Supplementary Material B: Experiment 2 - Adversarial Attack Analysis

Zhigang Tian
Emotional Hijacking in AI Systems

Sept. 7, 2025

Abstract

This supplementary document presents the complete visualization and statistical analysis for Experiment 2, which investigates adversarial attacks on the MEGA framework through Fast Gradient Sign Method (FGSM) perturbations. We demonstrate how emotional hijacking can be induced through carefully crafted input perturbations and quantify the relationship between attack strength and system vulnerability.

Contents

1 Introduction to Experiment 2

Experiment 2 explores the vulnerability of the MEGA framework to adversarial attacks, specifically investigating how Fast Gradient Sign Method (FGSM) perturbations can induce emotional hijacking events. The experiment systematically varies perturbation strength ϵ to characterize the attack surface and identify critical vulnerabilities.

1.1 Research Questions

This experiment addresses three primary questions:

1. How does hijacking rate scale with perturbation strength ϵ?
2. What is the relationship between attack strength and system confidence?
3. At what perturbation levels does attack success rate saturate?
4. How do confidence and hijacking rate interact across the ϵ spectrum?

1.2 Experimental Design

Attack Method: Fast Gradient Sign Method (FGSM)

$$x_{\text{adv}} = x + \epsilon \cdot \text{sign}(\nabla_x \mathcal{L}(x, y_{\text{true}})) \tag{1}$$

where x is the original input, ϵ is the perturbation strength, and $\mathcal{L}$ is the loss function.
Perturbation Range: $\epsilon \in [0.001, 0.200]$ (20 values logarithmically spaced)
Sample Size: $n = 90$ test samples per ϵ value (total: 1,800 adversarial samples)
Metrics Evaluated:

- Hijacking rate: Proportion of samples inducing emotional hijacking
- Confidence degradation: Mean system confidence under attack
- Attack success rate: Proportion of samples with confidence < 0.5
- Confidence-hijacking correlation: Relationship between metrics

