Perspicere at SemEval-2026 Task 2: Modeling Longitudinal Valence and Arousal via Dense Embeddings and Agentic Reasoning

Kamyar Moradian Zehab, Mohammad Sadegh Poulaei, Nasser Mozayani
Iran University of Science and Technology
{kamyar_moradian, m_poulaei}@comp.iust.ac.ir, mozayani@iust.ac.ir

Abstract

This paper presents our system for SemEval 2026 Task 2 (Subtask 1), modeling affect assessment as a longitudinal trajectory. We evaluate a tripartite affective framework of escalating contextual complexity, spanning zero-context feature extraction, latent temporal modeling via LSTM, and explicit semantic reasoning via the Teacher-Guided Clinical Reasoning Agent utilizing in-context learning. Our results show that robust static extraction outperforms explicit sequence modeling. Specifically, Matryoshka-distilled embeddings (Jasper) paired with XGBoost provided the best balance of speed and accuracy when utilizing the full training corpus (Valence composite $r = 0.654$, a 17.4% improvement compared with the baseline), mitigating the severe overfitting observed on partitions of the dataset. Additionally, we uncover a distinct agentic advantage: although the reasoning agent trailed mathematical regressors in tracking high-frequency fluctuations, its SOTA psychological profiling yielded the highest Between-User Valence correlation ($r = 0.725$), demonstrating its efficacy in user-level affective profiling. Finally, a persistent "arousal bottleneck" confirms the limitations of text-only modeling for physiological activation.

1 Introduction

Affective Natural Language Processing is moving beyond static sentiment classification to model emotion as a continuous, path-dependent trajectory governed by emotional inertia (Kuppens et al., 2010, 2012; Xu et al., 2024). Addressing this paradigm shift, SemEval-2026 Task 2 (Subtask 1) (Soni et al., 2026) challenges systems to predict continuous Valence and Arousal scores from chronological sequences of "ecological essays" and "feeling words." Grounded in Russell's Circumplex Model (Russell, 1980), the task requires tracking dynamic mood fluctuations over time while also generalizing to entirely new subjects. While standard Independent and Identically Distributed (i.i.d.) regressors ignore this temporal complexity, we hypothesize that when powered by modern, high-density embeddings, they remain highly effective for extracting an entry's immediate core semantic state. However, because these mathematical regressors act as black boxes, they lack the explainable nuance and clinical interpretability beneficial for a robust psychological assessment.

To bridge this gap, we propose a tripartite affective framework of escalating contextual complexity that integrates high-precision mathematical regression with interpretable clinical deduction. Our system isolates specific affective dimensions across three parallel paradigms: (1) a static paradigm handling zero-context isolated feature extraction, leveraging state-of-the-art dense embeddings to extract the immediate emotional state; (2) a longitudinal paradigm introducing latent temporal context by explicitly modeling temporal smoothing and path-dependencies via Long Short-Term Memory (LSTM) networks (Hochreiter and Schmidhuber, 1997); and (3) an agentic paradigm introducing explicit semantic reasoning and temporal dynamics via a Teacher-Guided Clinical Reasoning Agent (TG-CRA) that leverages in-context learning using few-shot exemplars from a frontier Large Language Model (LLM) for explainable user profiling.

To evaluate this framework, we rigorously benchmarked modern embedding architectures—including Jasper (1.9B) (Zhang et al., 2024), Qwen3 (0.6B/4B) (Zhang et al., 2025), and mE5-Large (Wang et al., 2024)—against legacy BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019) baselines. From these experiments, our results highlight three key contributions. First, we identify that Matryoshka-distilled embeddings (Jasper) (Kusupati et al., 2024) combined with gradient-boosted trees (XGBoost) (Chen and Guestrin, 2016) establish a superior density-efficiency frontier. While highly prone to overfit-

Proceedings of the The 20th International Workshop on Semantic Evaluation (2026), pages 671–685

ting in dense semantic spaces under low-data conditions, providing the complete training corpus mitigates this issue, allowing XGBoost to robustly map complex, non-linear affective boundaries. Second, we demonstrate that robust static extraction consistently dominates explicit recurrent sequence modeling via LSTMs, which struggle to map the sparse, irregular chronological intervals between entries. Third, we uncover a distinct agentic advantage: although the Clinical Agent trails vector-based regressors in tracking high-frequency fluctuations, it achieves the highest Between-User Correlation for baseline profiling among our evaluated systems, outperforming all mathematical regressors in establishing psychological traits. This quantitative success is underpinned by the agent's qualitative clinical robustness; unlike black-box regressors, it correctly interprets non-standard orthography and contextual mitigators, such as spiritual terms softening negative affect, demonstrating a vital capacity for the complex pragmatics of emotional expression.

2 Related Work

Longitudinal Affect Modeling Traditional approaches to Affective Computing have largely treated emotion classification as a static, isolated task. Early research relied on lexicon-based methods and feature engineering (Pang et al., 2008; Liu, 2012) or Convolutional Neural Networks (CNNs) (Kim, 2014), before shifting to pre-trained Transformers like BERT and RoBERTa (Devlin et al., 2019; Liu et al., 2019), which became the standard for sentiment analysis. However, these models typically capture "perceived" affect from third-party annotations rather than the subject's internal state. Psychological research posits that emotion is a dynamic, path-dependent process governed by *Emotional Inertia* (Kuppens et al., 2010, 2012). This motivates a shift from Independent and Identically Distributed (i.i.d.) regression to exploring longitudinal modeling. While Recurrent Neural Networks (RNNs) (Hochreiter and Schmidhuber, 1997) and context-aware architectures like DialogueRNN (Majumder et al., 2019) have been applied to conversation, SemEval-2026 Task 2 (Subtask 1) extends this to ecological, first-person essays for tracking long-term affective rhythms (Soni et al., 2026).

Clinical Agentic Frameworks The advent of Large Language Models (LLMs) (Brown et al., 2020) has transformed computational psychology from simple classification to complex reasoning (Xu et al., 2024; Giannos and Delardas, 2024). This has enabled agentic frameworks where models utilize Chain-of-Thought (CoT) prompting (Wei et al., 2022) to articulate the rationale behind a diagnosis, a capability recently demonstrated in mental health profiling tasks (Yang et al., 2023). In our work, we adopt this paradigm via the Clinical Agent (TG-CRA), bridging the gap between proprietary frontier models and open-weights systems using teacher-guided in-context learning. By leveraging a frontier teacher model to generate high-quality reasoning traces as few-shot anchors, we guide an open-weights LLM to perform robust, explainable psychological profiling and baseline assessment.

Dense Representation Learning While the BERT family (Devlin et al., 2019) remains a strong baseline, the landscape of text representation has evolved with models that leverage massive corpora to capture deeper semantic nuances, such as the billion-scale contrastively pre-trained mE5 (Wang et al., 2024) and LLM-derived embeddings like Qwen3 (Zhang et al., 2025). Critically, techniques like Matryoshka Representation Learning (MRL) (Kusupati et al., 2024) now allow models to encode coarse-to-fine information in initial dimensions, enabling effective truncation. Our system utilizes Jasper (Zhang et al., 2024), a compact 1.9B model that combines multi-stage knowledge distillation from massive teacher models with MRL dimensionality reduction. We posit that the resulting high semantic density of these embeddings scales exceptionally well with non-linear, tree-based ensembles like XGBoost (Chen and Guestrin, 2016).

3 System Overview

Our system architecture is designed to address the core challenge of SemEval 2026 Task 2 (Subtask 1): modeling emotion not as a static snapshot, but as a dynamic, lived experience grounded in longitudinal context (Soni et al., 2026). Instead of a sequential pipeline, we propose a tripartite affective framework consisting of three parallel paradigms of escalating contextual complexity: (1) **Zero-context Static Regression**, which establishes a high-performance foundation using strong pre-trained representations without historical context; (2) **Latent Temporal Contextualization**, which explicitly captures path-dependencies and temporal smoothing via LSTM architectures; and (3)

Explicit Semantic Reasoning (TG-CRA), which leverages LLMs to perform deductive clinical reasoning over user history and emotional trajectories.

