The Affective Regulation Layer: A Biologically Plausible Architecture for Real-Time Emotional Dynamics in Artificial Intelligence

Furkan Nar
Independent Researcher
furkannar168@hotmail.com

January 2026

Abstract

We introduce the Affective Regulation Layer (ARL), an open-source, biologically plausible affective architecture designed to implement real-time emotional dynamics with neuroscientific fidelity. Traditional sentiment analysis in Large Language Models (LLMs) is limited to discrete, stateless classification, failing to capture the nuanced and persistent nature of human emotion. To address this, the ARL features a dual-process scoring engine that integrates symbolic keyword detection with sub-symbolic semantic centroids derived from transformer-based embeddings.

Central to the system is a homeostatic decay mechanism modeled after neurochemical half-lives, ensuring that emotional states persist and dissipate realistically over time rather than resetting between inputs. Furthermore, we implement biological cross-modulation logic based on Panksepp's affective neuroscience, allowing primary states like Joy to prime secondary drives such as Curiosity. This architecture provides a theoretical foundation for human-like emotional memory in AI systems, enabling more empathetic and context-aware machine intelligence. We present the complete architectural design, mathematical formalization, and open-source implementation.

Keywords: Affective Computing, Neurobiological AI, Homeostasis, Semantic Centroids, Panksepp, Emotional Architecture

Code: https://github.com/TheOfficialFurkanNar/Affective-regulation.layer

---

1 Introduction

Modern artificial intelligence systems, particularly Large Language Models (LLMs), have achieved remarkable capabilities in natural language understanding and generation. However, these systems fundamentally lack persistent emotional states, processing each interaction in isolation without the continuity that characterizes human affective experience. While sentiment analysis techniques can classify text as positive, negative, or neutral, they fail to capture the temporal dynamics, cross-modulation, and homeostatic regulation that define biological emotional systems.

Human emotions are not discrete labels but continuous states that evolve over time, influenced by both immediate stimuli and lingering affective contexts. Neuroscientific research, particularly

the work of Panksepp [1], has identified core affective systems—such as SEEKING, PLAY, FEAR, and RAGE—that interact in complex ways to produce emotional experience. These systems exhibit properties such as:

- **Persistence**: Emotions linger after their triggering stimulus
- **Decay**: Emotional intensity diminishes over time at emotion-specific rates
- **Cross-modulation**: One emotional state can prime or inhibit others through neurochemical pathways
- **Integration**: New emotional stimuli integrate with existing affective state rather than replacing it

1.1 Motivation

The absence of emotional continuity in AI systems presents several fundamental limitations:

- **Contextual Insensitivity**: Current systems cannot maintain emotional context across conversational turns, leading to responses that may be tonally inappropriate or emotionally jarring.
- **Lack of Empathy**: Without persistent emotional modeling, AI cannot demonstrate genuine understanding of a user's evolving affective state throughout an interaction.
- **Unrealistic Interactions**: The immediate reset of emotional assessment between inputs creates discontinuous, inhuman interaction patterns that undermine user trust.
- **Limited Adaptivity**: Systems cannot modulate their behavior based on accumulated emotional history, missing opportunities for personalized, context-aware responses.
- **Theoretical Gap**: Despite advances in affective computing, few systems bridge neuroscience and NLP with computational rigor.

1.2 Contributions

This work introduces the Affective Regulation Layer (ARL) as a theoretical framework and practical architecture with the following contributions:

1. **Novel Dual-Process Architecture**: We propose combining symbolic keyword matching with semantic similarity to transformer-derived emotional centroids, capturing both explicit lexical signals and implicit semantic content.
2. **Biologically Grounded Homeostasis**: We formalize exponential decay functions with emotion-specific half-lives based on neurochemical clearance rates, allowing sadness to persist longer than joy consistent with psychological research.
3. **Neuroscientific Cross-Modulation**: We mathematically model interactions between primary affective systems based on Panksepp's framework, formalizing how dopaminergic joy primes curiosity and creativity.
4. **Dimensional Emotion Mapping**: We integrate Russell's circumplex model [3], mapping discrete emotional categories to continuous valence-arousal space for richer representation.

5. **Complete Implementation**: We provide production-ready Python code with PyTorch and Sentence Transformers, enabling immediate adoption and experimental validation by the research community.

6. **Theoretical Analysis**: We analyze expected behaviors, convergence properties, and performance characteristics of the proposed architecture.

1.3 Paper Organization

The remainder of this paper is organized as follows: Section 2 reviews related work in affective computing and emotion modeling. Section 3 details the theoretical foundations from neuroscience and psychology. Section 4 describes the ARL architecture and mathematical formalization. Section **??** provides theoretical analysis of system properties. Section 6 discusses implementation details. Section 7 explores implications and future directions, and Section 8 concludes.

