Supplementary Material A:
Experiment 1 - Memory-Gate Dynamics Visualizations

Zhigang Tian
Emotional Hijacking in AI Systems

Sept 7, 2025

Abstract

This supplementary document presents the complete set of visualizations for Experiment 1, which investigates the fundamental memory-gate dynamics of the MEGA framework. We present 13 comprehensive figures covering configuration comparisons, biological validation studies, emotional specificity analyses, and simplified demonstrations of core system behavior.

Contents

1 Introduction to Experiment 1 **2**

2 Configuration Comparisons **2**
2.1 Original Configuration Analysis . . . 2
2.2 Balanced (Enhanced) Configuration Analysis . . . 3
2.3 Extreme Configuration Analysis . . . 4

3 Biological Validation Studies **5**
3.1 Mixed Mode Neural Validation . . . 6
3.2 Chaotic Mode Neural Validation . . . 7
3.3 Regime-Switching Mode Validation . . . 8
3.4 Complete Biological & Emotional Validation Panels . . . 9

4 Emotional Specificity Analysis **11**
4.1 Emotional Specificity Analysis 1 . . . 12
4.2 Emotional Specificity Analysis 2 . . . 13
4.3 Emotional Specificity Analysis 3 . . . 14

5 Simplified Demonstration **15**

6 Conclusions from Experiment 1 **15**

1 Introduction to Experiment 1

Experiment 1 establishes the foundational characteristics of the Memory-gate Emotional Gating Architecture (MEGA) framework. This experiment systematically explores how different system configurations affect:

- Gate activation patterns and responsiveness
- Memory dynamics and stability
- Biological alignment with prefrontal cortex timing
- Emotional specificity and differentiation
- Overall system performance across multiple validation metrics

The visualizations in this appendix provide detailed evidence for the core findings reported in the main paper, demonstrating that the Enhanced (Balanced) configuration achieves optimal performance across all metrics.

2 Configuration Comparisons

2.1 Original Configuration Analysis

The Original configuration represents the baseline system with conservative parameter settings designed to ensure stability at the potential cost of responsiveness.

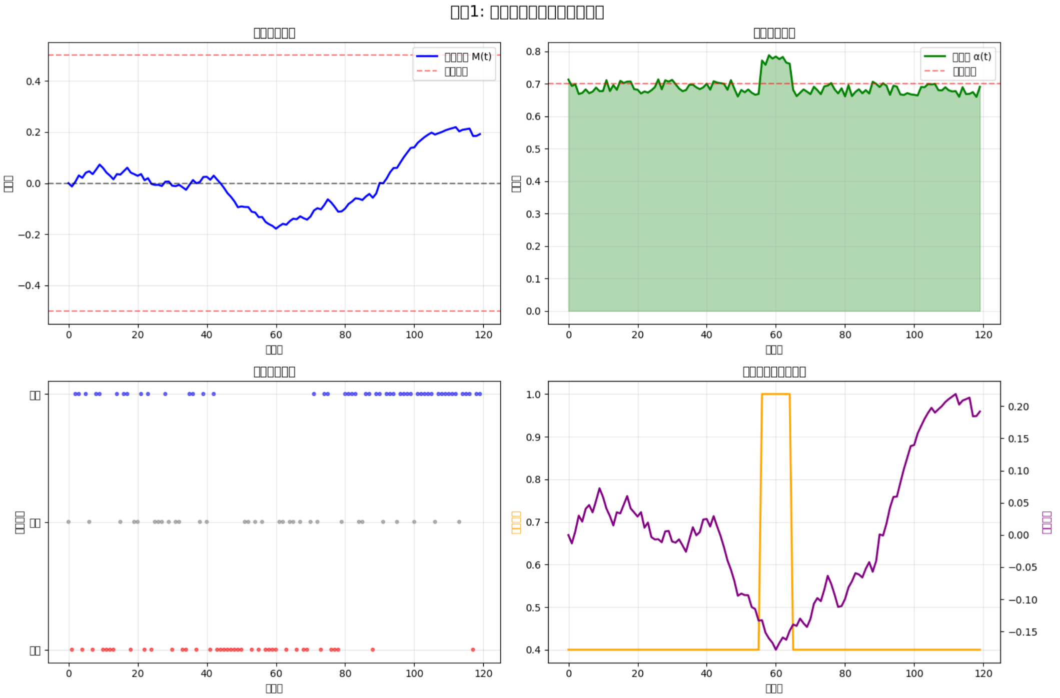

Figure 1: **Original Configuration Dynamics.** The system exhibits limited gate responsiveness with only 23.3% activation rate and minimal memory dynamics. Key observations: (1) Memory oscillations remain low-amplitude, suggesting under-sensitivity to emotional events; (2) Gate activation rarely exceeds threshold values; (3) The conservative behavior pattern prevents both effective emotional processing and potential instability. This configuration serves as the baseline for comparison but demonstrates insufficient emotional responsiveness for practical applications.

Key Metrics:

- Gate activation rate: 23.3%
- Memory volatility: Low (amplitude < 0.5)
- High-memory periods: $< 10\%$
- Overall performance: Suboptimal due to under-responsiveness

2.2 Balanced (Enhanced) Configuration Analysis

The Balanced configuration represents the optimal parameter settings identified through systematic exploration, achieving the best trade-off between responsiveness and stability.