3.1 Data Preprocessing and Representation

The dataset consists of two distinct text modalities: free-text ecological essays and lists of feeling words. Because our mathematical regressors and temporal networks (Paradigms 1 and 2) require a unified semantic space, we apply a template-based transformation prior to vectorization. Paradigm 3 bypasses this entire preprocessing and embedding pipeline to operate directly on the unadulterated raw text. For any input x_i representing a list of feeling words, we execute this transformation via a formal concatenation:

$$x_{\text{transformed},i} = \text{“I am feeling ”} \oplus x_i$$

This natural language prompt converts disjoint labels (e.g., “Tired, Calm”) into a syntactic structure comparable to the essays, mitigating domain shift for pre-trained models. Conversely, the ecological essays are preserved in their original free-text format without any template modification, ensuring their natural narrative structure remains intact.

Feature Extraction and Normalization: To encode the processed texts into dense vector representations, we utilize a diverse set of pre-trained Transformer backbones. To ensure robust coverage of both linguistic and affective features, we employ three classes of frozen backbones: (1) **Bidirectional Encoder Embeddings:** BERT-base-uncased (Devlin et al., 2019), RoBERTa (Base/Large) (Liu et al., 2019), and XLM-RoBERTa (Base/Large) (Conneau et al., 2020) serve as our foundational baselines, while E5-Large (Wang et al., 2024) scales up the bidirectional encoder architecture through extensive pre-training to better capture deep semantic dependencies; (2) **Generative LLM Embeddings:** Qwen3 (0.6B and 4B) (Zhang et al., 2025) represent a paradigm shift, utilizing causal attention and LLM reasoning capabilities to capture deeper contextual nuances; and (3) **Distilled Multimodal Embeddings:** Jasper (Zhang et al., 2024), a 1.9B-parameter student model distilled from NV-Embed-v2 and Stella, adapts Matryoshka Representation Learning (MRL) (Kusupati et al., 2024) to extract high-density features, offering performance comparable to 7B parameter models with significantly lower computational overhead.

All backbones are frozen during training. We extract the raw embedding $\mathbf{e}_t$ for the text at time t by applying a pooling operation over the token sequence. To ensure scale invariance across different backbones, we apply L2 normalization directly: $\hat{\mathbf{e}}_t = \mathbf{e}_t / \|\mathbf{e}_t\|_2$. These resulting normalized embeddings $\hat{\mathbf{e}}_t$ are then routed exclusively into the independent evaluation tracks of Paradigms 1 and 2.

3.2 Paradigm 1: Zero-Context Static Regression

Our first parallel paradigm treats each text entry x_t as an independent and identically distributed (i.i.d.) sample, ignoring the temporal sequence. This paradigm aims to determine the extent to which valence and arousal can be inferred from linguistic cues alone, without historical context. We implement three classes of regressors on top of the frozen embeddings: (1) **Linear Ridge Regression**, a regularized linear model to prevent overfitting on high-dimensional embeddings; (2) **XGBoost** (Chen and Guestrin, 2016), a gradient-boosted decision tree framework designed to capture non-linear interactions between embedding dimensions; and (3) **Support Vector Regression (SVR)** (Drucker et al., 1997), employing both linear and RBF kernels to model complex decision boundaries in the semantic space. For all Paradigm 1 models, the predicted valence and arousal scores $(\hat{v}_t, \hat{a}_t)$ are functions solely of the current normalized text embedding $\hat{\mathbf{e}}_t$.

3.3 Paradigm 2: Latent Temporal Contextualization

The core premise of longitudinal affect tracking suggests that an individual’s current emotional state is a function of their history. To capture this, we implement an LSTM network (Hochreiter and Schmidhuber, 1997). We specifically deploy LSTMs as a classical, explicit baseline for longitudinal temporal smoothing to strictly isolate the impact of the dense embeddings, reserving the complexity of transformer-based sequence models for future work. For a user u, we construct the history sequence $H_u = [\hat{\mathbf{e}}_1, \hat{\mathbf{e}}_2, \ldots, \hat{\mathbf{e}}_T]$, where $\hat{\mathbf{e}}_t$ is the normalized embedding of the t-th entry. The architecture is defined as:

$$\mathbf{h_t}, \mathbf{c_t} = \text{LSTM}(\hat{\mathbf{e}}_t, \mathbf{h}_{t-1}, \mathbf{c}_{t-1}),$$
$$(\hat{v}_t, \hat{a}_t) = \text{MLP}(\mathbf{h}_t),$$

where $\mathbf{h}_t$ is the hidden state representing the latent emotional trajectory at time t.

Handling Variable History: To dynamically process chronological user sequences of unequal duration, we employ a dual-stage padding mitigation approach. First, dynamic sequence packing ensures the recurrent network computes hidden states solely from valid, observed timesteps. Second, boolean masking during the loss calculation prevents padding artifacts from distorting backpropagation gradients, faithfully preserving the exact intra-subject temporal progression.

3.4 Paradigm 3: Teacher-Guided Clinical Reasoning Agent (TG-CRA)

Our most advanced approach frames longitudinal affect assessment as clinical deductive reasoning. Powered by GPT-OSS-120B (OpenAI et al., 2025), this paradigm eschews latent state vectors to operate directly on the raw, interpretable textual context window. For each query, the Teacher-Guided Clinical Reasoning Agent (TG-CRA; hereafter, the Clinical Agent) dynamically constructs a contextual prompt anchored in Russell's Circumplex Model of Affect (Russell, 1980). This prompt integrates three critical variables: the user's available chronological history, the exact time elapsed between entries to track emotional decay, and the input modality. By recognizing this modality, the agent seamlessly switches between evaluating explicit feeling words and inferring latent emotions from essays.

To prevent unstructured generation and enforce clinical rigor, the Clinical Agent is constrained via a strict JSON output schema. Before yielding numerical predictions—which are strictly typed to single-decimal floats—the model must articulate a CoT (Wei et al., 2022) reasoning schema progressing through four required diagnostic steps: **(1) Baseline Assessment:** Analyzing the user's general emotional disposition based on history; **(2) Recent Trajectory:** Identifying directional trends such as recovery or deterioration; **(3) Target Analysis:** Examining specific linguistic cues and modality constraints of the current text; and **(4) Time Factor Weighting:** Evaluating the impact of elapsed time on emotional continuity.

To bridge the gap between open-weights models and state-of-the-art proprietary deduction, we utilize teacher-guided in-context learning. High-fidelity rationales are generated by a stronger teacher model (Gemini-3.0-Pro) conforming to the exact same four-step JSON CoT schema expected of the agent. These rationales are dynamically injected as few-shot anchors to serve as algorithmic priors. Crucially, these anchors adapt based on the evaluation track: **Seen Users** receive anchors modeling multi-step historical trajectory tracking, while **Unseen Users** receive cold-start anchors teaching the agent to resolve mixed-affect states while gracefully defaulting historical fields to non-applicable states. The complete system persona and structured JSON prompt schemas are detailed in Appendix B.