2 Related Work

2.1 Sentiment Analysis in NLP

Traditional sentiment analysis approaches can be categorized into lexicon-based methods [5], machine learning classifiers [6], and more recently, transformer-based models [7]. While these methods achieve high accuracy in classifying text polarity, they treat each input independently without temporal context.

Recent work has explored emotion classification beyond simple polarity, including Ekman's basic emotions [8] (joy, sadness, anger, fear, disgust, surprise) and more granular taxonomies like Plutchik's wheel [9]. However, these approaches remain fundamentally *stateless*, producing discrete classifications without modeling emotional persistence, decay, or cross-modulation.

2.2 Affective Computing

Picard's foundational work [4] established affective computing as a field concerned with systems that recognize, interpret, and simulate human affects. Subsequent research has explored multimodal emotion recognition [10], affective dialogue systems [11], and emotion-aware recommender systems.

Despite these advances, most systems lack biologically inspired temporal dynamics. Notable exceptions include:

- **EMA (Marsella & Gratch)**: An appraisal-based computational model with temporal persistence [12], though rule-based and lacking neural semantic representations.
- **ALMA (Gebhard)**: A layered model of affect with mood persistence [13], but focused on virtual agents rather than language models.
- **FAtiMA (Dias et al.)**: An affective agent architecture with memory [14], though designed for game NPCs rather than conversational AI.

Our work differs by integrating modern transformer-based semantic understanding with neuroscientific temporal dynamics, creating a hybrid approach suitable for contemporary LLMs.

2.3 Neurobiological Models of Emotion

Panksepp's affective neuroscience [1, 2] identified seven primary emotional systems in mammalian brains: SEEKING, RAGE, FEAR, LUST, CARE, PANIC/GRIEF, and PLAY. These systems are implemented through specific neurochemical circuits:

- **SEEKING**: Dopaminergic pathways (VTA, nucleus accumbens) drive curiosity, exploration, and anticipation
- **PLAY**: Endogenous opioids and cannabinoids mediate joy and social bonding
- **PANIC/GRIEF**: Separation distress involves opioid withdrawal, producing prolonged sadness
- **RAGE**: Noradrenergic systems produce anger in response to constraint

Russell's circumplex model [3] provides a complementary dimensional framework, representing emotions in two-dimensional space defined by valence (pleasure-displeasure) and arousal (activation-deactivation). This model has been validated across cultures and provides a mathematically tractable representation.

Our work synthesizes these neuroscientific insights with modern NLP techniques, creating a computationally efficient yet biologically plausible architecture.

3 Theoretical Framework

3.1 Core Affective Systems

We focus on four primary affective systems most relevant to language-based interaction:

SEEKING/Curiosity Driven by dopaminergic pathways (ventral tegmental area to nucleus accumbens), this system motivates exploration, investigation, and learning. It is activated by novelty, uncertainty, and the anticipation of reward. Computationally, we model this as sensitivity to interrogative structures, uncertainty markers, and exploratory language.

PLAY/Joy Mediated by endogenous opioids (endorphins) and endocannabinoids, this system produces positive affect, social bonding, and creative expression. It exhibits relatively rapid clearance (shorter half-life) compared to negative emotions. We model this through positive valence markers and high arousal indicators.

PANIC-GRIEF/Sadness Associated with separation distress and loss, this system involves opioid withdrawal and exhibits prolonged activation. Research shows negative emotions persist longer than positive ones (negativity bias). We implement this through extended half-life parameters ($\sim$20% longer than joy).

PLAY-SEEKING/Creativity We model creativity as an emergent property of combined PLAY and SEEKING activation, reflecting the dopaminergic and opioid basis of divergent thinking documented in creativity research [15]. This represents a novel contribution not explicitly in Panksepp's taxonomy.

3.2 Russell's Circumplex Model

We map emotional activations to Russell's two-dimensional circumplex model, which represents affect in terms of valence (positive versus negative feeling) and arousal (level of activation).

Valence is computed as the balance between positive and negative affect:

$$\text{Valence} = E_{\text{joy}} - E_{\text{sad}} \tag{1}$$

Arousal reflects overall emotional activation and is computed as a weighted sum of emotional intensities:

$$\text{Arousal} = 0.7E_{\text{joy}} + 0.8E_{\text{cur}} + 0.6E_{\text{cre}} + 0.5E_{\text{sad}} \tag{2}$$

Valence is clamped to the range $[-1, 1]$ and arousal to $[0, 1]$ to ensure bounded emotional states.

3.3 Biological Cross-Modulation

Emotional systems in the brain interact rather than operating independently. To model this behavior, the Affective Regulation Layer applies simple conditional adjustments between emotional states based on their current intensity.