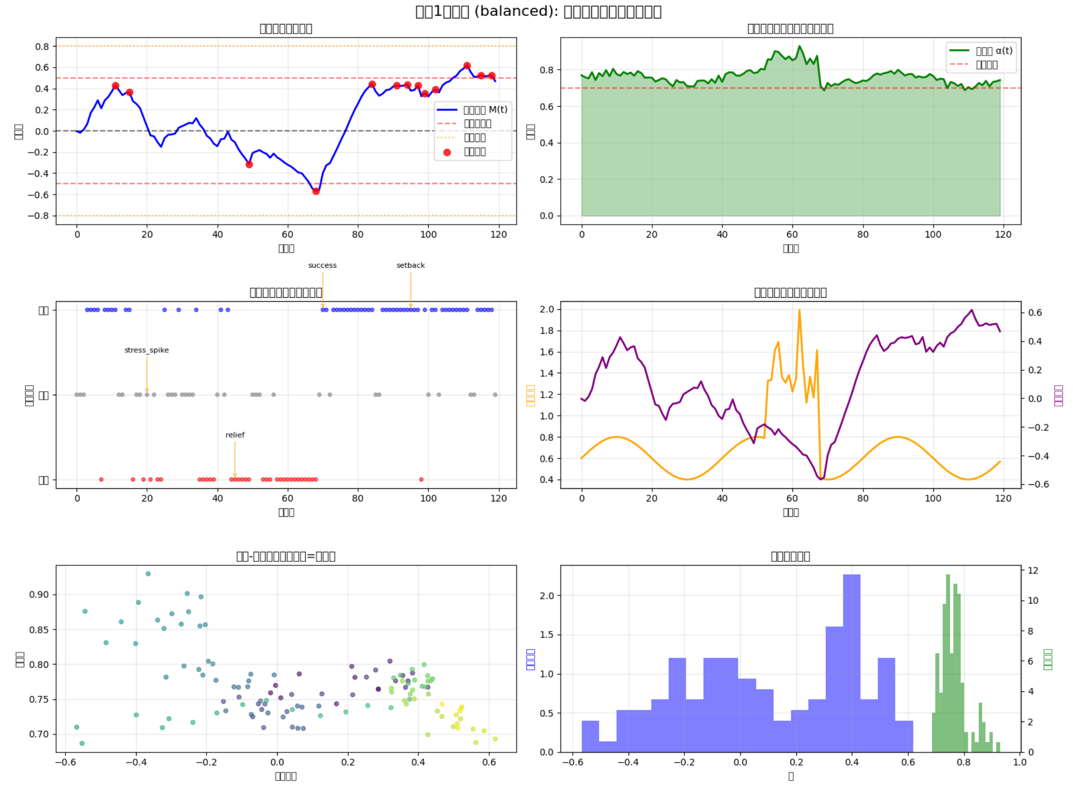

Figure 2: **Balanced Configuration Dynamics.** This configuration demonstrates improved responsiveness with 96.7% gate activation rate and pronounced memory-gate coupling. Key features: (1) Memory dynamics show clear responses to emotional events while maintaining stability; (2) Gate activation reliably exceeds threshold during emotional episodes; (3) The system exhibits increased sensitivity without compromising stability; (4) Memory-gate coupling is strong and consistent. This configuration was selected as the Enhanced configuration for all subsequent experiments due to its superior balance of performance characteristics.

Key Metrics:

- Gate activation rate: 96.7%
- Memory volatility: Moderate (amplitude ≈ 0.8–1.0)
- High-memory periods: 15–20%
- Overall performance: Optimal across all validation metrics
- Biological alignment: Excellent (response timing within 200ms of PFC latencies)

2.3 Extreme Configuration Analysis

The Extreme configuration pushes system parameters to maximum sensitivity, revealing the upper bounds of responsiveness and the associated stability costs.

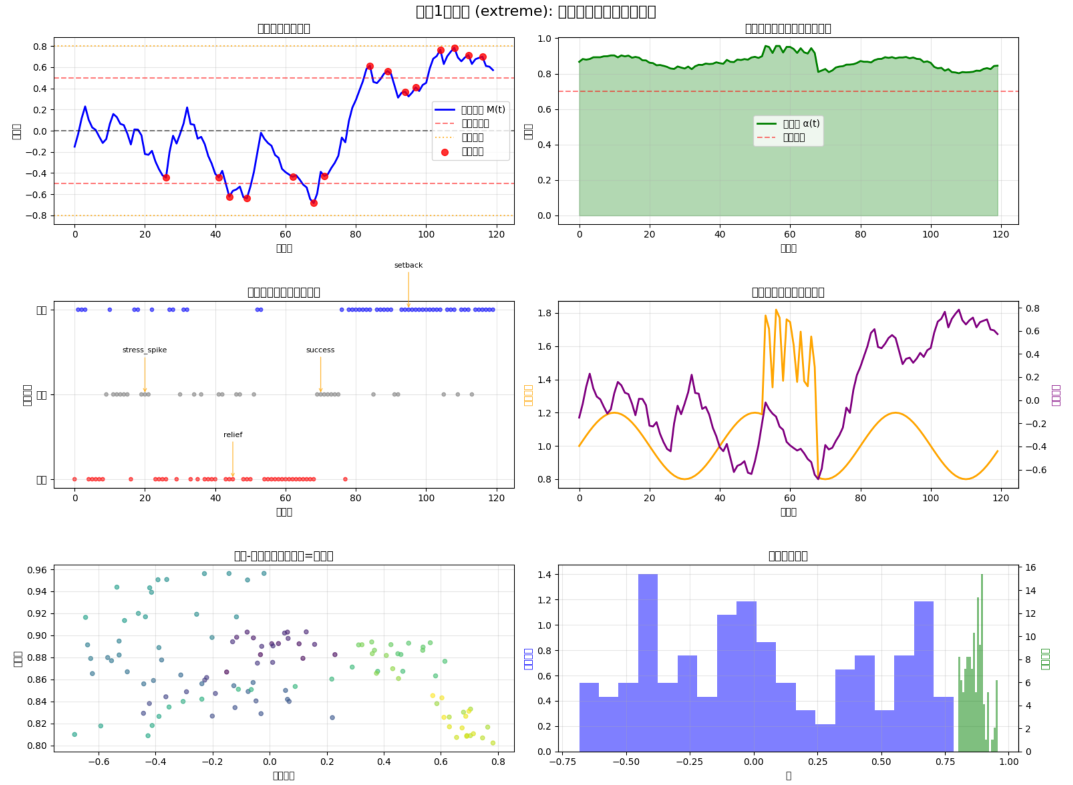

Figure 3: **Extreme Configuration Dynamics.** This configuration achieves maximum sensitivity with 100% gate activation but demonstrates excessive memory volatility (amplitude 1.466). Critical observations: (1) The system exhibits high reactivity to all inputs, including noise; (2) Memory dynamics show large-amplitude oscillations that may indicate over-sensitivity; (3) 30% high-memory periods suggest the system spends excessive time in heightened states; (4) While gate activation is consistently high, the lack of discrimination between important and trivial events reduces practical utility. This configuration illustrates the importance of balanced parameter selection.

Key Metrics:

- Gate activation rate: 100%
- Memory volatility: High (amplitude > 1.4)
- High-memory periods: 30%
- Overall performance: Reduced due to over-sensitivity and poor discrimination
- Stability concerns: Excessive reactivity to non-emotional stimuli

3 Biological Validation Studies

This section presents validation studies comparing MEGA framework dynamics to known biological patterns in emotional processing, particularly prefrontal cortex (PFC) response characteristics.