4 Experimental Setup

4.1 Dataset and Evaluation Metrics

We utilized the official SemEval-2026 Task 2 (Subtask 1) (Soni et al., 2026). To prevent data leakage, we employed an internal validation split (80/20 user-level), ensuring each user's entries remain in a single fold. For Paradigm 1 and 2 models, hyperparameter grid searches were conducted on this held-out set before full retraining. The test set evaluates two groups: **Seen Users ($\mathcal{U}_{seen}$)** (future timesteps of training users) to assess longitudinal modeling $P(y_{t+k}|y_{1:t})$, and **Unseen Users ($\mathcal{U}_{unseen}$)** (completely held-out subjects) to evaluate zero-shot generalization $P(y|x)$. We adopt the official evaluation metrics proposed by the task organizers (Soni et al., 2026), prioritizing both trait-level assessment and state-level dynamics. For Valence and Arousal, we report three primary Pearson correlation metrics: Between-User ($r_{between}$), Within-User (r_{within}), and the primary ranking Composite Correlation ($r_{composite}$), which aggregates them via Fisher's z-transformation. We also report their Mean Absolute Error (MAE) counterparts to quantify prediction error magnitude.

4.2 Implementation and Optimization

Our system was implemented in Python 3.10. We utilized PyTorch (Paszke et al., 2019) for sequence modeling, alongside scikit-learn (Pedregosa et al., 2011) and XGBoost (Chen and Guestrin, 2016) for static baseline regressors. Pre-trained Transformer backbones were sourced via the HuggingFace libraries (Wolf et al., 2020). Agentic inference and teacher-guided rationale generation were conducted using the Groq API[1] and the Google Gemini platform[2], respectively. Experiments for Paradigms 1 and 2 were executed on a single NVIDIA T4 GPU via Google Colab. For optimization, Paradigm 2 LSTMs utilized the AdamW opti-

[1] https://groq.com/
[2] https://gemini.google.com/app

Backbone	Method	Valence (r)			Arousal (r)		
		W	B	C	W	B	C
BERT-Base (Soni et al., 2026)	Ridge	0.435	0.659	0.557	0.253	0.343	0.299
BERT-Base	Ridge	0.504	0.651	0.583	0.312	0.531	0.428
	SVR	0.508	0.667	0.593	0.355	0.533	0.448
	XGBoost	0.493	0.675	0.592	0.345	0.540	0.447
RoBERTa-L	Ridge	0.517	0.619	0.570	0.294	0.538	0.424
	SVR	0.512	0.614	0.565	0.294	0.621	0.474
	XGBoost	0.491	0.633	0.567	0.306	0.443	0.376
Jasper	Ridge	0.562	0.684	0.627	0.419	**0.629**	**0.532**
	SVR†	0.555	0.682	0.623	0.392	0.590	0.497
	XGBoost	**0.579**	0.718	**0.654**	0.421	0.562	0.495
	LSTM	0.527	0.702	0.622	**0.423**	0.547	0.488
Qwen3-4B	Ridge	0.550	0.685	0.622	0.418	0.590	0.509
	SVR	0.551	0.687	0.623	0.405	0.529	0.470
	XGBoost	0.528	0.691	0.616	0.390	0.516	0.456
	LSTM	0.448	0.662	0.564	0.391	0.533	0.465
GPT-OSS-120B	TG-CRA	0.529	**0.725**	0.637	0.362	0.565	0.470

Table 1: Primary correlation metrics (r) for representative backbones. W, B, and C refer to *Within-User*, *Between-User*, and *Composite*. †Official submission for SemEval-2026 Task 2 (Subtask 1); see Section 5 for submission details. See Appendix C for MAE scores and the exhaustive tabulation of all evaluated models.

mizer (Loshchilov and Hutter, 2019) with a masked Mean Squared Error (MSE) loss to ignore padding artifacts, alongside gradient clipping (max norm 1.0). The inference-only Paradigm 3 Agent utilized a 20-entry sliding context window, a 0.85 generation temperature, and JSON-mode constrained decoding. Grid search spaces, final selected hyperparameters for Paradigms 1 and 2, and full configuration search details for Paradigm 3 are provided in Appendix A and Appendix B.4, respectively.

5 Results and Analysis

5.1 Quantitative Performance Overview

Table 1 presents the primary correlation metrics (r) for our representative backbones—BERT-Base and RoBERTa-Large (Legacy Transformers) (Devlin et al., 2019; Liu et al., 2019), Jasper and Qwen3-4B (SOTA Embeddings) (Zhang et al., 2024, 2025)—evaluated across the parallel tracks of our tripartite framework: Zero-Context Static Regression (Paradigm 1, utilizing Ridge, SVR, and XGBoost regressors) (Drucker et al., 1997; Chen and Guestrin, 2016), Latent Temporal Contextualization (Paradigm 2, utilizing LSTM) (Hochreiter and Schmidhuber, 1997), and Explicit Semantic Reasoning (Paradigm 3, via the Agentic Framework) (OpenAI et al., 2025; Wei et al., 2022; Brown et al., 2020). A complete tabulation including Mean Absolute Error (MAE) for all evaluated models and configurations is provided in Appendix C.

Backbone Dominance: Modern architectures consistently outperform the legacy BERT family in extracting immediate affective states via Paradigm 1. Specifically, Jasper, leveraging Matryoshka-style distillation (Kusupati et al., 2024), establishes the superior density-efficiency frontier among mathematical regressors (see Section 5.3 for hardware metrics). Jasper paired with XGBoost (Chen and Guestrin, 2016) achieves the highest Valence state-level tracking ($r_{within} = 0.579$) and Composite score ($r_{composite} = 0.654$). For the mathematically resistant Arousal dimension, Jasper with Ridge regression secures the top Composite score ($r_{composite} = 0.532$).

The Agentic Trade-off (Profiling vs. Tracking): The Clinical Agent (Paradigm 3) exhibits a critical performance divergence. While trailing Jasper XGBoost in tracking Valence state fluctuations (r_{within} 0.529 vs. 0.579), it achieves the highest Valence Between-User correlation score across all models ($r_{between}$ 0.725), though Jasper XGBoost retains the edge in the overall Composite metric ($r_{composite}$ 0.654 vs. 0.637). The explicit reasoning agent clearly excels at user-level affective profiling, whereas mathematical regressors remain superior for overall high-frequency tracking.

5.2 Longitudinal and Head Sensitivity Analysis

Static Dominance over Sequence Modeling: Empirical evaluation reveals that explicit sequence modeling via LSTM (Paradigm 2) generally fails to improve upon optimized static baselines (Paradigm 1). For instance, Qwen3-4B's (Zhang et al., 2025) Valence $r_{composite}$ drops from 0.623 (SVR) (Drucker et al., 1997) to 0.564 (LSTM), and Jasper's (Zhang et al., 2024) static XGBoost (0.654) (Chen and Guestrin, 2016) outpaces its recurrent counterpart (0.622). We posit that while static regressors robustly isolate immediate affective states, classical recurrent architectures likely overfit to sparse, irregular intervals. This suggests that fully leveraging longitudinal context in high-density semantic spaces requires Transformer-based temporal integration systems capable of modeling complex time dependencies.

Ablation on Data Volume and Head Scalability: Embedding density interacts strongly with data volume, revealing a distinct performance inversion between regression heads. In preliminary ablations on restricted, stratified data subsets (detailed in Appendix D), regularized and margin-

based models demonstrated superior robustness against the noise of sparse semantic spaces. Under these low-data conditions, Jasper paired with SVR achieved a Valence $r_{composite}$ of 0.654, decisively outperforming the severe overfitting of XGBoost ($r_{composite} = 0.572$). However, scaling to the complete training corpus entirely inverted this hierarchy. The full data manifold mitigated tree-based overfitting, allowing XGBoost to robustly map complex non-linear affective boundaries and establish the overall state-of-the-art for state-level tracking (Valence $r_{within} = 0.579$). Consequently, our official SemEval submission featured the Jasper-SVR architecture optimized during these initial subset evaluations, whereas the superior XGBoost configuration represents a post-submission refinement achieved by utilizing the complete training corpus.