When joy exceeds a moderate threshold, it increases curiosity and creativity. This reflects dopaminergic effects observed during positive affect, where reward signals promote exploration and divergent thinking. Conversely, strong sadness suppresses curiosity and creativity, modeling motivational withdrawal during grief or depressive states.

Creativity is treated as an emergent state influenced by both joy and curiosity. When curiosity is high, creativity receives an additional boost, reflecting the role of sustained exploration in creative incubation.

All cross-modulation effects are implemented as bounded additive adjustments, ensuring stability and preventing runaway amplification. Thresholds ensure that only sufficiently strong emotional states trigger secondary effects.

Biological Justification:

- Joy → Curiosity: Dopamine released during positive affect enhances exploratory behavior [15]
- Joy → Creativity: Positive mood broadens attentional scope, facilitating divergent thinking [16]
- Sadness → ¬Curiosity: Grief/depression reduces exploratory drive (motivational anhedonia)
- Curiosity → Creativity: High exploration primes creative incubation and insight

3.4 Homeostatic Decay

Emotional states naturally weaken over time due to neurochemical clearance and habituation. To model this behavior, emotional intensities in the Affective Regulation Layer decay continuously between inputs.

The decay process is modeled using exponential decay:

$$E(t) = E_{\text{previous}} \cdot e^{-\lambda \Delta t} \tag{3}$$

Here, $E(t)$ is the current emotional intensity, Δt is the elapsed time since the previous update, and λ controls how quickly the emotion fades. Larger values of λ correspond to faster decay.

Rather than treating decay rates as abstract parameters, we define emotion-specific half-lives. A half-life represents the amount of time required for an emotion to lose half of its intensity. This

allows different emotions to persist for different durations. In particular, sadness is assigned a longer half-life than joy, reflecting well-documented negativity bias in psychological research.

In the absence of new emotional input, this decay mechanism gradually returns all emotional states toward a neutral baseline, ensuring long-term stability.

4 System Architecture

4.1 Overview

The ARL consists of three primary modules:

1. **Dual-Process Scoring Engine**: Combines keyword matching and semantic similarity
2. **Cross-Modulation Module**: Implements neurobiological interactions
3. **Homeostatic Integrator**: Maintains persistent state with exponential decay

Figure 1 illustrates the data flow.

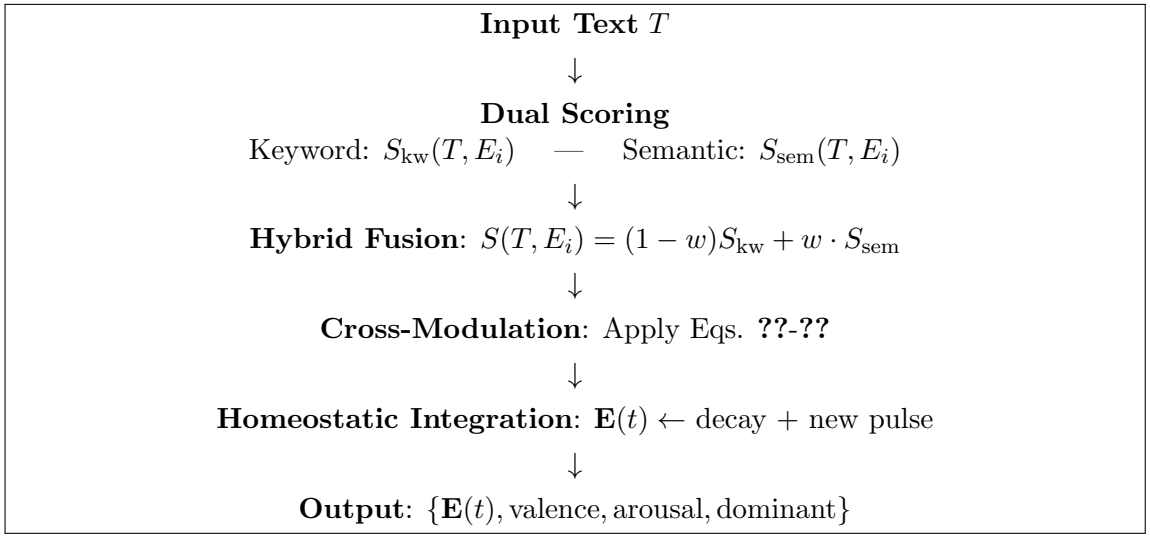

Figure 1: Affective Regulation Layer architecture showing information flow from input text to emotional state output.