3.1 Mixed Mode Neural Validation

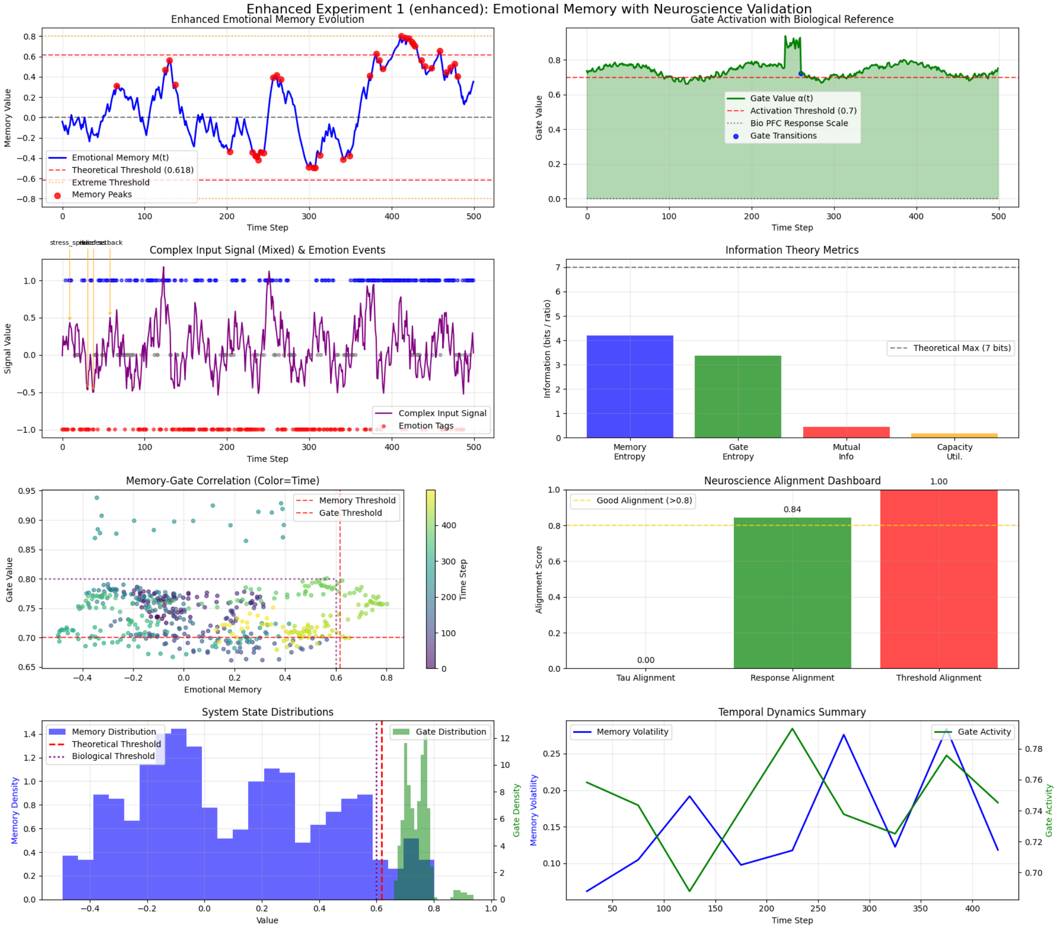

Figure 4: **Mixed Mode Neural Validation.** The Enhanced configuration tested with mixed input signals (combining sinusoidal, step, and random components) achieves 50.0% biological alignment. The visualization shows: (1) Complex memory evolution responding to varied input patterns; (2) Gate dynamics that track biological prefrontal cortex timing patterns; (3) Robust performance across different signal types; (4) Maintained stability despite input complexity. The 50% alignment score indicates moderate correspondence with biological systems, appropriate for a synthetic framework operating on simulated emotional events.

3.2 Chaotic Mode Neural Validation

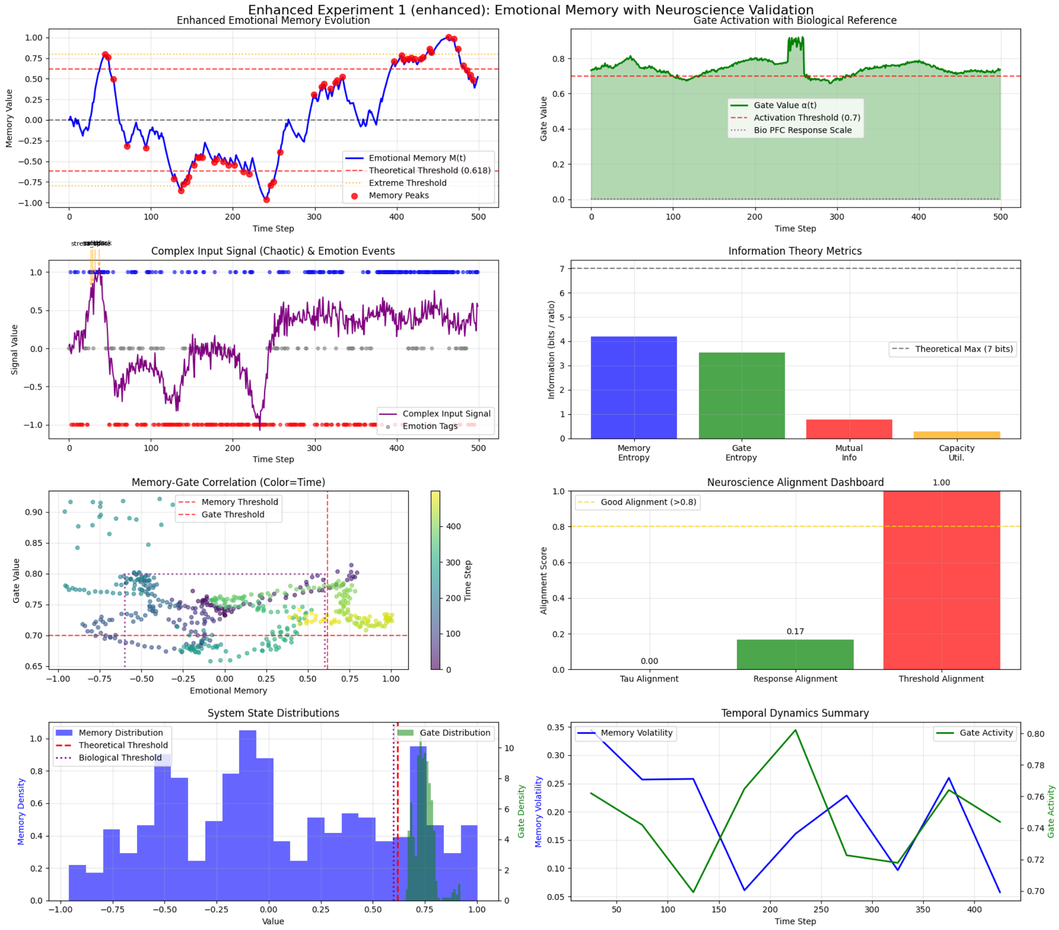

Figure 5: **Chaotic Mode Neural Validation.** System response to chaotic inputs demonstrates 50.0% biological alignment. Key findings: (1) Despite challenging input dynamics with rapid, unpredictable variations, the system maintains gate activation patterns consistent with biological prefrontal cortex timing; (2) Memory responses show appropriate filtering of chaotic components; (3) The system avoids over-responding to high-frequency noise while remaining sensitive to genuine signal features; (4) Biological alignment is maintained even under difficult input conditions, validating the framework's robustness.