Valence/Arousal Gap: Across all models, predicting Arousal remains significantly more challenging (peaking at $r_{composite} \approx 0.53$) than Valence (peaking at $r_{composite} \approx 0.65$). This aligns with established affective computing literature demonstrating that arousal or activation is better accessed via acoustic features like speech prosody, whereas valence is more effectively captured by linguistic or semantic features (Schuller, 2018; Atmaja and Akagi, 2021). Because our models rely purely on text, they lack the multimodal physiological or acoustic signals necessary to resolve this arousal bottleneck.

5.3 Computational Efficiency

Backbone	Size	Time/Batch	Rel. Speed
BERT-Base	∼110M	0.20s	1.0x
RoBERTa-Base	∼125M	0.19s	∼1.0x
RoBERTa-Large	∼355M	0.67s	3.3x Slower
XLM-R-Large	∼560M	0.68s	3.4x Slower
mE5-Large	∼560M	0.31s	1.5x Slower
Qwen3-0.6B	∼0.6B	0.64s	3.2x Slower
Jasper	∼1.9B	1.41s	7.0x Slower
Qwen3-4B	∼4B	5.37s	26.8x Slower

Table 2: Inference Latency Comparison (Batch Size 32). Metrics were recorded on a single NVIDIA T4 GPU to ensure standardized environmental constraints and reproducible baselines.

As shown in Table 2, while BERT-family models (Devlin et al., 2019; Liu et al., 2019) offer peak throughput, Jasper (Zhang et al., 2024) establishes the optimal Pareto frontier. It delivers state-of-the-art accuracy with manageable latency (1.41s/batch), whereas Qwen3-4B (Zhang et al., 2025) incurs a massive computational cost ($4\times$ slower), restricting its use to offline processing.

5.4 Qualitative Analysis of Reasoning

Qualitative analysis (see Appendix B for full reasoning traces) reveals capabilities critical for clinical trust that quantitative metrics miss. The Clinical Agent (Paradigm 3) (OpenAI et al., 2025; Wei et al., 2022; Brown et al., 2020) demonstrates distinct orthographic and temporal awareness, correctly interpreting non-standard emphasis (e.g., "goooooood") within a longitudinal recovery trajectory rather than as isolated sentiment. Furthermore, it handles semantic buffering by recognizing spiritual terms as protective factors against negative valence, and exhibits robust cold-start inference by deducing complex mixed states, such as grief combined with duty-driven arousal, without any historical baseline.

6 Conclusion

We evaluated a tripartite affective framework for longitudinal affect assessment spanning three parallel paradigms: Zero-Context Static Regression, Latent Temporal Contextualization via LSTMs, and Explicit Semantic Reasoning. Empirically, Matryoshka-distilled embeddings (Zhang et al., 2024; Kusupati et al., 2024) paired with XGBoost (Chen and Guestrin, 2016) established a superior density-efficiency frontier, outperforming larger models (Zhang et al., 2025). Surprisingly, robust static extraction consistently dominated explicit sequence modeling via LSTMs, which struggled with sparse, irregular intervals. Furthermore, while the Clinical Agent trailed vector-based regressors in tracking high-frequency fluctuations, it achieved state-of-the-art psychological profiling, yielding the highest Between-User Valence performance and confirming the value of explicit deductive reasoning for trait assessment. Finally, a persistent "arousal bottleneck" confirms the limitations of text-only modeling for physiological activation (Atmaja and Akagi, 2021; Schuller, 2018). Future work will address these constraints by replacing LSTMs with Transformer-based architectures to better capture long-range dependencies, integrating multimodal signals, and exploring domain-adapted psychological LLMs to enhance nuanced affective reasoning.

References

Bagus Tris Atmaja and Masato Akagi. 2021. Two-stage dimensional emotion recognition by fusing predictions of acoustic and text networks using svm. *Speech Communication*, 126:9–21.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. In *NeurIPS*.

Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794. Association for Computing Machinery.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451. 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 (NAACL-HLT)*, pages 4171–4186.

Harris Drucker, Chris J.C. Burges, Linda Kaufman, Alex Smola, and Vladimir Vapnik. 1997. Support vector regression machines. *Advances in Neural Information Processing Systems*, 9.

Panagiotis Giannos and Orestis Delardas. 2024. Applications of large language models in psychiatry: A systematic review. *Frontiers in Psychiatry*, 15:1422807.

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

Yoon Kim. 2014. Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882*.

Peter Kuppens, Nicholas B Allen, and Lisa B Sheeber. 2010. Emotional inertia and psychological maladjustment. *Psychological science*, 21(7):984–991.

Peter Kuppens, Zita Oravecz, and Francis Tuerlinckx. 2012. The nature of emotional inertia: A matter of temporal structure rather than intensity or variability. *Psychological Review*, 119(3):1161–1174.

Aditya Kusupati, Gantavya Bhatt, Aniket Rber, Matthew Wallingford, Aditya Sinha, Vivek Prasanna, Ali Farhadi, Sham Kakade, and Prateek Jain. 2024. Matryoshka representation learning. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Bing Liu. 2012. *Sentiment analysis and opinion mining*. Morgan and Claypool Publishers.

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*.

Navonil Majumder, Soujanya Poria, Devamanyu Hazarika, Rada Mihalcea, Alexander Gelbukh, and Erik Cambria. 2019. Dialoguernn: An attentive rnn for emotion detection in conversations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6818–6825.

OpenAI, Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K. Arora, Yu Bai, Bowen Baker, and 1 others. 2025. GPT-OSS-120B & GPT-OSS-20B model card. *arXiv preprint arXiv:2508.10925*.

Bo Pang, Lillian Lee, and 1 others. 2008. Opinion mining and sentiment analysis. *Foundations and Trends® in Information Retrieval*, 2(1–2):1–135.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, and 2 others. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12:2825–2830.

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

Björn W Schuller. 2018. Speech emotion recognition: Two decades in a nutshell, benchmarks, and ongoing trends. *Communications of the ACM*, 61(5):90–99.

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.

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*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pages 24824–24837.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 others. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45. Association for Computational Linguistics.

Xuhai Xu, Bingshen Yao, Yuanzhe Dong, Saadia Gabriel, Hong Yu, James Hendler, Marzyeh Ghassemi, Anind K. Dey, and Dakuo Wang. 2024. Mental-LLM: Leveraging large language models for mental health prediction via online text data. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 8(1):1–32.

Kailai Yang, Shaoxiong Ji, Tianlin Zhang, Qianqian Xie, and Sophia Ananiadou. 2023. Mentallama: Interpretable mental health analysis on social media with large language models. *arXiv preprint arXiv:2309.13567*.

Dun Zhang, Jiacheng Li, Ziyang Zeng, and Fulong Wang. 2024. Jasper and stella: Distillation of sota embedding models. *arXiv preprint arXiv:2412.19048*.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. 2025. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*.

A Hyperparameter Specifications

A.1 Search Spaces

We conducted grid searches to optimize the hyperparameters for all Paradigm 1 (Static) and Paradigm 2 (Longitudinal) models. The search spaces were defined as follows:

1. Ridge Regression:

- **Bidirectional Encoder Embeddings:**
  - *BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), XLM-RoBERTa (Conneau et al., 2020):* **Alpha:** 0.0, 1e-05, 0.0001 to 0.005, 0.01 to 0.09 (step 0.01), 0.1 to 0.5, 0.7 to 1.0.
  - *mE5-Large (Wang et al., 2024):* **Alpha:** Standard range (0.0–1.0) plus extended range: 1.1 to 2.0 (step 0.1).
- **Distilled Multimodal Embeddings (Jasper (Zhang et al., 2024)):**
  - **Alpha:** Standard range (0.0–1.0) plus extended range: 2.0 to 20.0 (step 1.0), 25.0, 50.0.
- **Generative LLM Embeddings (Qwen3 (Zhang et al., 2025)):**
  - **Alpha:** Standard range (0.0–1.0) plus extended range: 1.1 to 2.0 (step 0.1).