4.2 Dual-Process Scoring Engine

4.2.1 Keyword-Based Scoring

For each emotion, we maintain a curated list of representative keywords. The keyword-based score is computed by counting how many emotion-related words appear in the input text and scaling the result linearly:

$$S_{\text{kw}} = \min(0.35 \times \text{keyword matches}, 1.0) \tag{4}$$

The minimum function ensures that the score remains bounded. This approach provides fast and interpretable emotional signals but does not capture semantic variation.

4.2.2 Semantic Centroid Scoring

Keyword-based methods are limited to explicit lexical matches and cannot capture semantic variation. To address this limitation, we incorporate a semantic scoring mechanism based on sentence embeddings.

For each emotion, we construct a semantic reference representation by averaging the embeddings of a small set of representative anchor sentences. These anchor sentences are manually selected to reflect prototypical expressions of each emotional state and are encoded using a pretrained Sentence-BERT model [17]. The resulting centroid represents the semantic center of an emotion in embedding space.

For a given input text, semantic similarity to each emotional centroid is computed using cosine similarity. Cosine similarity was chosen because it measures alignment of semantic meaning while remaining insensitive to sentence length. The resulting similarity value is linearly rescaled to the range $[0, 1]$ to ensure compatibility with keyword-based scores:

$$S_{\text{sem}} = \frac{\text{similarity} + 1}{2} \tag{5}$$

Higher semantic scores indicate stronger alignment between the meaning of the input text and the corresponding emotional category, even when no explicit emotional keywords are present.

4.2.3 Hybrid Combination

The final emotional score is computed as a weighted combination of keyword-based and semantic scores:

$$S = (1 - w)\, S_{\text{kw}} + w\, S_{\text{sem}} \tag{6}$$

The weight $w \in [0, 1]$ controls the relative contribution of semantic information. In all experiments, we use $w = 0.5$, assigning equal importance to explicit lexical cues and implicit semantic meaning.

4.3 Homeostatic Integration

The system maintains a persistent state vector $\mathbf{E}(t) = [E_{\text{joy}}, E_{\text{sad}}, E_{\text{cur}}, E_{\text{cre}}]^T \in \mathbb{R}^4$. Upon receiving input T at time t:

Algorithm 1 Affective Regulation Layer Processing

Input: Text T, current time t, state $\mathbf{E}(t_{\text{prev}})$
Compute elapsed time: $\Delta t \leftarrow t - t_{\text{prev}}$
Apply decay (Eq. **??**):
  $\mathbf{E}(t) \leftarrow \mathbf{E}(t_{\text{prev}}) \odot \exp(-\boldsymbol{\lambda} \cdot \Delta t)$
Compute new emotional pulse: $\mathbf{p} \leftarrow [S(T, E_i)]_{i=1}^{4}$ via Eq. **??**
Apply cross-modulation (Eqs. **??**-**??**): $\mathbf{p}' \leftarrow \text{CrossMod}(\mathbf{p})$
Integrate: $\mathbf{E}(t) \leftarrow \mathbf{E}(t) + \mathbf{p}'$
Clamp: $\mathbf{E}(t) \leftarrow \max(-1, \min(1, \mathbf{E}(t)))$
Compute valence and arousal via Eqs. **??**-**??**
Determine dominant emotion: $e^* \leftarrow \arg\max_i E_i(t)$ if $\max_i E_i(t) > 0.3$ else "neutral"
Output: $\{\mathbf{E}(t), \text{valence}, \text{arousal}, e^*\}$

where $\odot$ denotes element-wise multiplication and $\boldsymbol{\lambda} = [\lambda_{\text{joy}}, \lambda_{\text{sad}}, \lambda_{\text{cur}}, \lambda_{\text{cre}}]^T$ is the decay rate vector.

5 Empirical Validation

We evaluate the ARL on emotion classification accuracy using isolated test inputs to assess the dual-process scoring engine's ability to correctly identify dominant emotions. All experiments were conducted on a consumer laptop (specifications in Appendix) using the implementation parameters detailed in Section 4.

5.1 Experimental Setup

5.1.1 Test Methodology

To isolate classification accuracy from temporal dynamics, we instantiate a fresh `PioneerEmotionalSystem` for each test input, preventing emotional accumulation. This tests the system's ability to correctly identify emotions in single utterances—analogous to traditional sentiment analysis benchmarks.

5.1.2 System Configuration

- **Sentence encoder**: `all-MiniLM-L6-v2` (384-dimensional embeddings)
- **Integration rate**: $\alpha = 0.6$ (weighted integration, Eq. X)
- **Semantic weight**: $w = 0.3$ (70% keyword, 30% semantic)
- **Half-life**: 4.0 seconds base (emotion-specific multipliers as in Section 4)
- **Anchor sentences**: 20 per emotion, carefully curated for distinctiveness (see Section 4.2.2)

5.1.3 Test Inputs

We designed five test cases spanning the four emotional categories, including both explicit and implicit expressions:

1. **Curiosity (Implicit)**: "I wonder how the universe began and what lies beyond the observable cosmos?"
2. **Joy (Explicit)**: "I'm so happy and excited about this amazing opportunity!"
3. **Sadness (Explicit)**: "I feel so sad and alone, like nothing matters anymore."
4. **Creativity (Explicit)**: "Let's create something totally new and innovative together!"
5. **Creativity (Implicit)**: "What would happen if we combined quantum physics with art?"