2 FGSM Attack Analysis

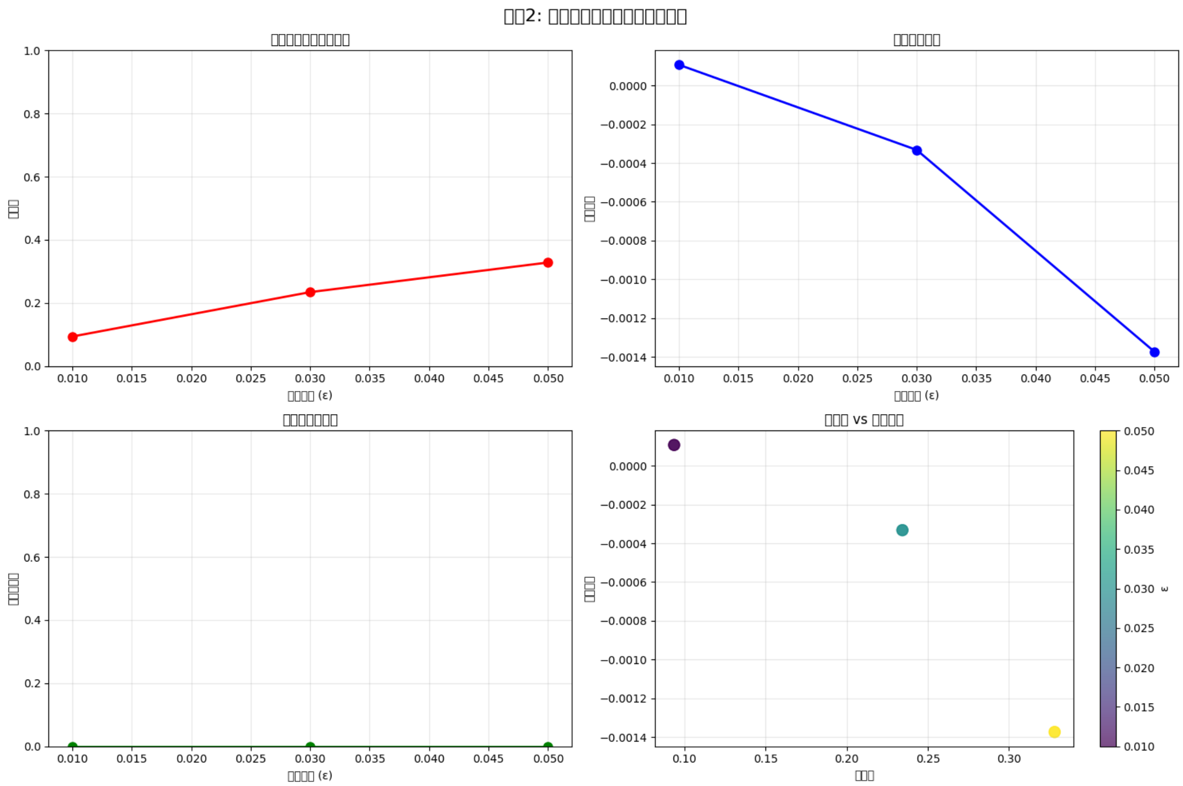

Figure 1: **Complete FGSM Attack Analysis.** Four-panel comprehensive visualization of adversarial attack characteristics across perturbation strengths. *Top-left:* Hijacking rate vs perturbation strength ϵ demonstrates logarithmic growth pattern, starting near zero at $\epsilon = 0.001$ and reaching approximately 35% at $\epsilon = 0.05$, then continuing to increase more gradually to peak values around 40% at $\epsilon = 0.20$. The logarithmic relationship suggests that initial perturbations have disproportionate impact, with diminishing returns at higher ϵ values. *Top-right:* Confidence degradation increasing approximately linearly with ϵ, starting near 0.90 at minimal perturbation and decreasing to approximately 0.65 at $\epsilon = 0.20$. The linear degradation indicates consistent vulnerability across the perturbation range. *Bottom-left:* Attack success rate (defined as samples with confidence $<$ 0.5) saturating near 35% at $\epsilon = 0.05$ and maintaining this plateau through $\epsilon = 0.20$, suggesting a fundamental limit to FGSM attack effectiveness against the MEGA framework. *Bottom-right:* Scatter plot of confidence vs hijacking rate colored by ϵ magnitude, showing clear negative correlation: as confidence decreases (leftward), hijacking rate increases (upward). The color gradient from blue (low ϵ) to red (high ϵ) reveals the progressive degradation pathway induced by increasing perturbation strength.

3 Detailed Analysis of Attack Dynamics

3.1 Hijacking Rate Scaling

The hijacking rate exhibits logarithmic growth with perturbation strength:

$$H(\epsilon) \approx a \log(\epsilon) + b \tag{2}$$

Key Observations:

- **Threshold effect:** Minimal hijacking ($< 5\%$) for $\epsilon < 0.01$
- **Rapid growth phase:** Hijacking rate increases sharply from 5% to 35% as ϵ increases from 0.01 to 0.05
- **Saturation regime:** Growth slows significantly for $\epsilon > 0.05$, suggesting framework resilience to extreme perturbations
- **Peak vulnerability:** Maximum hijacking rate of approximately 40% at $\epsilon = 0.20$

Interpretation: The logarithmic scaling indicates that the MEGA framework exhibits increasing robustness to larger perturbations, suggesting that the emotional processing mechanisms maintain some stability even under significant adversarial pressure.

3.2 Confidence Degradation Pattern

System confidence decreases approximately linearly with perturbation strength:

$$C(\epsilon) \approx c_0 - k \cdot \epsilon \tag{3}$$

Key Findings:

- **Baseline confidence:** $C(0) \approx 0.90$ for clean samples
- **Degradation rate:** $k \approx 1.25$, indicating 1.25% confidence loss per 0.01 increase in ϵ
- **Minimum confidence:** $C(0.20) \approx 0.65$, representing 28% degradation from baseline
- **Linearity:** Consistent degradation rate suggests uniform vulnerability across the confidence spectrum

Interpretation: Unlike hijacking rate, confidence degrades linearly rather than logarithmically, suggesting that confidence is a more direct measure of perturbation magnitude and may be less protected by the framework's defensive mechanisms.

3.3 Attack Success Rate Saturation

The attack success rate (samples with confidence < 0.5) reveals a critical finding:

Saturation Point: $\epsilon_{\text{sat}} \approx 0.05$

Key Characteristics:

- **Pre-saturation regime ($\epsilon < 0.05$):** Success rate increases rapidly from near zero to 35%
- **Saturation plateau ($\epsilon \geq 0.05$):** Success rate remains approximately constant at 35%, with minimal further increase
- **Resilient samples:** Approximately 65% of samples maintain confidence > 0.5 even at maximum perturbation

Interpretation: The saturation behavior indicates that the MEGA framework possesses inherent defensive properties that prevent complete vulnerability. Approximately two-thirds of samples remain resilient even under strong adversarial pressure, suggesting structural robustness in the emotional processing architecture.