2. XGBoost (Chen and Guestrin, 2016):

- **Learning Rate:** 0.01, 0.05
- **Max Depth:** 3, 4, 6
- **Min Child Weight:** 4, 6, 8
- **Subsample:** Fixed at 0.8

3. Support Vector Regression (SVR) (Drucker et al., 1997):

- **Bidirectional Encoder Embeddings:**
  - *BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), XLM-RoBERTa (Conneau et al., 2020):* **Kernel:** Linear, RBF; **C (Regularization):** 0.01, 0.1, 1.0, 5.0, 10.0; **Epsilon (Margin):** 0.01, 0.1, 0.2, 0.3.
  - *mE5-Large (Wang et al., 2024):* **Kernel:** Linear, RBF; **C:** 0.1, 1.0, 2.0; **Epsilon:** 0.1 to 0.7 (step 0.1).
- **Distilled Multimodal Embeddings (Jasper (Zhang et al., 2024)):**
  - **Kernel:** Linear, RBF
  - **C (Extended):** Added 2.0, 2.5, 3.0, 4.0, 15.0, 20.0
  - **Epsilon (Extended):** Added 0.4 to 0.9 (step 0.1)
- **Generative LLM Embeddings (Qwen3 (Zhang et al., 2025)):**
  - **Kernel:** Linear, RBF
  - **C:** 0.1, 1.0, 2.0
  - **Epsilon:** 0.1 to 0.7 (step 0.1)

4. Long Short-Term Memory (LSTM) (Hochreiter and Schmidhuber, 1997):

- **Hidden Dimension:** 64, 128, 256
- **Layers:** 1, 2
- **MLP Units:** 32, 64, 128
- **Epochs:** 20, 30
- **Batch Size:** 16, 32
- **Dropout:** 0.1, 0.3
- **Weight Decay:** 0.01, 0.001

A.2 Final Selected Hyperparameters

Tables 3 through 6 detail the optimal hyperparameters identified for each backbone and method combination.

Backbone	Alpha (α)
BERT-Base	0.3
RoBERTa-Base	0.05
XLM-RoBERTa	0.0005
RoBERTa-Large	0.01
XLM-RoBERTa-Large	0.0005
mE5-Large	2.0
Jasper	2.0
Qwen3-0.6B	0.9
Qwen3-4B	1.8

Table 3: Ridge Regression Hyperparameters

Backbone	LR	Depth	MinChild	Sub
BERT-Base	0.05	6	8	0.8
RoBERTa-Base	0.05	3	6	0.8
XLM-RoBERTa	0.05	6	8	0.8
RoBERTa-Large	0.05	3	6	0.8
XLM-RoBERTa-Large	0.05	4	6	0.8
mE5-Large	0.05	6	8	0.8
Jasper	0.05	6	4	0.8
Qwen3-0.6B	0.05	4	4	0.8
Qwen3-4B	0.05	4	8	0.8

Table 4: XGBoost Hyperparameters

Backbone	Kernel	C	Eps
BERT-Base	RBF	1.0	0.01
RoBERTa-Base	RBF	10.0	0.01
XLM-RoBERTa	RBF	10.0	0.01
RoBERTa-Large	RBF	10.0	0.3
XLM-RoBERTa-Large	RBF	10.0	0.01
mE5-Large	RBF	1.0	0.8
Jasper	Linear	0.7	0.7
Qwen3-0.6B	Linear	1.0	0.7
Qwen3-4B	Linear	1.0	0.6

Table 5: SVR Hyperparameters

Backbone	Hid	Lyr	MLP	Drop	WD	Ep
mE5-Large	256	2	128	0.3	0.01	30
Jasper	256	1	64	0.1	0.01	30
Qwen3-0.6B	256	2	64	0.3	0.001	30
Qwen3-4B	256	2	32	0.1	0.01	30

Table 6: LSTM Architecture & Optimization

B Agentic Framework Configuration

This appendix details the configuration of the Clinical Agent. We utilize a specialized System Persona for domain alignment and a dynamic Prompt Template incorporating Chain-of-Thought (CoT) reasoning (Wei et al., 2022) to ensure interpretability in longitudinal affect assessment.

B.1 System Persona

The system prompt instantiates the agent as a domain expert, grounding the analysis in Russell's Circumplex Model of Affect (Russell, 1980). It defines the affective ontology (Valence/Arousal) and instructs the model to handle multimodal inputs—differentiating between explicit diagnostic descriptors (`feeling_words`) and implicit narrative inference (`essay`).

```
You are an expert Affective Computing Scientist
    and Clinical Psychologist specialized in
    Longitudinal Affect Assessment. Your goal is
     to predict the emotional state of a subject
     based on their text entries.

<theoretical_framework>
You adhere to Russell's (1980) Circumplex Model
    of Affect:
1. Valence (V): Measures pleasantness. Range:
    -2.0 to 2.0.
2. Arousal (A): Measures activation/energy.
    Range: 0.0 to 2.0.

Your analysis must account for:
* Baseline: The user's "default" emotional state
     (if history is available).
* Dynamics: How the user reacts to events and
    the time elapsed.
* Modality:
    * If type="feeling_words": Explicit
    diagnostic emotions.
    * If type="essay": Narrating events. Infer
    latent emotion.
</theoretical_framework>
```

B.2 Prompt Template & Task Definition

The prompt template enforces a strict few-shot CoT reasoning (Wei et al., 2022) structure. It directs the model to explicitly model the user's emotional baseline and trajectory before outputting a prediction. To ensure computational compatibility, the output is constrained to a strict JSON schema.

```
<task_definition>
You will be provided with a `Target` entry and,
    if available, a `History` of past entries.

Let's think step-by-step:
1. Analyze Context: If `History` is provided,
    use it to model the user's emotional
    baseline and recent trajectory.
2. Evaluate Dynamics: Consider the `Time_Delta`.
3. Predict: Estimate the Valence and Arousal.
</task_definition>

<output_format>
Provide your response in JSON format.
Constraint: Predict `valence` and `arousal` as
    floats with 1 decimal place.
{
  "reasoning": {
    "baseline_assessment": "...",
    "recent_trajectory": "...",
    "target_analysis": "...",
    "time_factor": "..."
  },
  "prediction": {
    "valence": <float_1_decimals>,
```

```
    "arousal": <float_1_decimals>
  }
}
</output_format>

<few_shot_examples>
{{INSERT_REPRESENTATIVE_EXAMPLES_FROM_TRAINING_
SET}}
</few_shot_examples>

<input_context>
User History: {{INSERT_DYNAMIC_USER_HISTORY}}

Target Entry:
- Timestamp: {{TARGET_TIMESTAMP}}
- Time Since Last Entry: {{
    PRE_CALCULATED_TIME_DELTA}}
- Type: {{IS_WORDS_BOOLEAN}}
- Text: "{{TARGET_TEXT}}"
</input_context>
```

Structured Clinical Reasoning Constraints The `<output_format>` schema enforces a strict 4-step diagnostic progression:

1. **Baseline Assessment:** Analyzing the user's general emotional disposition based on their history.
2. **Recent Trajectory:** Identifying directional trends (e.g., recovery vs. deterioration).
3. **Target Analysis:** Examining the specific linguistic and semantic cues of the current text.
4. **Time Factor:** Weighing the impact of the elapsed time on the continuity of the emotional state.

Dynamic Context Variables The `<input_context>` and `<few_shot_examples>` blocks are populated dynamically for each query to construct the context window:

- **User History (`INSERT_DYNAMIC_USER_HISTORY`):** A chronological sequence of the user's recent entries, including previously assessed affective states.
- **Temporal Dynamics (`PRE_CALCULATED_TIME_DELTA`):** The explicit time difference between entries (e.g., "4 hours later" vs. "4 days later"), enabling the model to distinguish between transient moods and persistent affective episodes.
- **Adaptive Anchoring (`INSERT_REPRESENTATIVE_EXAMPLES...`):** Teacher-generated few-shot examples (see Section B.3) dynamically injected into the prompt based on whether the user is "seen" or "unseen".
- **Target Entry Variables (`TARGET_TIMESTAMP`, `IS_WORDS_BOOLEAN`, `TARGET_TEXT`):** The raw inputs representing the immediate entry to be predicted, including its precise temporal marker, explicit modality flag, and textual content.