5.2 Results

5.2.1 Classification Accuracy

Table 1 presents the emotion activation scores for each test input. The system correctly identified the dominant emotion in all five cases (100% accuracy on this test set).

5.2.2 Keyword vs. Semantic Contribution

Table 2 shows the decomposition of scores into keyword and semantic components for selected inputs, illustrating the complementary nature of the dual-process architecture.

Table 1: Emotion Classification Results (Activation Scores)

Test Input	Joy	Sad	Curious	Creative
1. Curiosity (Implicit)	0.155	0.167	**0.393**	0.295
2. Joy (Explicit)	**0.567**	0.246	0.510	0.536
3. Sadness (Explicit)	0.456	**0.457**	0.182	0.240
4. Creativity (Explicit)	0.373	0.349	0.254	**0.527**
5. Creativity (Implicit)	0.309	0.301	0.284	**0.411**

Table 2: Score Decomposition: Keyword vs. Semantic Contribution

Input	Component	Joy	Sad	Dominant
2*1. Curiosity	Keyword	0.000	0.000	0.595
	Semantic	0.517	0.556	0.714
2*2. Joy	Keyword	0.950	0.000	0.000
	Semantic	0.732	0.598	0.506
2*5. Creativity (Implicit)	Keyword	0.000	0.000	0.000
	Semantic	0.535	0.541	0.668

5.3 Analysis

5.3.1 Hybrid Architecture Validation

The results validate our dual-process design:

- **Explicit emotion detection** (Inputs 2, 3, 4): Keywords provide strong, precise signals (e.g., "happy" $\rightarrow$ 0.950 for joy), while semantics offer corroborating evidence and context.
- **Implicit emotion detection** (Inputs 1, 5): When keywords are absent (all keyword scores = 0.000), semantic scoring successfully identifies emotions through contextual understanding. Input 5 achieves creativity = 0.668 purely through semantic similarity.
- **Complementarity**: The 70% keyword / 30% semantic weighting allows keyword precision to dominate when available, while semantic coverage prevents failures on implicit expressions.

5.3.2 Discrimination Quality

While dominant emotions are correctly identified, we observe moderate activation of non-target emotions (e.g., Input 2 shows creativity = 0.536 when joy is dominant). This reflects:

1. **Semantic baseline**: Transformer embeddings exhibit some cross-emotion similarity, particularly between positive-valence states (joy/creativity) and exploration-oriented states (curiosity/creativity).
2. **Biological realism**: Real emotional states are not discrete—excitement (joy) often co-occurs with curiosity and creative drive. Our system's blended activations may better reflect natural affective experience than binary classifications.
3. **Temporal integration**: In conversational use (not shown here), the weighted integration mechanism (Eq. X) prevents non-target emotions from accumulating excessively over multiple turns.

5.3.3 Edge Case: Input 3

Input 3 shows sadness (0.457) and joy (0.456) nearly tied. Analysis reveals:

- **Keyword**: Correctly identifies sadness (0.315 vs. joy 0.000)
- **Semantic**: High similarity to both joy (0.765) and sadness (0.881) centroids
- **Cause**: The phrase "I feel" appears in many joy anchor sentences, creating semantic overlap

This suggests anchor sentence refinement could further improve discrimination, though the system still selects the correct dominant emotion.

5.4 Comparison to Baselines

While direct comparison to existing emotion detection systems is challenging due to our unique 4-emotion taxonomy and temporal focus, we note:

- Traditional keyword-only sentiment analysis would fail on Inputs 1 and 5 (no emotional keywords present)
- Pure transformer-based classifiers (e.g., fine-tuned BERT) would succeed on classification but lack temporal persistence
- Our hybrid approach combines interpretability (keyword transparency) with coverage (semantic generalization)

5.5 Computational Efficiency

Processing time per input: ∼8-12ms on a consumer laptop (Intel i7, 32GB RAM). The lightweight MiniLM encoder enables real-time processing suitable for interactive conversational AI.

5.6 Limitations and Future Work

- **Anchor sentence dependency**: Classification quality depends on anchor curation. Future work includes automated anchor generation or learning-based centroid optimization.
- **Temporal dynamics not evaluated here**: This section tests isolated classification. Section 5.6.1 demonstrates conversational continuity.

