Gold from Bronze: Training a Stable Affective Neural Network by Refining Noisy LLM-Generated Data

First Author: ***Shivansh Sharma***

Abstract

Affective computing systems traditionally rely on rule-based engines or large pre-defined datasets, both of which limit adaptability, scalability, and emotional nuance. In this work, I propose a neural method for modeling stateful emotion transitions using synthetic hormone-inspired data generated via large language models (LLMs). The objective is to learn a continuous, physiologically-motivated function mapping social-emotional inputs to next-state hormone vectors without ground-truth physiological data.

I present a teacher–student framework where a curated prompt induces an LLM (Gemini 2.5 Flash) to act as a pseudo-expert in affective reasoning, generating 400 initial samples and 20 hand-crafted edge-case samples. After feature cleanup and dataset expansion to ~1000 samples, I train a compact feedforward neural network to map (hormones, user intent logits) → next hormones.

Through a sequence of ablations, I identify the core challenges: (1) a “dying ReLU” failure mode causing convergence plateaus at MAE ≈ 9, (2) high-volatility synthetic labels that penalize stable predictions, and (3) output activation saturation leading to a zero-bias collapse near the lower bound. I systematically resolve these issues by (i) replacing ReLU with ELU, (ii) removing noisy engineered features, (iii) augmenting with targeted edge-case samples, (iv) applying learning-rate decay, and (v) implementing a Parametric Bounded Linear Activation (PBLA), a learnable hard-sigmoid output function.

The final model achieves **5.60 training MAE and 5.25 test MAE**, with Test < Train indicating strong regularization (notably due to dropout deactivation during inference). I validate the model using a deterministic physiological mapping layer, demonstrating that the network effectively learns complex behaviors such as emotional inertia, de-escalation, and stability filtering. Error analysis reveals that most high test errors arise not from model failures but from intentional rejections of implausibly “spiky” labels generated by the teacher model—a phenomenon I call **Gold from Bronze**, where a model trained on noisy synthetic signals learns a smoother, more coherent emotional logic than its labels.

This work demonstrates that small neural networks can learn complex affective state-transition dynamics from noisy synthetic data when guided by principled activation design, data-centric iteration, and tailored regularization.

---

1. Introduction

Modeling human emotion in computational systems remains challenging due to the difficulty of collecting labeled affective data and the inherently ambiguous, continuous nature of emotional states. Most production systems rely on rule-based emotional engines: long sequences of if/else statements encoding hand-designed transitions across emotions. Such systems are brittle, unscalable, and incapable of emerging behavior.

My initial system was a rule engine, containing many conditional branches. Efforts to extend it quickly revealed the limitations of symbolic logic for continuous affective modeling.

To move beyond brittle rules, I sought to learn emotional state transitions directly from data. However, real hormone-level or physiological datasets are scarce, private, or ethically sensitive. This led us to a novel idea; use a large language model as a *synthetic teacher* to generate plausible “hormone vectors” representing modeled internal emotional states.

This paper describes the process of constructing such a dataset, training a small neural network on it, and discovering several subtle failure modes. Including dying ReLUs, noisy labels, activation saturation, and non-grounded volatility, and ultimately solving them using a combination of architectural innovations and data-centric refinement.

---

2. Related Work

Affective Computing

Traditional affective computing models rely on symbolic rules or classification-based sentiment systems. Few works attempt *continuous* hormone-style regression modeling due to lack of real data.

Synthetic Labeling via LLMs

Teacher–student pipelines have emerged in programming, reasoning, and preference modeling. Here, I extend the idea to *affective state transitions*, using an LLM to simulate a plausible emotional physiology.

Activation Function Pathologies

ReLU's susceptibility to dying units is well known, particularly under L1 loss, which induces large gradients for outliers. ELU and variants address this by ensuring negative-region gradients remain non-zero.

Optimizers

AdamW's decoupled weight decay yields faster convergence and better generalization than vanilla Adam. I validate this on small-model regression.

