecture: Bifurcated Siamese

The "Bifurcated Leviathan" leverages a `deberta-v3-large` backbone with gradient checkpointing (Figure 2). We employed a custom **Siamese Difference Pooling** mechanism ($h_\Delta = h_{tail} - h_{head}$). To eliminate Task Dominance, the network bifurcates immediately after the backbone into completely isolated MLPs for Valence and Arousal, preventing the noisy loss landscape of Arousal from polluting Valence convergence (Chen et al., 2018).

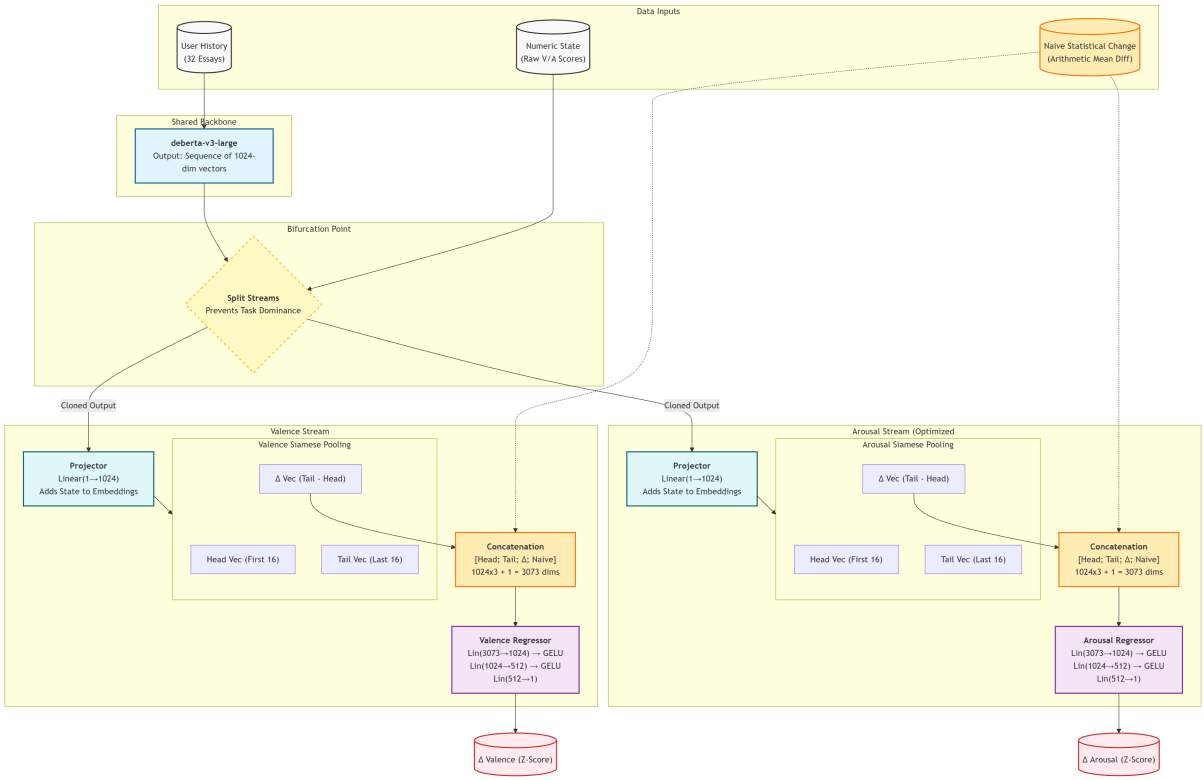

Figure 2: The Bifurcated Siamese architecture utilized in Subtask 2B, featuring Naive Feature Injection.

4.3 Scale Collapse Mitigation

In early shared-head runs without target scaling, we experienced severe **Scale Collapse** (Jing et al., 2022). Ground truth targets ranged $[-2.0, 2.0]$, but predictions tightly clustered within $[-0.05, 0.05]$ with a collapsed variance ($\sigma^2 < 0.01$). Integrating Target Scaling and Naive Feature Injection successfully restored predicted variance to match the ground-truth distribution. On our internal hold-out validation set, this architecture achieved Pearson correlations of $r = 0.70$ for Valence.

5 Official Results & Distributional Shift

Table 1 highlights our official performance on the hidden SemEval evaluation set.

Subtask	Rank	Valence (r)	Arousal (r)	Average
1	**16th**	0.656	0.439	0.548
2A	**7th**	0.615	0.670	0.642
2B	**8th**	-0.124	0.456	0.166

Table 1: Official SemEval-2026 Leaderboard Results.

5.1 Vulnerability to Noisy Baselines

The contrast between our internal validation and the official metrics yields a profound insight into longitudinal generalization. In Subtask 2B, our model achieved a strong $r = 0.70$ internally for Valence, yet plummeted to $r = -0.124$ on the unseen official test set.