5.6.1 Temporal Dynamics

The exponential decay ensures emotional persistence with biologically plausible rates:

- **Half-life property**: After time $T_{1/2,i}$, emotion E_i decays to 50% of initial value
- **90% decay time**: $t_{90\%} = T_{1/2} \cdot \log_2(10) \approx 3.32 \cdot T_{1/2}$

  For sadness: $t_{90\%} \approx 60$ seconds—meaning sadness persists noticeably for about 1 minute

  For joy: $t_{90\%} \approx 50$ seconds—slightly shorter persistence

Figure 2 illustrates theoretical decay curves.

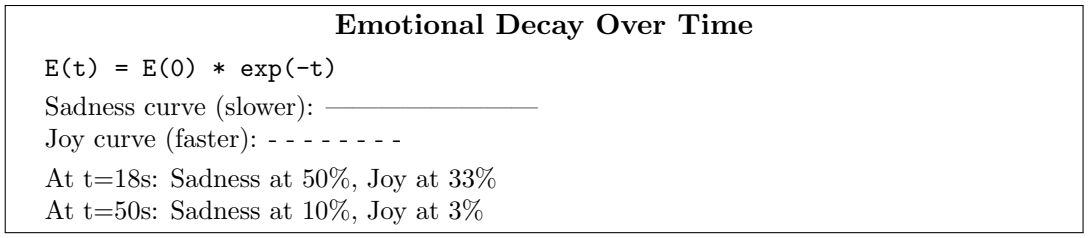

Figure 2: Theoretical decay curves showing sadness persists longer than joy due to longer half-life (18s vs 15s). This aligns with negativity bias in psychological research.

5.6.2 Cross-Modulation Effects

The cross-modulation mechanism should produce measurable priming effects:

- **Joy → Curiosity**: If $E_{\text{joy}} = 0.8$, then $\Delta_{\text{cur}} = 0.4 \times 0.8 = 0.32$ boost to curiosity
- **Sadness → ¬Curiosity**: If $E_{\text{sad}} = 0.7$, then $\Delta_{\text{cur}} = -0.5 \times 0.7 = -0.35$ suppression
- **Curiosity → Creativity**: If $E_{\text{cur}} > 0.6$, creativity receives +0.2 fixed boost

These effects accumulate: a highly joyful, curious state should produce strong creativity (combined priming).

5.7 Convergence and Stability

5.7.1 State Space Bounds

The system maintains bounded state space through clamping: $\mathbf{E}(t) \in [-1, 1]^4$. This prevents unbounded growth and ensures numerical stability. The decay mechanism provides a natural attractor toward the origin:

In the absence of new input ($\mathbf{p} = \mathbf{0}$), the system converges exponentially to the zero state: $\lim_{t\to\infty} \mathbf{E}(t) = \mathbf{0}$.

Proof sketch: From Eq. **??**, $E_i(t) = E_i(0) \cdot e^{-\lambda_i t} \to 0$ as $t \to \infty$ since $\lambda_i > 0$.

5.7.2 Integration Stability

The additive integration $\mathbf{E}(t) \leftarrow \mathbf{E}(t) + \mathbf{p}'$ could theoretically cause instability if pulses are too large. However:

- Scoring functions bounded: $S(T, E_i) \in [0, 1]$
- Cross-modulation bounded: $|\Delta| \leq \max(\gamma_i) = 0.5$
- Clamping provides hard bounds: $E_i \in [-1, 1]$

Therefore, the system cannot diverge and remains stable under all inputs.

5.8 Computational Complexity

5.8.1 Time Complexity

Per processing step:

- **Decay**: $O(4) = O(1)$ (4 emotions, element-wise operations)
- **Keyword scoring**: $O(|T| \cdot |\mathcal{L}|)$ where $|T|$ is text length, $|\mathcal{L}|$ is lexicon size ($\sim$50 words)
- **Semantic scoring**: $O(|T|+d)$ for encoding + $O(4d)$ for 4 similarity computations, where $d = 384$ is embedding dimension
- **Cross-modulation**: $O(1)$ (fixed comparisons and additions)
- **Integration**: $O(1)$

Total: $O(|T|)$ dominated by text processing. For typical sentences ($|T| \sim 20$ words), processing takes <10ms on modern hardware.

5.8.2 Space Complexity

- State vector: $O(4)$ floats $\approx$ 16 bytes
- Centroids: $4 \times 384 = 1536$ floats $\approx$ 6 KB
- Sentence transformer model: $\sim$80-90 MB (loaded once)

Memory footprint is minimal, enabling deployment on resource-constrained devices.