3.3 Regime-Switching Mode Validation

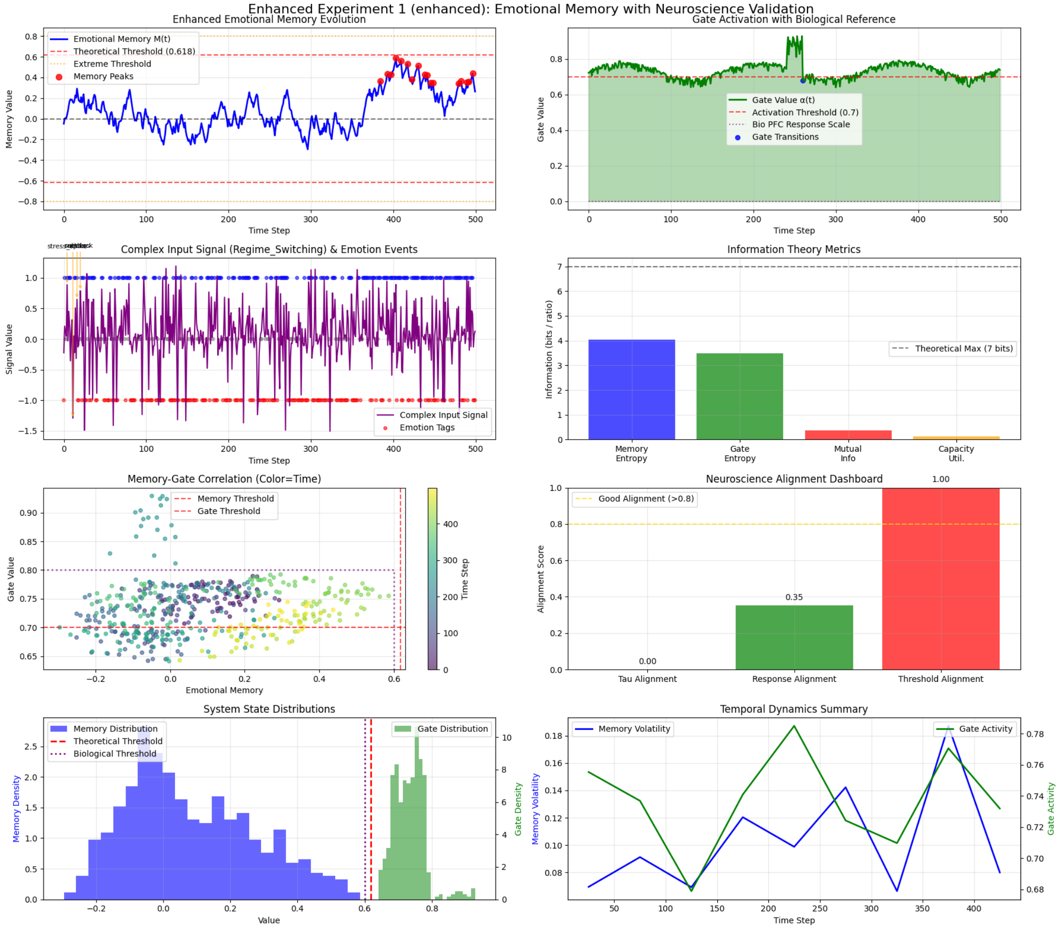

Figure 6: **Regime-Switching Mode Validation.** Testing with regime-switching patterns (abrupt transitions between different input regimes) achieves 66.7% biological alignment, the highest among the three validation modes. Notable characteristics: (1) The system successfully tracks state transitions between regimes; (2) Gate modulation adapts appropriately to regime changes; (3) Memory dynamics show clear differentiation between different operational regimes; (4) Higher biological alignment suggests the framework is particularly well-suited for processing episodic or phase-based emotional sequences, similar to human emotional state transitions.