B.3 Teacher-Guided Few-Shot Anchors

To ground the agent's structured reasoning, we inject a small set of worked examples into the prompt as few-shot anchors (Brown et al., 2020). These high-fidelity rationales were generated using **Gemini-3.0-Pro** as a teacher model. Crucially, the teacher was constrained to the exact same 4-step diagnostic schema defined in Section B.2, explicitly guiding the open-weights model (GPT-OSS-120B (OpenAI et al., 2025)) via in-context learning on how to link linguistic cues and temporal gaps to emotional trajectories.

For **Seen** Users, we provide examples drawn from a representative user's own timeline (`FEW_SHOT_SEEN`, 3 examples), demonstrating how to leverage history and temporal dynamics. For **Unseen** Users, we provide population-level exemplars (`FEW_SHOT_UNSEEN`, 3 examples) that teach the model to reason from text alone, without historical context. Below, we present one representative anchor from each set.

FEW_SHOT_SEEN (Representative Example): This anchor demonstrates a multi-step recovery trajectory. The model is taught to (1) track a user's emotional arc across sequential entries, (2) recognize stressor resolution via lexical shift, and (3) use a long time gap (overnight) to justify a departure from a prior negative trend.

```
{
  "type": "essay",
  "history": [
    {
      "text": "Disappointed , Worried ,
    Apprehensive , Concerned",
      "timestamp": "2024-11-06 12:56:50",
      "type": "feeling_words",
      "label": {"valence": -1.0, "arousal": 1.0}
    },
    {
      "text": "Uncertain , Motivated , Tentative
     , Hopeful , Worried",
      "timestamp": "2024-11-06 17:01:38",
      "type": "feeling_words",
      "label": {"valence": 0.0, "arousal": 1.0}
    }
  ],
  "target": {
```

```
    "text": "Feeling calm and content . Feeling
    hopeful for the future and proud of my baby
    daughter and her growth .",
    "timestamp": "2024-11-07 12:57:55",
    "time_delta": "19 hours later",
    "label": {"valence": 2.0, "arousal": 0.0}
  },
  "reasoning": {
    "baseline_assessment": "User 137 was
    struggling yesterday (negative/mixed).",
    "recent_trajectory": "Strong upward recovery
     overnight.",
    "target_analysis": "The user has fully reset.
     The negative words from yesterday are gone.
     Strong positive markers ('proud', 'hopeful
    ', 'content') dominate. The shift from '
    Worried' (High Arousal) to 'Calm' (Low
    Arousal) indicates a resolution of the
    stressor.",
    "time_factor": "The overnight gap (19 hours)
     allowed for a sleep cycle and a 'fresh
    start', justifying the break from yesterday'
    s negative trend."
  }
}
```


FEW_SHOT_UNSEEN (Representative Example): This anchor teaches mixed-affect resolution without access to user history. The model learns to weigh conflicting semantic signals—positive mental states against negative physical states—and to let the dominant signal class determine the final prediction.


```
{
  "type": "feeling_words",
  "text": "Tired, Sick, Calm, Content, Unwell",
  "label": {"valence": -1.0, "arousal": 0.0},
  "reasoning": {
    "baseline_assessment": "N/A (Few-Shot
    Example)",
    "recent_trajectory": "N/A (Few-Shot Example)
    ",
    "target_analysis": "This entry lists
    conflicting emotions. 'Calm' and 'Content'
    are positive, but 'Sick', 'Unwell', and '
    Tired' are strong negative physical markers.
     The negative physical state outweighs the
    mental calm, resulting in a moderately
    negative Valence (-1.0). The overall energy
    is clearly low (0.0).",
    "time_factor": "N/A (Few-Shot Example)"
  }
}
```


B.4 Inference Configuration Search

Unlike the mathematical regressors optimized via gradient descent, the performance of the Clinical Agent is governed by inference-time generation parameters. To identify the optimal configuration for longitudinal deduction, we conducted a targeted parameter sweep over two primary axes: **Context Window Size** ($w \in \{10, 20, 30\}$), determining the maximum number of historical entries injected into the prompt, and **Generation Temperature** ($\tau \in \{0.7, 0.85, 1.0\}$), controlling the diversity and creativity of the CoT reasoning (Wei et al., 2022).

Due to the significant computational and financial costs associated with API-based LLM inference, this configuration search was executed on a highly restricted subset of the training dataset, with 30 total entries. Crucially, to evaluate the agent's longitudinal reasoning accurately, entries were randomly sampled by specific user IDs and strictly ordered by timestamp. This targeted sampling strategy ensured that the chronological continuity of the historical context remained unbroken, establishing a realistic and methodologically sound directional baseline before the full-scale test set evaluation. While this limited 30-entry subset was used strictly to find the optimal inference configuration, the final test set evaluation reported in Section 5 was executed on the complete 1,737-instance dataset.

		Valence (r)			Arousal (r)		
Window	**Temp**	W	B	C	W	B	C
10	0.85	0.048	0.836	0.557	-0.120	0.500	0.211
10	1.00	0.035	0.821	0.535	-0.062	0.479	0.226
20	0.70	0.007	**0.875**	0.593	-0.004	0.381	0.196
20	0.85	**0.079**	0.824	0.555	**0.174**	**0.667**	**0.455**
20	1.00	0.016	0.841	0.552	-0.039	0.580	0.302
30	0.85	0.034	0.874	**0.599**	-0.076	0.579	0.284

Table 7: Agentic inference configuration search on a constrained 30-entry training subset. W, B, and C refer to *Within-User*, *Between-User*, and *Composite* Pearson correlations.

Configuration Analysis: As detailed in Table 7, while higher context windows (e.g., $w = 30$) yielded the highest Valence Composite score ($r_{composite} = 0.599$), they triggered a catastrophic collapse in the model's ability to track Arousal fluctuations ($r_{within} = -0.076$). We hypothesize that overly large context windows introduce "temporal noise," causing the agent to lose focus on the immediate target entry's linguistic cues. Conversely, overly low temperatures ($\tau = 0.7$) resulted in rigid reasoning traces that failed to adapt to nuanced state changes.

The configuration of *Window Size = 20* and *Temperature = 0.85* was selected for the final evaluation because it provided the most robust equilibrium. While maintaining strong Valence profiling, it was the only configuration capable of successfully generating positive, stabilized tracking for the highly resistant Arousal dimension ($r_{composite} = 0.455$).