This work combines these threads into a coherent affective regression framework.

---

3. Methodology

3.1 Dataset Generation (LLM Distillation)

Initial 400 Samples

I structured prompts instructing Gemini 2.5 Flash to act as an "expert emotional state modeler" given inputs:

- current hormone vector (5 values)
- user intent logits (8 values from an NLP classifier)

and output:

- next hormone vector (oxytocin, serotonin, cortisol, adrenaline, dopamine)

These 400 samples were split 370/30 (train/test).

Edge-Case Augmentation

Synthetic data tends to cluster around "common-case" scenarios. I hand-generated 20 edge-case examples with rare combinations (e.g., high stress + happiness, low oxytocin + high affection) to prevent blind spots. New dataset: 420 samples → 378 train / 42 test.

Feature Cleanup

Originally, 5 engineered features (e.g., cortisol × angry_logit) were appended, increasing input dimension to 18. Experiments revealed these features degraded performance, so I reverted to the 13-dimensional base input: 5 hormones + 8 intent logits.

Dataset Expansion to 1000

For final training, I expanded to ~1000 samples to stabilize gradients for PBLA and AdamW.

3.2 Model Architecture

The final network is a compact FNN:

- **Input:** 13 features (current hormones + user intent logits)
- **Hidden layers:** one layer of 320 ELU-activated units
- **Dropout:** 0.45
- **Output:** 5 hormone predictions, each passed through PBLA

Activation Functions

- **Hidden:** ELU — solves dying-ReLU gradients.
- **Output: Parametric Bounded Linear Activation (PBLA)**
  A learnable HardSigmoid variant:
  `PBLA(x) = clamp(k * x + b, min=L, max=U)`
- where `k` (slope) and `b` (bias) are *per-output-unit* learnable parameters, and `L` and `U` are fixed lower and upper bounds (e.g., 0 and 100) for the hormone values.

PBLA avoids zero-floor saturation, supports flexible slopes near boundaries, and stabilizes small-signal predictions (e.g., 1.7 instead of 0).

3.3 Training Procedure

- **Loss:** L1 (MAE), robust to outliers.
- **Optimizer:** AdamW, weight_decay=0.02.
- **Batch Size:** empirically optimal range 100–180.
- **Epochs:** 4000 (converged visually, no overfit).
- **LR Decay:** step decay at 2000 epochs × 0.1.

The train/test split remained **static 90/10** across all experiments.

3.4 Interpretation Layer (Physiological Mapping)

To ensure the predicted hormone vectors represent coherent emotional states, I defined a deterministic mapping layer between hormones and emotions. This layer converts the continuous 5-dimensional hormone vector into categorical emotion states based on simplified physiological hypotheses. This ensures that the model's numeric outputs are grounded in interpretable behavior rather than arbitrary regression targets.

The mapping rules used to evaluate the model are:

- **Happy:** Modeled as reward-seeking (Serotonin *S*, Dopamine *D*) gated by the absence of stress (Cortisol *C*).
    - *Equation:* Happy = [(S + 0.3D) / 1.3] × (1 – C)
- **Sad:** Modeled as high stress (Cortisol) in the absence of well-being (Serotonin).
    - *Equation:* Sad = C × (1 – S)
- **Angry:** Modeled as a combination of threat (Adrenaline *A*) and stress (Cortisol *C*), normalized by their combined drivers.
    - *Equation:* Angry = (A + 0.7C) / 1.7
- **Caring:** A pro-social state driven by Oxytocin (*O*), suppressed only by panic (Adrenaline *A*).
    - *Equation:* Caring = O × (1 – A)
- **Love:** A peak state requiring both bonding (*O*) and well-being (*S*), highly fragile to any stressor.
    - *Equation:* Love = (O × S) × (1 – max(C, A))

This mapping reveals that "Love" is mathematically distinct from "Caring" in my framework. It requires the simultaneous presence of high bonding and high mood, making it a naturally rare state in the dataset.