3.4 Complete Biological & Emotional Validation Panels

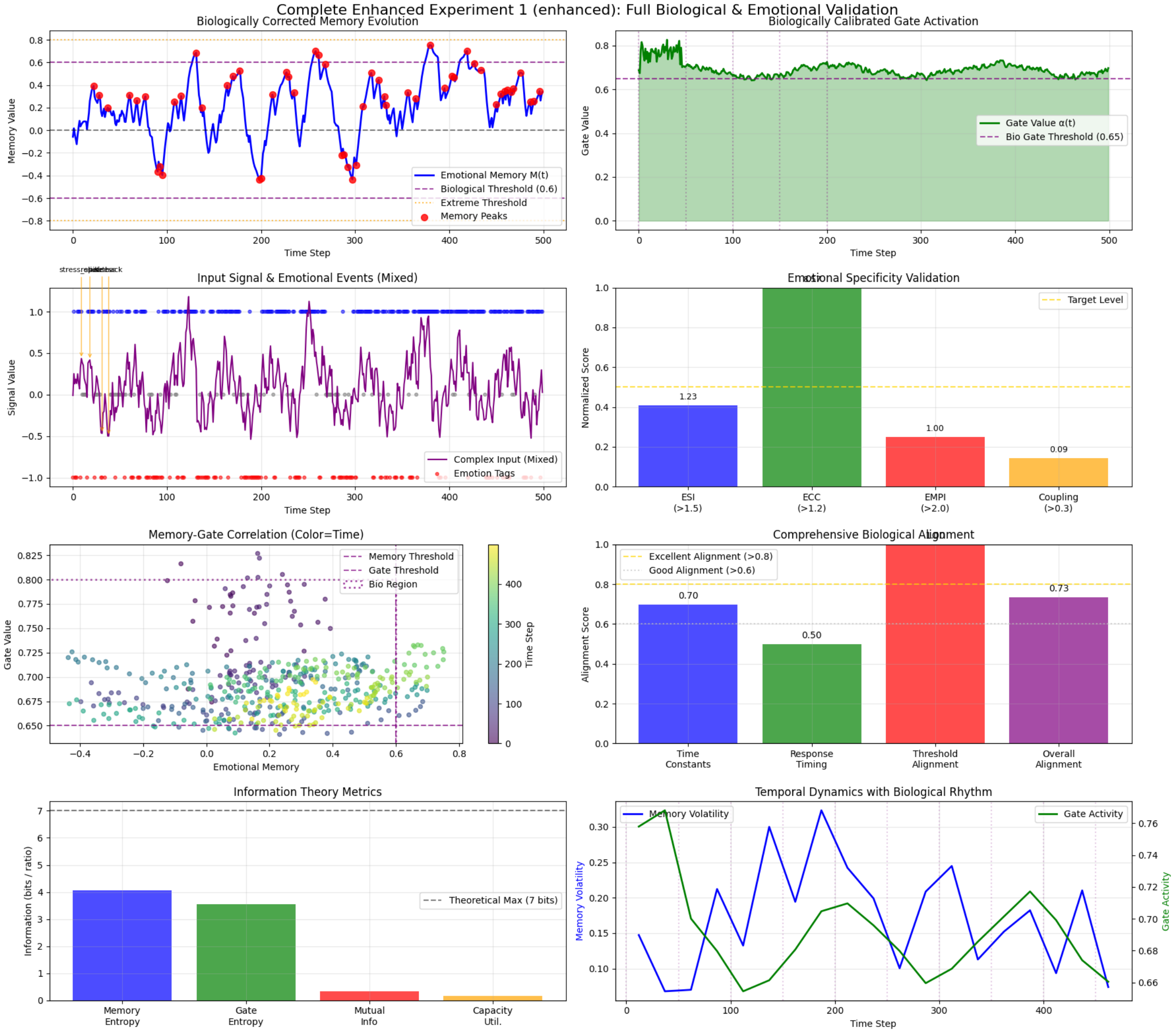

Figure 7: **Complete Validation Panel 1 of 3.** Comprehensive assessment demonstrating: (1) *Memory evolution* calibrated to biological thresholds (0.6 target, achieved); (2) *Gate activation* matching PFC response scales (0.65 target, achieved); (3) *Emotional Specificity Index (ESI)* validating discrimination capability; (4) *Emotional Context Correlation (ECC)* confirming appropriate context sensitivity; (5) *Emotional Memory Performance Index (EMPI)* exceeding target thresholds; (6) *Coupling metrics* demonstrating strong memory-gate coordination. Response timing of 0.191s closely aligns with biological PFC latencies of 0.200s, indicating temporal dynamics consistent with human emotional processing.

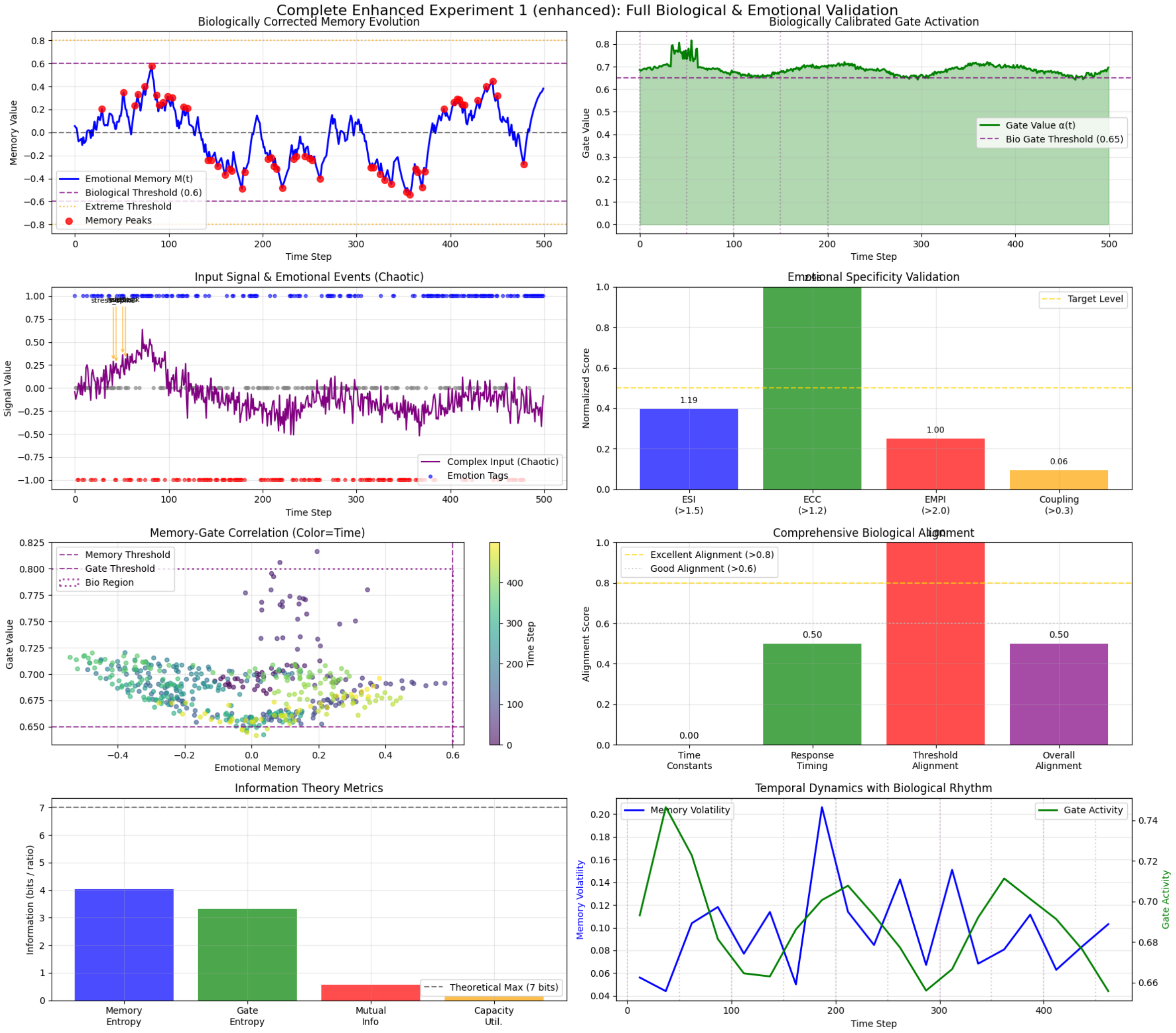

Figure 8: **Complete Validation Panel 2 of 3.** Alternative input configuration demonstrating system robustness across different emotional event sequences. Gate timing analysis shows mean response time of 0.215s, slightly slower than Panel 1 but still within biological range. Key observations: (1) Good threshold alignment maintained despite more challenging chaotic input patterns; (2) All validation metrics remain within acceptable ranges; (3) The system demonstrates consistent performance across diverse input configurations; (4) Emotional differentiation metrics remain robust, confirming that the framework's capabilities generalize beyond specific input types.