B.5 Qualitative Case Studies

The following examples demonstrate the agent's reasoning capabilities in complex scenarios, highlighting its ability to integrate longitudinal history with immediate linguistic cues. *Note: The JSON blocks below are composite representations combining the dynamic inputs and the model's generated outputs for readability. They have been slightly truncated for brevity (e.g., shortening long history summaries) while preserving the complete logical structure of the agent's reasoning.*

Case Study 1: Informal Orthography & Recovery

- **Context:** 5 hours after reporting "fatigue".
- **Input Text:** "Life is gooooooooooooooooooooood"
- **Agent Output:**


```
{
  "history_summary": "User moved from low
    arousal/fatigue (Valence 0.2) to high energy
     over 5 hours.",
  "target": {
    "time_delta": "5 hours later",
    "text": "Done with work... Life is
    gooooooooo[...]ooood"
  },
  "model_output": {
    "reasoning": {
      "baseline_assessment": "Across the three
    recent entries the user shows a generally
    positive baseline (valence 1.0) that dipped
    ... and then recovered.",
      "recent_trajectory": "The trajectory over
    the last few hours is upward, moving from a
    mildly negative mood to a small positive
    shift...",
      "target_analysis": "The target text
    reports finishing work... and an emphatic '
    Life is gooooooooood'. This signals high
    pleasantness and a mild boost in energy/
    excitement...",
      "time_factor": "Only 5 hours have passed
    since the last entry, so the recent upward
    trend is likely to continue..."
    },
    "prediction": { "valence": 1.8, "arousal":
    0.4 }
  }
}
```


Case Study 2: Semantic Buffering

- **Context:** 14 hours after a High Positive state.
- **Input Text:** "Sad, Worried, Sacred, Prayers, Anxious"
- **Agent Output:**


```
{
  "history_summary": "User dropped from High
    Positive (2.0) to Deep Negative (-2.0) over
    24 hours.",
  "target": {
    "time_delta": "14 hours later",
    "text": "Sad , Worried , Sacred , Prayers ,
    Anxious"
  },
  "model_output": {
    "reasoning": {
      "baseline_assessment": "The user's
    baseline shifted from a highly positive
    state... to a strongly negative state by the
     last entry.",
      "recent_trajectory": "Valence has been
    trending sharply downward over the past 24
    hours...",
      "target_analysis": "The current feeling-
    words list contains three negative emotions
    (Sad, Worried, Anxious) and two spiritually-
    oriented terms (Sacred, Prayers). The
    negatives dominate valence, but the
    spiritual words provide slight mitigation...
     All negative terms are activating...",
      "time_factor": "The 14-hour interval is
    short enough that the prior negative episode
     likely persists, but the elapsed time
    allows for some reflection..."
    },
    "prediction": { "valence": -1.5, "arousal":
    1.0 }
  }
}
```


Case Study 3: Longitudinal Decay

- **Context:** 3 days after reporting "High Energy".
- **Input Text:** "Tired, Sleepy, Calm, Relaxed, Careless"
- **Agent Output:**


```
{
  "history_summary": "Consistently high Valence
    (2.0) with varying arousal.",
  "target": {
    "time_delta": "3 days later",
    "text": "Tired , Sleepy , Calm , Relaxed ,
    Careless"
  },
  "model_output": {
    "reasoning": {
      "baseline_assessment": "The user has
    consistently reported a maximally positive
    valence (2.0) across all three prior entries
    ...",
      "recent_trajectory": "While valence
    remained high, arousal varied... The most
    recent entry showed high arousal.",
      "target_analysis": "The listed words point
     to low arousal overall. Calm and Relaxed
    contribute slight positive pleasantness,
    whereas Tired, Sleepy and Careless pull the
    valence toward neutral. The net impression
    is a modestly positive but near-neutral
    valence.",
```

```
      "time_factor": "A three-day interval
    provides enough time for the user's mood to
    drift away from the previous high-energy
    state..."
    },
    "prediction": { "valence": 0.2, "arousal":
    0.3 }
  }
}
```


Case Study 4: Cold-Start Inference (Unseen User)

- **Context:** No history available (Empty History).
- **Input Text:** "...my friends dad died so i have been... helping him book travel..."
- **Agent Output:**


```
{
  "history_summary": "N/A (Empty History)",
  "target": {
    "time_delta": "N/A",
    "text": "over the weekend , my friends dad
    and dog both died... feeling sorry for him
    ..."
  },
  "model_output": {
    "reasoning": {
      "baseline_assessment": "N/A (first entry)",

      "recent_trajectory": "N/A (first entry)",
      "target_analysis": "The user describes the
     death of a friend's father and dog... This
    conveys a predominantly sad and grieving
    affect, indicating low pleasantness. The
    language is calm... suggesting low arousal,
    though the caregiving action adds a slight
    activation.",
      "time_factor": "N/A (first entry)"
    },
    "prediction": { "valence": -1.5, "arousal":
    0.5 }
  }
}
```


C Complete Experimental Results

This appendix presents the exhaustive performance metrics for all backbone architectures, regression heads, and the agentic framework evaluated across our tripartite study, including those omitted from the main text for brevity; full results are detailed in Table 8.

D Ablation Study: Restricted Data Volume Results

As analyzed in Section 5, the choice of regression head is highly sensitive to the available training data volume. Table 9 presents the exhaustive performance metrics for all models evaluated during our preliminary experimental phase on restricted, stratified partitions of the training dataset. These results explicitly demonstrate the severe overfitting of tree-based ensembles (XGBoost) in low-data conditions and the superior noise-filtering capabilities of margin-based models (SVR)—the robust architecture utilized for our official SemEval submission. As detailed in the main text, this performance hierarchy completely inverts when models are scaled to the full training corpus.

Backbone	Method	Metric	Val (W)	Val (B)	Val (C)	Aro (W)	Aro (B)	Aro (C)
BERT-Base	Ridge	r	0.504	0.651	0.583	0.312	0.531	0.428
		MAE	0.829	0.479	0.693	0.550	0.264	0.418
	XGBoost	r	0.493	0.675	0.592	0.345	0.540	0.447
		MAE	0.823	0.451	0.678	0.539	0.269	0.413
	SVR	r	0.508	0.667	0.593	0.355	0.533	0.448
		MAE	0.827	0.469	0.688	0.543	0.275	0.418
RoBERTa-B	Ridge	r	0.513	0.644	0.582	0.331	0.507	0.423
		MAE	0.839	0.478	0.701	0.555	0.273	0.424
	XGBoost	r	0.490	0.651	0.576	0.341	0.495	0.421
		MAE	0.843	0.474	0.703	0.556	0.283	0.429
	SVR	r	0.497	0.663	0.586	0.313	0.508	0.415
		MAE	0.840	0.468	0.698	0.556	0.277	0.426
XLM-RoBERTa	Ridge	r	0.502	0.639	0.575	0.283	0.497	0.395
		MAE	0.848	0.473	0.707	0.565	0.260	0.424
	XGBoost	r	0.499	0.694	0.606	0.275	0.445	0.363
		MAE	0.819	0.461	0.679	0.563	0.280	0.432
	SVR	r	0.484	0.656	0.577	0.295	0.384	0.340
		MAE	0.861	0.488	0.724	0.563	0.298	0.440
RoBERTa-L	Ridge	r	0.517	0.619	0.570	0.294	0.538	0.424
		MAE	0.833	0.489	0.700	0.561	0.273	0.427
	XGBoost	r	0.491	0.633	0.567	0.306	0.443	0.376
		MAE	0.851	0.486	0.714	0.567	0.290	0.439
	SVR	r	0.512	0.614	0.565	0.294	0.621	0.474
		MAE	0.844	0.502	0.713	0.550	0.272	0.421
XLM-R-Large	Ridge	r	0.492	0.711	0.613	0.309	0.396	0.353
		MAE	0.825	0.456	0.682	0.564	0.273	0.429
	XGBoost	r	0.486	0.695	0.601	0.236	0.498	0.375
		MAE	0.815	0.463	0.675	0.559	0.275	0.428
	SVR	r	0.491	0.627	0.563	0.268	0.404	0.338
		MAE	0.875	0.504	0.741	0.557	0.296	0.435
Jasper	Ridge	r	0.562	0.684	0.627	0.419	**0.629**	**0.532**
		MAE	0.793	0.447	0.653	**0.507**	**0.233**	**0.378**
	XGBoost	r	**0.579**	0.718	**0.654**	0.421	0.562	0.495
		MAE	**0.788**	0.439	**0.646**	0.518	0.255	0.395
	SVR	r	0.555	0.682	0.623	0.392	0.590	0.497
		MAE	0.798	0.445	0.656	0.533	0.254	0.403
	LSTM	r	0.527	0.702	0.622	**0.423**	0.547	0.488
		MAE	0.820	0.426	0.667	0.516	0.263	0.397
Qwen3-0.6B	Ridge	r	0.523	0.673	0.603	0.409	0.581	0.500
		MAE	0.809	0.445	0.665	0.522	0.243	0.391
	XGBoost	r	0.563	0.704	0.638	0.386	0.518	0.455
		MAE	0.812	0.453	0.670	0.529	0.266	0.406
	SVR	r	0.532	0.679	0.611	0.390	0.531	0.463
		MAE	0.816	0.453	0.673	0.544	0.258	0.411
	LSTM	r	0.523	0.709	0.625	0.392	0.568	0.485
		MAE	0.803	**0.424**	0.653	0.540	0.266	0.412
Qwen3-4B	Ridge	r	0.550	0.685	0.622	0.418	0.590	0.509
		MAE	0.802	0.435	0.656	0.519	0.243	0.389
	XGBoost	r	0.528	0.691	0.616	0.390	0.516	0.456
		MAE	0.813	0.436	0.665	0.526	0.257	0.400
	SVR	r	0.551	0.687	0.623	0.405	0.529	0.470
		MAE	0.811	0.438	0.664	0.541	0.257	0.409
	LSTM	r	0.448	0.662	0.564	0.391	0.533	0.465
		MAE	0.873	0.451	0.724	0.529	0.253	0.400
mE5-Large	Ridge	r	0.528	0.688	0.614	0.385	0.569	0.482
		MAE	0.812	0.452	0.670	0.540	0.262	0.411
	XGBoost	r	0.518	0.681	0.606	0.371	0.537	0.458
		MAE	0.811	0.442	0.666	0.542	0.270	0.415
	SVR	r	0.525	0.708	0.625	0.362	0.542	0.456
		MAE	0.801	0.437	0.655	0.560	0.283	0.432
	LSTM	r	0.523	0.694	0.615	0.369	0.401	0.385
		MAE	0.866	0.521	0.738	0.566	0.298	0.442
GPT-OSS-120B	TG-CRA	r	0.529	**0.725**	0.637	0.362	0.565	0.470
		MAE	0.807	0.494	0.680	0.558	0.271	0.425