The constants in these equations (e.g., 0.3, 0.7) act as damping factors. I determined these values empirically to prevent the calculated emotions from easily saturating at 100%. This reflects the biological reality that extreme emotional intensity is rare; it typically requires multiple hormonal drivers to peak simultaneously, rather than a single input spiking alone.

---

4. Chronological Progress & Ablation Findings

Phase 1 — Baseline Model (Failure)

Result: Train ≈ 8.0, Test ≈ 9.0 MAE
ReLU activations caused dying neurons, compounded by L1 loss spikes. Hyperparameter searches (batch sizes 12–370, dropout, LR sweeps, architectures 64→512) all failed to break the plateau.

Phase 2 — Fixing the Activation (Breakthrough #1)

Replacing ReLU with ELU removed the gradient-dead regions.

New result: Train ≈ 7.0, Test ≈ 7.0
This was the first successful convergence.

Phase 3 — Data-Centric Refinement (Breakthrough #2)

- Removed engineered features (18 → 13 inputs).
- Augmented dataset with the 20 edge-case samples described in Section 3.1.
- Used learning-rate decay to refine small-signal behavior.

New result: Train ≈ 5.2, Test ≈ 6.0

I discovered the **Gold from Bronze** phenomenon:
The model often rejected extreme artificial spikes in labels. These labels created large MAE but better qualitative behavior.

Phase 4 — Final Model (Breakthrough #3)

- Introduced **PBLA** to replace relu6_scaled.
- Switched to **AdamW**.
- Increased dataset to **~1000 samples**.
- Found optimal batch size ~100–180.

Final result:

- **Train MAE:** 5.60
- **Test MAE: 5.25**

Importantly: **Test < Train**, indicating ideal regularization under label noise.

---

5. Results

5.1 Training Procedure

Phase	Key Change	Train MAE	Test MAE
Baseline	ReLU	~8.0	~9.0
Phase 2	ELU	~7.0	~7.0
Phase 3	Data cleanup + LR decay	~5.2	~6.0
Phase 4	PBLA + AdamW + larger dataset	**~5.60**	**~5.25**

Additional observations:

- ~80% of samples have ≤2 MAE.
  ~5% of samples (high-volatility labels) inflate total MAE by 0.7–1.0.
- PBLA nearly eliminates zero-bias predictions.
- The model generalizes better than it trains, a sign of correct regularization (notably due to dropout deactivation during inference).

5.2 Qualitative Analysis of State Transitions

While MAE measures numerical convergence, it does not guarantee emotional intelligence. To validate the model's learned behavior, I analyzed specific transition samples to verify if the "Gold from Bronze" effect resulted in socially coherent interactions.

I observed three distinct emergent behaviors that suggest the model learned to generalize rather than memorize:

Table 2: Qualitative Failure & Success Modes

Scenario	Current State	User Input (Logits)	Predicted Next State	Interpretation
De-escalation	Angry (44.0)	Loving (94.1)	**Happy (30.9)**	**Success:** The model correctly identifies that high affection can resolve moderate anger, flipping the valence.
Inertia (Resistance)	Angry (94.5)	Thankful (74.9)	**Angry (81.8)**	**Success:** The model exhibits emotional inertia. A simple "thank you" is insufficient to resolve extreme rage immediately, resulting in a dampened but still angry state.
Stability Filtering	Happy (50.2)	Angry (96.0)	**Angry (45.0)**	**Success (Gold from Bronze):** While the user input was extremely negative (96.0), the model predicted a moderate response (45.0) rather than an extreme spike, rejecting the volatility of the teacher labels.

This qualitative audit confirms that the low Test MAE corresponds to smooth, explainable affective dynamics, validating the architecture's ability to act as a stability filter for noisy synthetic data.

6. Discussion

6.1 Learning in the Absence of Ground Truth

Real physiological signals are unobservable in this context. The model learns an *approximation* of emotional dynamics, not biology. Yet it captures consistent, human-interpretable transitions.