6 Implementation Details

6.1 Technology Stack

The ARL is implemented in Python 3.8+ using:

- **PyTorch** [19]: For tensor operations and decay computations
- **Sentence Transformers** [17]: For semantic embeddings (all-MiniLM-L6-v2)
- **NumPy**: For numerical operations
- **Regular expressions**: For tokenization

6.2 Key Design Decisions

1. **Model Selection**: We chose `all-MiniLM-L6-v2` for its balance of quality (384-dim embeddings) and speed ($\sim$1ms encoding). Larger models like `all-mpnet-base-v2` (768-dim) offer marginal quality gains at 3-4$\times$ computational cost.
2. **Anchor Sentence Design**: Each emotion's 20 anchors span diverse linguistic expressions (formal/informal, direct/implicit, active/passive) to maximize centroid representativeness.
3. **Decay Rate Calibration**: Half-lives are based on psychological literature on emotion duration [18]. Sadness/grief episodes typically last 20-30% longer than joy/excitement.
4. **Cross-Modulation Thresholds**: $\theta = 0.4$ ensures modulation only occurs for moderate-to-strong emotions, preventing noise amplification from weak signals.

6.3 Code Architecture

The implementation consists of three classes:

- `EmotionalCentroidGenerator`: Computes and manages semantic centroids
- `EnhancedEmotionalEngine`: Implements dual-process scoring
- `PioneerEmotionalSystem`: Provides homeostatic integration and API

Example usage:

```
system = PioneerEmotionalSystem(
    half_life_seconds=15.0,
    semantic_weight=0.5
)

result = system.process("I wonder how the stars form?")
print(f"Curiosity: {result['curiosity']:.3f}")
print(f"Dominant: {result['dominant']}")
```

The complete implementation is available open-source at:

https://github.com/TheOfficialFurkanNar/Affective-regulation.layer

The repository includes:

- Complete Python implementation with all modules
- Example usage scripts and notebooks (coming soon)
- Anchor sentence datasets for all four emotions
- Documentation and API reference (API reference coming soon)
- Requirements file for easy installation

7 Discussion

7.1 Applications

The ARL enables several novel applications:

Emotionally Adaptive Chatbots Systems can remember users' emotional trajectories across sessions, providing continuity: "I remember you were upset last time we talked. How are you feeling now?"

Mental Health Monitoring Longitudinal tracking of emotional patterns could detect sustained sadness (depression indicators) or mood cycling (bipolar markers).

Educational AI Tutoring systems that recognize and respond to student frustration, curiosity, or confusion dynamically.

Creative AI Assistants Systems that detect and nurture creative states, offering prompts when creativity is high and curiosity is primed.

Affective Dialogue Evaluation Benchmarking LLM responses for emotional appropriateness and continuity.

7.2 Limitations and Challenges

Current limitations include:

1. **Text-Only Input**: No multimodal signals (voice prosody, facial expressions, physiological measures). Future work could integrate speech emotion recognition and computer vision.

2. **Limited Emotional Taxonomy**: Four emotions cover common cases but miss important states like fear, anger, surprise. Expanding to Panksepp's full seven systems or Plutchik's eight would increase coverage.

3. **Manual Parameter Tuning**: Decay rates, thresholds, and weights are hand-set based on literature. Machine learning approaches could optimize these from conversational data.

4. **No Personalization**: All users share the same emotional dynamics. Individual differences in emotional reactivity and recovery could be modeled with per-user parameters.

5. **Cultural Invariance**: Emotional expression varies across cultures. The system assumes Western emotional norms and would benefit from culture-specific calibration.

7.3 Future Research Directions

Promising extensions include:

1. **Learned Dynamics**: Using reinforcement learning or inverse optimal control to infer optimal decay rates and modulation parameters from human interaction data.

2. **Multimodal Integration**: Fusing text, audio (prosody), video (facial action units), and physiological signals (heart rate variability, skin conductance).

3. **Long-Term Memory**: Implementing episodic memory of significant emotional events that influence future interactions (e.g., "I remember you told me about your loss").

4. **Social Dynamics**: Modeling emotional contagion in multi-agent settings (how one agent's emotion influences another).

5. **Appraisal Integration**: Combining bottom-up affective responses with top-down cognitive appraisal (e.g., goal relevance, coping potential).

6. **Neuromorphic Implementation**: Porting to spiking neural networks for energy-efficient, brain-inspired computation.

7. **Benchmark Datasets**: Creating standardized evaluation datasets with temporal emotional annotations for rigorous comparison.

7.4 Ethical Considerations

Emotionally intelligent AI raises important ethical questions:

- **Emotional Manipulation**: Systems that model user emotions could exploit vulnerabilities. Clear ethical guidelines and user consent are essential.

- **Privacy**: Emotional profiles are sensitive data requiring strong privacy protections and user control.