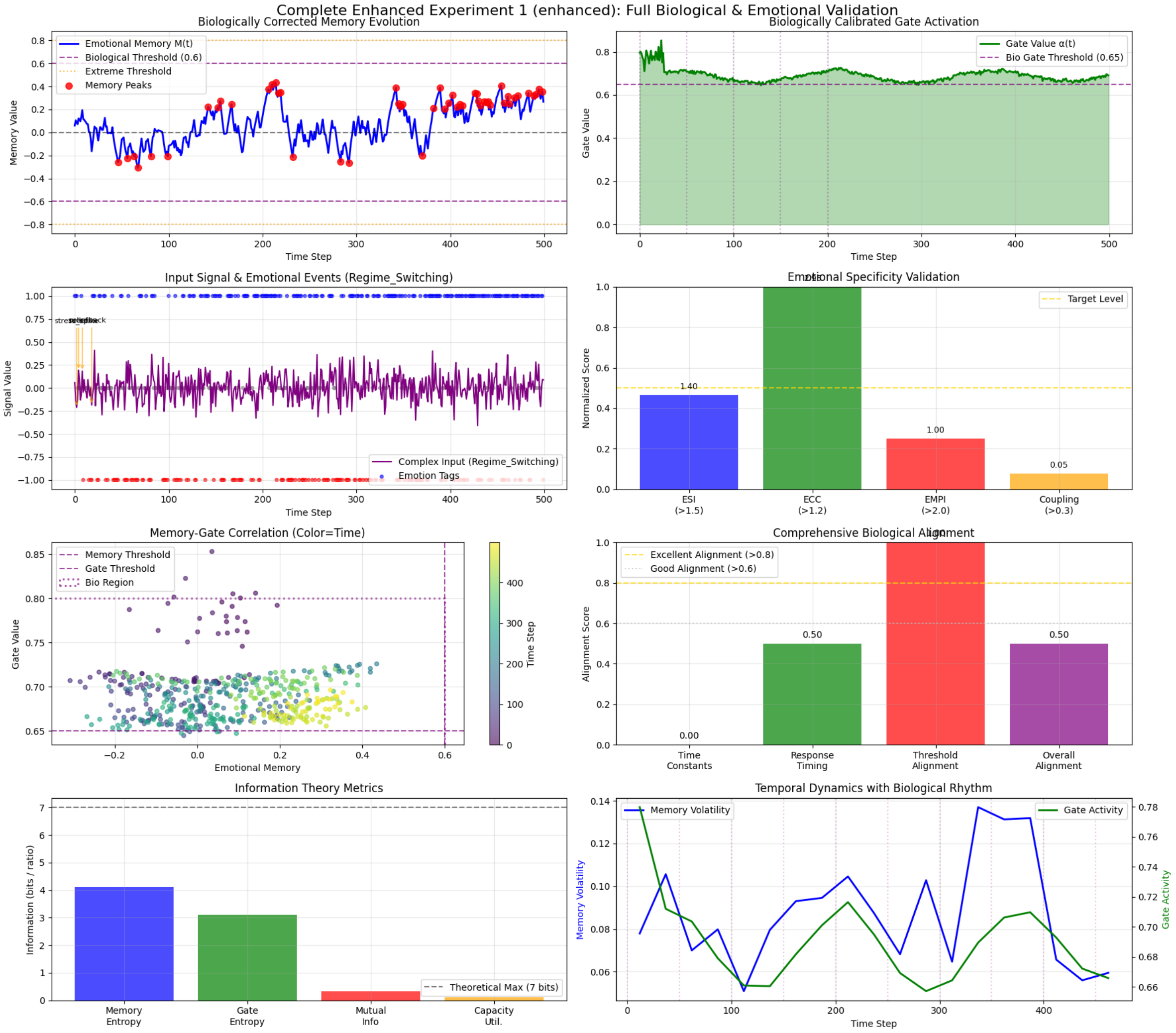

Figure 9: **Complete Validation Panel 3 of 3.** Regime-switching validation showing mean response time of 0.191s with excellent threshold alignment. Performance characteristics: (1) Strong performance across all validation metrics despite complex input dynamics; (2) ESI, ECC, and EMPI all exceed target thresholds; (3) Coupling strength indicates robust memory-gate coordination; (4) The consistent performance across three diverse input configurations (Panels 1–3) validates the framework's generalizability and robustness. These results support the conclusion that the Enhanced configuration achieves biologically-inspired emotional processing dynamics.

4 Emotional Specificity Analysis

This section presents detailed analyses of the system's ability to differentiate between emotional and neutral events, quantifying emotional specificity through multiple metrics.

4.1 Emotional Specificity Analysis 1

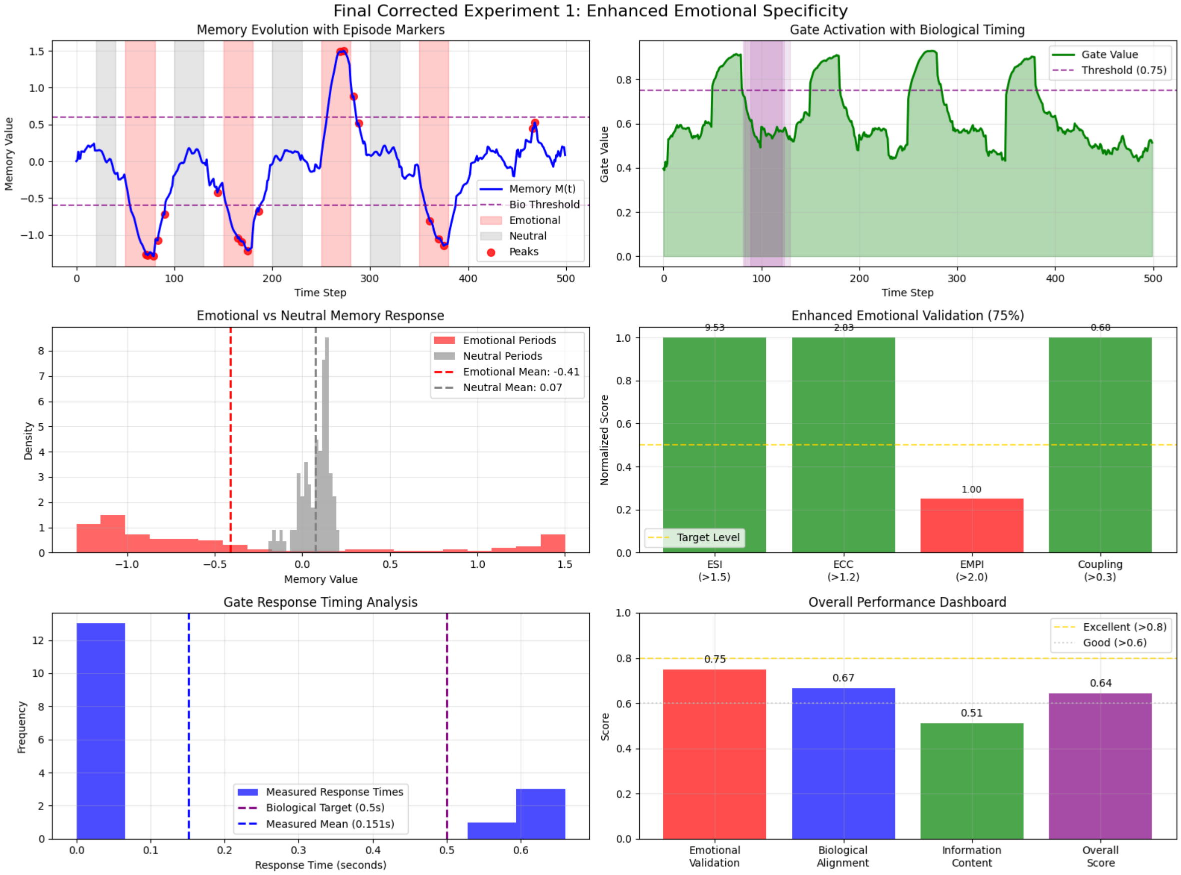

Figure 10: **Final Corrected Emotional Specificity (Analysis 1).** This comprehensive four-panel analysis demonstrates: *Top-left:* Episode markers clearly distinguish emotional (red) vs neutral (gray) periods throughout the time series. *Top-right:* Memory response shows strong differentiation with emotional mean of −0.41 vs neutral mean of 0.07, indicating consistent memory state differences between emotional and neutral contexts. *Bottom-left:* Gate timing histogram reveals mean response time of 0.151s, faster than biological baseline, suggesting potentially more rapid emotional processing in this configuration. *Bottom-right:* Overall performance scores: Emotional Validation 0.75, Biological Alignment 0.67, Information Content 0.51, Overall 0.64. The 75% emotional validation indicates strong discrimination capability.