Furthermore, the successful decoding of predicted vectors via the Interpretation Layer (Section 3.4) confirms that the learned manifold is structurally coherent. For instance, the model consistently predicts low Cortisol levels when the context implies 'Love' or 'Happiness,' effectively internalizing the physiological gating mechanisms (e.g., that stress inhibits well-being) purely through regression on the synthetic dataset.

6.2 Rejecting Noise: The Gold from Bronze Principle

In several cases, teacher labels exhibit implausible hormone swings (e.g., oxytocin 90 → 15 in one step).
The model correctly *refuses* to imitate these, predicting softer transitions instead.
This increases MAE but **increases realism**.

6.3 Why This Model Works Better Than the Teacher

LLMs generate emotionally coherent but sometimes inconsistent hormone values.
A small neural network, trained with strong regularization and successful architecture choices, becomes a **stability filter**.

6.4 Lessons for Affective Computing

This work suggests a new workflow:

1. Use LLMs to generate approximate emotional data.
2. Train a small grounded model to *refine* that data.
3. Architect activations to prevent pathological failure modes.
4. Embrace smoothing of noisy supervision.

---

7. Conclusion

I have demonstrated a compact and effective affective state-transition model trained entirely on synthetic hormone-inspired data created using a teacher LLM. Through systematic experimentation, architectural changes, data-centric iteration, and a novel output activation (PBLA), I reduced MAE from ~9.0 to ~5.25. Crucially, qualitative validation via a deterministic physiological mapping layer confirms that this numerical improvement corresponds to

interpretable, socially intelligent behavior, successfully distinguishing between transient high-excitement states and stable bonding dynamics.

Beyond performance improvement, this work provides principled methods for training models on inherently noisy, synthetic, soft-labeled data.

The approach is suitable for real-time conversational agents, affective simulators, and reinforcement-learning agents requiring internal emotional states.

---

8. Future Work

- Expand dataset using multi-teacher prompting (Gemini + GPT-4 + Claude)
    - Because each teacher has a unique bias, using multiple teachers will reduce the total bias in the student model.
- Extend the model to a recurrent or attention-based architecture
    - Incorporating GRUs or lightweight attention would allow the system to capture higher-order temporal dynamics. While the current model successfully maintains state inertia via the hormone vector, a recurrent or attention-based architecture would enable the agent to model trajectory-dependent phenomena, such as emotional accumulation, sensitization, or refractory periods.
- Evaluate long-horizon stability with simple biological constraints
    - Testing multi-step rollouts and integrating basic physiological coupling rules may improve long-term coherence and interpretability.

---

9. References

Clevert, D. A., Unterthiner, T., & Hochreiter, S. (2016). Fast and Accurate Deep Network Learning by Exponential Linear Units (ELUs). *In Proceedings of the 4th International Conference on Learning Representations (ICLR)*.

Gao, Q., He, L., & Chen, Z. (2021). A Survey on Dialogue State Tracking. *arXiv:2105.02388*.

Gemini Team, Google. (2025). Gemini 2.5: Pushing the Frontier with Advanced Reasoning, Multimodality, Long Context, and Next Generation Agentic Capabilities. *arXiv:2507.06261*.

Kingma, D. P., & Ba, J. (2014). Adam: A Method for Stochastic Optimization. *arXiv:1412.6980*.

Lu, L., Shin, Y., Su, Y., & Karniadakis, G. E. (2019). Dying ReLU and Initialization: Theory and Numerical Examples. *arXiv:1903.06733*.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., ... & Chintala, S. (2019). PyTorch: An Imperative Style, High-Performance Deep Learning Library. *Advances in Neural Information Processing Systems (NeurIPS)*, 32.

Rashkin, H., Smith, E. M., Li, M., & Boureau, Y. L. (2019). Towards Empathetic Open-domain Conversation Models: A New Benchmark and Dataset. *In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Wang, Y., Kaddour, J., & Winstanley, L. (2024). Large Language Models as Data Augmenters. *In Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)*.