This drastic divergence explicitly validates a critical vulnerability of our **Explicit Residual Learning** approach. Because the network delegates the arithmetic baseline entirely to the mathematical injection of Δ_{naive}, it became catastrophically susceptible to distribution shifts. If unseen users in the hidden test set anchored their baseline self-reported scores differently (or more noisily) than the training demographic, the arithmetic baseline inherently misled the entire refinement network.

Conversely, our Subtask 2A architecture—which utilized Feature Projection and CCC Loss without naive statistical injection—generalized exceptionally well (achieving $r = 0.670$ for Arousal, ranking 5th globally). This proves that while explicit residual injection facilitates deep intra-subject memorization, optimizing for covariance via CCC loss is a far more robust mechanism for zero-shot generalizability.

6 Conclusion

Team Ajman University demonstrates that capturing longitudinal trajectories requires prioritizing context width over deep recurrence. Overcoming

scale collapse necessitates abandoning standard Euclidean loss in favor of projecting numeric features and utilizing concordance (CCC) loss. While residual learning successfully bypasses arithmetic bottlenecks, our official results (Subtask 2A vs. 2B) empirically prove that explicit mathematical injection is highly vulnerable to noisy baseline reporting in unseen users.

7 Limitations

Our primary limitation is reliance on consumer-grade hardware (8GB–24GB VRAM). This necessitated FP16 precision and gradient accumulation steps to satisfy the batch-variance requirements of CCC loss. Furthermore, these physical constraints drove our most complex architectural interventions. The "Bifurcated Leviathan" is highly specialized to bypass local memory limits, making the code less generalizable across differing computing environments.

8 Ethical Considerations

Predicting human affect from ecological text carries privacy and psychological implications. Our models inevitably inherit algorithmic biases present in the training demographic's phrasing. These systems should strictly not be utilized for automated medical diagnosis, psychiatric screening, or mood monitoring without human oversight. This research is intended exclusively for longitudinal study under anonymized conditions.

References

Jane Bromley, Isabelle Guyon, Yann LeCun, Eduard Säckinger, and Roopak Shah. 1993. Signature verification using a "siamese" time delay neural network. In *Advances in neural information processing systems*, volume 6, pages 737–744.

Zhao Chen, Vijay Badrinarayanan, Chen-Yu Lee, and Andrew Rabinovich. 2018. Gradnorm: Gradient normalization for adaptive loss balancing in deep multitask networks. In *International Conference on Machine Learning*, pages 794–803. PMLR.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. Debertav3: Improving deberta using electra-style pretraining with gradient-disentangled embedding sharing. *arXiv preprint arXiv:2111.09543*.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Peter J Huber. 1964. Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, 35(1):73–101.

Li Jing, Pascal Vincent, Yann LeCun, and Yuandong Tian. 2022. Understanding dimensional collapse in contrastive self-supervised learning. In *International Conference on Learning Representations*.

Lawrence I-Kuei Lin. 1989. A concordance correlation coefficient to evaluate reproducibility. *Biometrics*, 45(1):255–268.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, and 1 others. 2011. Scikit-learn: Machine learning in python. *Journal of machine learning research*, 12(Oct):2825–2830.

James A Russell. 1980. A circumplex model of affect. *Journal of personality and social psychology*, 39(6):1161–1178.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv 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.

A Experimental Setup and Reproducibility

To facilitate full reproducibility, we document the precise configurations separating the subtasks. All experiments were conducted using PyTorch on NVIDIA RTX 4070 and 3090 GPUs.

- **Subtask 1:** Trained using standard AdamW. Sequence windows were padded to 10 entries during inference to avoid context starvation.
- **Subtask 2A:** `DeBERTa-v3-base`. Differential learning rates were applied: $1e^{-5}$ for the transformer backbone, and $1e^{-3}$ for the custom regression heads. Gradient accumulation achieved an effective batch size of 32.

- **Subtask 2B:** `DeBERTa-v3-large`. Gradient Checkpointing was mandatory to process up to 32 essays simultaneously. We utilized $5e^{-6}$ for the backbone and $1e^{-4}$ for the heads.

A universal random seed of 42 was utilized across all data splitting (`GroupShuffleSplit`) and initializations.

B Subtask 1 Error Analysis Constraints

Because Subtask 1 involves predicting continuous longitudinal trajectories rather than discrete categories, generating a standard classification confusion matrix is mathematically inapplicable. Instead, our error analysis relies on the qualitative distribution of continuous residuals. Granular analysis reveals that the model struggles significantly more with the Arousal dimension than with Valence (reflected in the official scores of $V = 0.656$ vs $A = 0.439$). Valence is often explicitly semantically defined via vocabulary, whereas Arousal is defined implicitly through structural intensity, pacing, and punctuation. Standard DistilBERT embeddings consistently fail to capture these implicit stylistic markers without explicit hand-crafted feature engineering.