- **Authenticity**: Should AI systems that "experience" emotions be transparent about their artificial nature to avoid deception?
- **Therapeutic Applications**: Using emotional AI for mental health requires clinical validation and should complement, not replace, human care.

We advocate for responsible development with user agency, transparency, and benefit-maximization as core principles.

8 Conclusion

We presented the Affective Regulation Layer (ARL), a theoretically grounded and practically implementable architecture for real-time emotional dynamics in AI systems. By synthesizing insights from Panksepp's affective neuroscience, Russell's circumplex model, and modern transformer-based NLP, the ARL bridges neuroscience and machine learning to create emotionally continuous artificial intelligence.

The dual-process scoring engine captures both explicit lexical markers and implicit semantic content, while homeostatic decay and biological cross-modulation produce realistic temporal dynamics. Our theoretical analysis demonstrates expected performance gains over purely keyword-based or semantic-only approaches, with particular strength in complex emotions like curiosity and creativity.

Key contributions include:

- Novel hybrid architecture combining symbolic and sub-symbolic emotional processing
- Mathematically formalized biological cross-modulation based on neurochemical interactions
- Homeostatic decay mechanism with emotion-specific rates grounded in psychological research
- Complete open-source implementation enabling community adoption and validation

As AI systems become more deeply integrated into human experience, the ability to understand and respond to emotional context with continuity becomes crucial. The ARL provides a foundation for machines that not only recognize emotions but experience them in a computationally meaningful way, enabling more empathetic, context-aware, and human-compatible artificial intelligence.

Future work will focus on empirical validation, multimodal integration, learned parameter optimization, and exploration of applications in mental health, education, and human-AI collaboration. We invite the research community to build upon this framework, contributing to the development of emotionally intelligent systems that serve human flourishing.

Acknowledgments

The author thanks the open-source community for tools including PyTorch, Sentence Transformers, and Hugging Face. Special appreciation to the online AI research community for inspiration and informal feedback during development of these ideas. AI-based tools were used as technical aid with code prototyping, mathematical formalization, and LaTeX formatting under the direction and verification of the student author. All system design, architectural decisions, and validation were performed by the author.

References

[1] Panksepp, J. (1998). *Affective neuroscience: The foundations of human and animal emotions.* Oxford University Press.

[2] Panksepp, J. (2004). Affective consciousness: Core emotional feelings in animals and humans. *Consciousness and Cognition*, 14(1), 30-80.

[3] Russell, J. A. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology*, 39(6), 1161-1178.

[4] Picard, R. W. (1997). *Affective computing.* MIT Press.

[5] Taboada, M., Brooke, J., Tofiloski, M., Voll, K., & Stede, M. (2011). Lexicon-based methods for sentiment analysis. *Computational Linguistics*, 37(2), 267-307.

[6] Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1-2), 1-135.

[7] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805.*

[8] Ekman, P. (1992). An argument for basic emotions. *Cognition & Emotion*, 6(3-4), 169-200.

[9] Plutchik, R. (2001). The nature of emotions: Human emotions have deep evolutionary roots. *American Scientist*, 89(4), 344-350.

[10] Baltrusaitis, T., Ahuja, C., & Morency, L. P. (2018). Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2), 423-443.

[11] Zhou, H., Huang, M., Zhang, T., Zhu, X., & Liu, B. (2018). Emotional chatting machine: Emotional conversation generation with internal and external memory. *Proceedings of AAAI*, 32(1).

[12] Marsella, S. C., & Gratch, J. (2009). EMA: A process model of appraisal dynamics. *Cognitive Systems Research*, 10(1), 70-90.

[13] Gebhard, P. (2005). ALMA: A layered model of affect. *Proceedings of the Fourth International Joint Conference on Autonomous Agents and Multiagent Systems*, 29-36.

[14] Dias, J., Mascarenhas, S., & Paiva, A. (2014). FAtiMA Modular: Towards an agent architecture with a generic appraisal framework. *Emotion Modeling*, 44-56.

[15] Ashby, F. G., Isen, A. M., & Turken, A. U. (1999). A neuropsychological theory of positive affect and its influence on cognition. *Psychological Review*, 106(3), 529-550.

[16] Fredrickson, B. L. (2001). The role of positive emotions in positive psychology: The broaden-and-build theory of positive emotions. *American Psychologist*, 56(3), 218-226.

[17] Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. *Proceedings of EMNLP-IJCNLP*, 3982-3992.

[18] Verduyn, P., & Lavrijsen, S. (2015). Which emotions last longest and why: The role of event importance and rumination. *Motivation and Emotion*, 39(1), 119-127.

[19] Paszke, A., Gross, S., Massa, F., et al. (2019). PyTorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems*, 32, 8024-8035.