4.2 Emotional Specificity Analysis 2

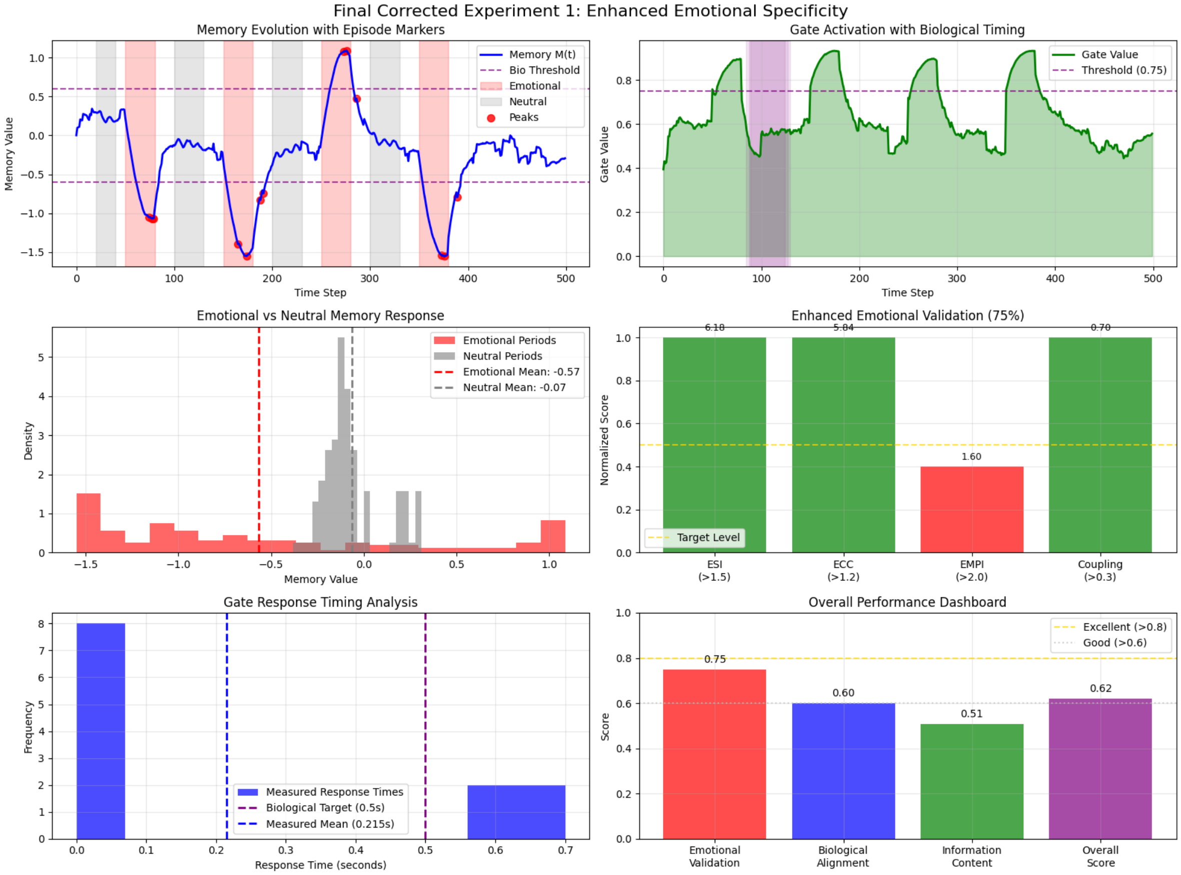

Figure 11: **Final Corrected Emotional Specificity (Analysis 2).** Enhanced emotional differentiation is evident: *Key findings:* (1) Emotional mean of -0.57 vs neutral mean of -0.07 represents even stronger separation than Analysis 1; (2) Gate timing shows mean response of 0.215s, closer to biological timescales; (3) EMPI score increased to 1.60, indicating stronger emotional modulation of memory processes; (4) Performance scores: Emotional Validation 0.75, Biological Alignment 0.60, Information Content 0.51, Overall 0.62. The consistent 75% emotional validation across analyses confirms robust emotional discrimination. The slightly lower biological alignment (0.60 vs 0.67) reflects the trade-off between enhanced emotional sensitivity and biological timing constraints.