Table 8: Comprehensive performance metrics for all Backbone and Method combinations.

Backbone	Method	Metric	Val (W)	Val (B)	Val (C)	Aro (W)	Aro (B)	Aro (C)
BERT-Base	Ridge	r	0.411	0.650	0.541	0.366	0.456	0.412
		MAE	0.952	0.889	0.927	0.551	0.488	0.520
	XGBoost	r	0.371	0.653	0.526	0.557	0.404	0.484
		MAE	0.968	0.900	0.943	0.571	0.507	0.540
	SVR	r	0.322	0.660	0.510	0.306	0.477	0.395
		MAE	0.942	0.880	0.916	**0.529**	**0.469**	**0.499**
RoBERTa-B	Ridge	r	0.445	0.669	0.567	**0.617**	0.484	**0.554**
		MAE	0.952	0.879	0.923	0.546	0.492	0.520
	XGBoost	r	0.504	0.596	0.552	0.401	0.286	0.345
		MAE	1.034	0.976	NaN	0.601	0.533	0.568
	SVR	r	0.445	0.659	0.561	0.600	0.436	0.523
		MAE	0.932	0.864	0.903	0.564	0.518	0.542
XLM-RoBERTa	Ridge	r	0.355	0.596	0.485	0.277	0.394	0.337
		MAE	1.060	0.986	NaN	0.591	0.523	0.558
	XGBoost	r	0.346	0.626	0.499	0.369	0.306	0.338
		MAE	1.000	0.933	0.998	0.586	0.519	0.554
	SVR	r	0.269	0.574	0.434	0.273	0.319	0.296
		MAE	1.080	1.004	NaN	0.585	0.522	0.554
RoBERTa-L	Ridge	r	0.483	0.619	0.555	0.305	0.494	0.404
		MAE	0.982	0.917	0.961	0.552	0.489	0.521
	XGBoost	r	0.333	0.577	0.463	0.164	0.358	0.264
		MAE	1.028	0.963	NaN	0.587	0.518	0.553
	SVR	r	0.495	0.614	0.557	0.398	0.462	0.431
		MAE	0.994	0.927	0.979	0.562	0.502	0.532
XLM-R-Large	Ridge	r	0.561	0.606	0.584	0.445	**0.518**	0.482
		MAE	0.942	0.884	0.918	0.534	0.474	0.504
	XGBoost	r	0.482	0.602	0.545	0.376	0.397	0.386
		MAE	1.008	0.949	NaN	0.574	0.502	0.539
	SVR	r	0.364	0.564	0.470	0.363	0.355	0.359
		MAE	1.072	1.006	NaN	0.574	0.511	0.544
Jasper	Ridge	r	0.575	0.629	0.602	0.289	0.496	0.398
		MAE	0.956	0.904	0.935	0.554	0.486	0.520
	XGBoost	r	0.395	**0.708**	0.572	0.495	0.447	0.471
		MAE	**0.914**	0.862	0.891	0.547	0.475	0.512
	SVR	r	**0.681**	0.625	**0.654**	0.441	0.485	0.463
		MAE	0.955	0.908	0.936	0.567	0.496	0.533
	LSTM	r	0.495	0.684	0.598	0.441	0.463	0.452
		MAE	0.967	0.908	0.944	0.566	0.502	0.535
Qwen3-0.6B	Ridge	r	0.541	0.639	0.592	0.483	0.472	0.478
		MAE	0.974	0.918	0.954	0.565	0.505	0.536
	XGBoost	r	0.499	0.661	0.586	0.489	0.337	0.416
		MAE	0.949	0.889	0.925	0.581	0.516	0.549
	SVR	r	0.530	0.621	0.577	0.292	0.448	0.372
		MAE	1.008	0.949	NaN	0.562	0.498	0.531
	LSTM	r	0.537	0.666	0.605	0.367	0.509	0.441
		MAE	0.965	0.905	0.942	0.566	0.500	0.534
Qwen3-4B	Ridge	r	0.539	0.632	0.587	0.399	0.507	0.455
		MAE	0.946	0.891	0.923	0.562	0.494	0.529
	XGBoost	r	0.451	0.654	0.561	0.478	0.495	0.486
		MAE	0.944	0.889	0.921	0.559	0.490	0.525
	SVR	r	0.512	0.604	0.560	0.393	0.477	0.436
		MAE	0.991	0.931	0.975	0.560	0.493	0.527
	LSTM	r	0.549	0.676	0.616	0.410	0.502	0.457
		MAE	0.915	**0.853**	**0.888**	0.572	0.501	0.537
mE5-Large	Ridge	r	0.491	0.640	0.570	0.407	0.413	0.410
		MAE	0.960	0.901	0.937	0.594	0.523	0.559
	XGBoost	r	0.506	0.654	0.585	0.506	0.386	0.448
		MAE	0.951	0.876	0.922	0.589	0.523	0.557
	SVR	r	0.451	0.609	0.534	0.421	0.344	0.383
		MAE	1.018	0.955	NaN	0.598	0.529	0.565
	LSTM	r	0.503	0.654	0.584	0.411	0.401	0.406
		MAE	0.952	0.893	0.928	0.608	0.543	0.576

Table 9: Performance metrics evaluated on a restricted (chunked) subset of the training data, demonstrating the underperformance of XGBoost under low-data conditions.

Note on Missing Values (NaN): The Composite Correlation metric depends on the Fisher z-transformation ($arctanh$), which is undefined for input values ≥ 1.0. Several models (e.g., XLM-RoBERTa with SVR/XGBoost, Qwen3-0.6B SVR) produced Within-User MAE scores exceeding 1.0 under these restricted training conditions, resulting in mathematically undefined Composite MAE scores. These are denoted as `NaN` in the table.