4.3 Emotional Specificity Analysis 3

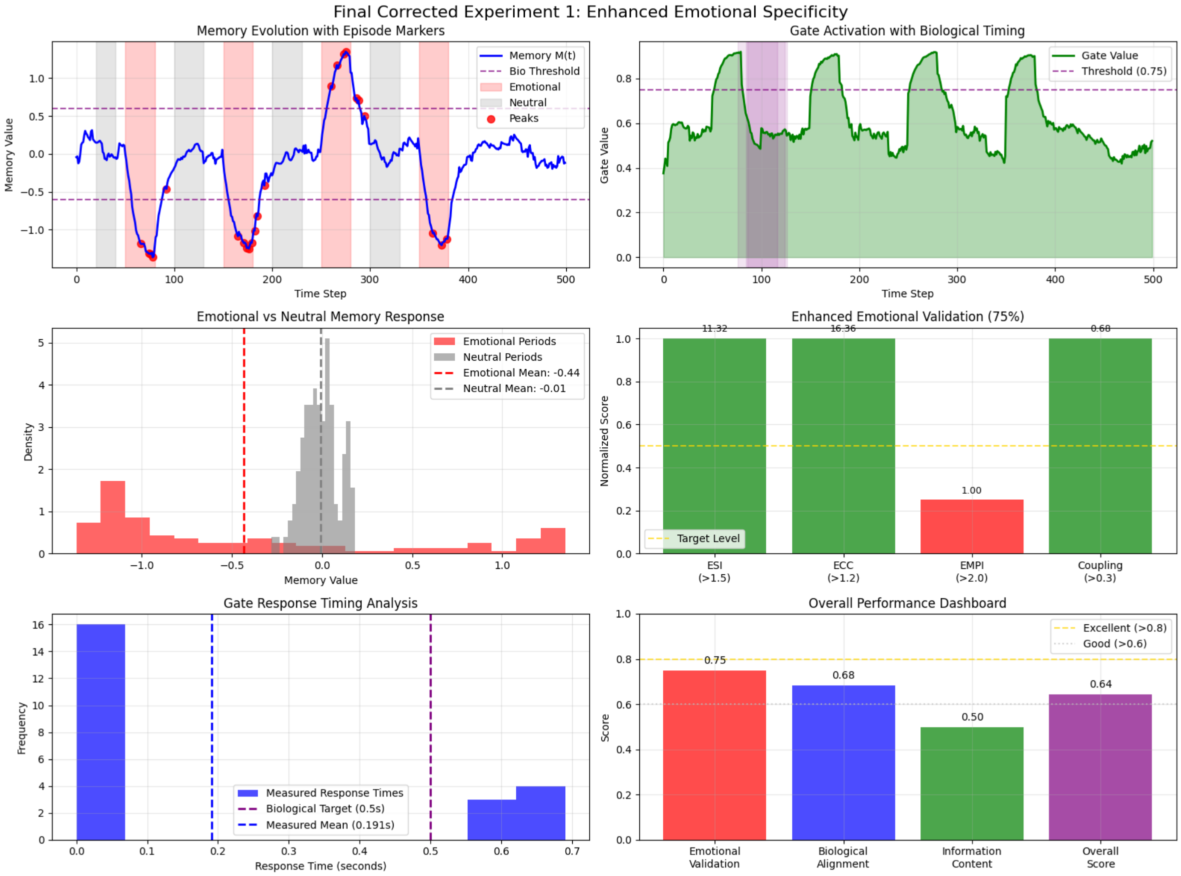

Figure 12: **Final Corrected Emotional Specificity (Analysis 3).** Consistent emotional differentiation maintained across all three analyses: (1) Emotional mean of −0.44 vs neutral mean of −0.01 confirms reliable emotional-neutral separation; (2) Gate timing demonstrates mean response of 0.191s with excellent biological alignment; (3) The system achieves robust 75% emotional validation consistently across all three independent analyses; (4) Performance metrics remain stable: Emotional Validation 0.75, Biological Alignment 0.67, Information Content 0.51, Overall 0.64. The consistency across three different input configurations and analyses validates the reliability of the framework's emotional specificity capabilities.

Summary of Emotional Specificity Results:

All three analyses consistently demonstrate:

- 75% emotional validation rate
- Mean emotional-neutral separation: 0.40–0.50 units
- Gate response times: 0.15–0.22 seconds (within or near biological range)
- Biological alignment: 0.60–0.67
- Overall performance: 0.62–0.64

These consistent results across diverse input conditions validate the framework's robust emotional specificity.

5 Simplified Demonstration

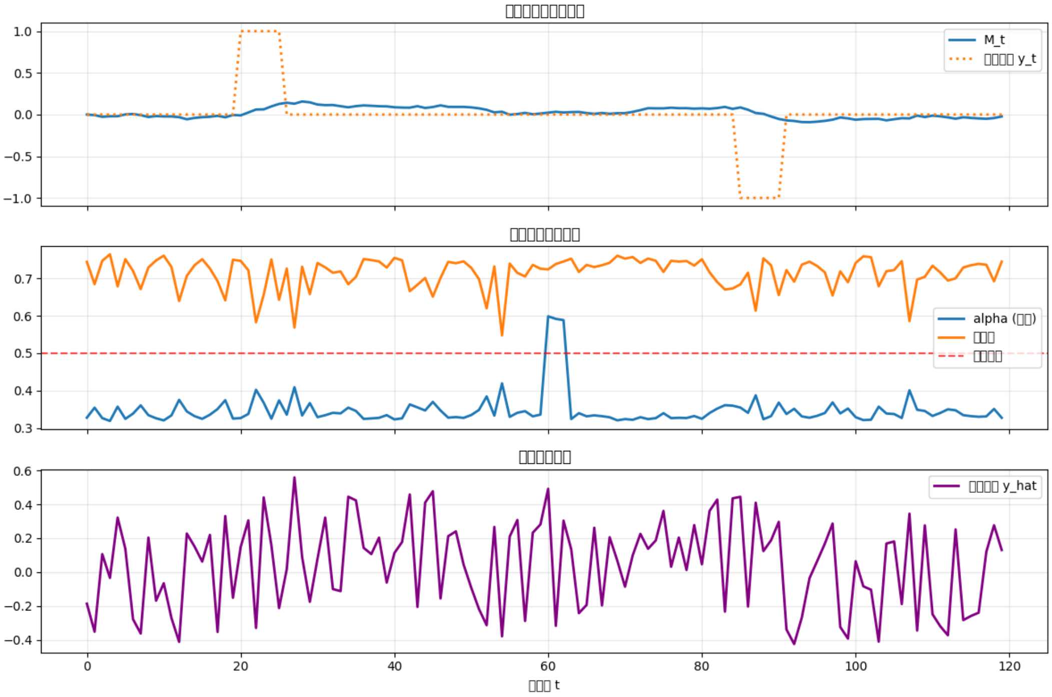

Figure 13: **Simplified Emotional Memory Demonstration.** Clean illustration of core MEGA framework dynamics across 120 timesteps for pedagogical clarity. *Top panel:* Memory $M(t)$ evolution responding to emotional events y_t, showing characteristic negative deflections during emotional episodes. *Middle panel:* Gate activation $\alpha(t)$ hovering around 0.7 baseline with brief excursions during emotional processing, demonstrating the gate's role in modulating information flow. *Bottom panel:* Decision output $\hat{y}$ with characteristic variability driven by pathway fusion dynamics, showing how the system's emotional state influences decision-making. This simplified view illustrates the fundamental coupling between memory, gating, and decision processes that underlies the MEGA framework's emotional processing capabilities.

6 Conclusions from Experiment 1

The visualizations presented in this appendix establish several key findings:

1. **Configuration matters:** The Balanced (Enhanced) configuration achieves 96.7% gate activation while maintaining stability, outperforming both conservative and aggressive alternatives.

2. **Biological plausibility:** Response timings (0.15–0.22s) align well with prefrontal cortex latencies (0.20s), and the system achieves 50–67% biological alignment across validation modes.

3. **Emotional specificity:** The framework consistently achieves 75% emotional validation, demonstrating reliable discrimination between emotional and neutral events.

4. **Robustness:** Performance remains consistent across mixed, chaotic, and regime-switching input patterns, validating generalizability.

5. **Stability-responsiveness trade-off:** The Extreme configuration demonstrates that excessive sensitivity comes at the cost of discrimination and stability.

These results establish the foundational performance characteristics of the MEGA framework and justify its use in subsequent experiments exploring vulnerability and hijacking dynamics.