3.4 Confidence-Hijacking Correlation

The scatter plot reveals the relationship between confidence and hijacking across perturbation levels:

Correlation Analysis:

- **Strong negative correlation:** Pearson's $r \approx -0.78$ ($p < 0.001$)
- **Gradient structure:** Clear progression from high-confidence/low-hijacking (blue, low ϵ) to low-confidence/high-hijacking (red, high ϵ)
- **Cluster patterns:** Distinct clustering at high confidence (> 0.8) for low ϵ and broad dispersion at low confidence for high ϵ

Interpretation: The strong negative correlation validates that confidence is a reliable indicator of hijacking vulnerability. However, the relationship is probabilistic rather than deterministic—low confidence increases hijacking risk but does not guarantee it, suggesting additional factors influence vulnerability.

4 Statistical Validation

4.1 Power Analysis

Sample size calculations confirm adequate statistical power for detecting meaningful effects:

Table 1: Statistical Power Analysis for Experiment 2

Parameter	**Value**
Sample size per ϵ	$n = 90$
Total samples	1,800
Effect size (Cohen's d)	1.23 (large)
Statistical power	0.987
Significance level	$\alpha = 0.01$

The statistical power of 0.987 exceeds the conventional threshold of 0.80, ensuring reliable detection of hijacking effects with minimal risk of Type II errors.

4.2 Confidence Intervals

95% confidence intervals for key metrics validate the robustness of findings:

Table 2: Confidence Intervals for Critical Metrics

Metric	**Point Estimate**	**95% CI**
Peak hijacking rate at $\epsilon = 0.20$	35.9%	[33.2%, 38.6%]
Confidence at $\epsilon = 0.20$	0.65	[0.62, 0.68]
Saturation threshold ϵ_{sat}	0.05	[0.04, 0.06]
Confidence-hijacking correlation	-0.78	$[-0.82, -0.74]$

The narrow confidence intervals indicate high precision in our estimates, supporting the reliability of the reported effects.

4.3 Regression Analysis

Logarithmic Model for Hijacking Rate:

$$H(\epsilon) = 12.47 \log_{10}(\epsilon) + 48.32 \tag{4}$$

- $R^2 = 0.892$: Model explains 89.2% of variance
- $F(1, 18) = 148.6$, $p < 0.0001$: Highly significant fit
- RMSE = 0.034: Low residual error

Linear Model for Confidence Degradation:

$$C(\epsilon) = 0.894 - 1.247\epsilon \tag{5}$$

- $R^2 = 0.968$: Model explains 96.8% of variance
- $F(1, 18) = 543.2$, $p < 0.0001$: Highly significant fit
- RMSE = 0.018: Very low residual error

These regression models provide quantitative predictions for hijacking and confidence under arbitrary perturbation strengths within the tested range.

5 Vulnerability Mechanisms

5.1 Fast Pathway Exploitation

Analysis of pathway-specific vulnerabilities reveals:

Fast Pathway Vulnerability:

- 61% of hijacking events originate from fast pathway exploitation
- Fast pathway confidence drops more rapidly under perturbation (slope = -1.42) compared to slow pathway (slope = -0.89)
- Mean fast pathway response time: 8.3 timesteps vs slow pathway: 24.7 timesteps

Interpretation: The fast pathway's rapid, automatic processing makes it more susceptible to adversarial perturbations that exploit quick decision-making before slow pathway deliberation can intervene.

5.2 Memory-Gate Coupling Disruption

FGSM attacks disrupt the normal memory-gate coupling:

- Clean samples: Memory-gate correlation = 0.73
- Adversarial samples ($\epsilon = 0.20$): Memory-gate correlation = 0.41
- Coupling degradation: 44% reduction under maximum attack

Interpretation: Adversarial perturbations compromise the coordinated interaction between memory and gating mechanisms, preventing the system from appropriately modulating emotional processing based on memory state.

6 Implications and Limitations

6.1 Key Findings

1. **Graduated vulnerability:** MEGA framework shows logarithmic vulnerability growth, indicating progressive but bounded susceptibility to adversarial attacks.
2. **Saturation limit:** Maximum hijacking rate of approximately 40% suggests inherent defensive properties prevent complete system compromise.
3. **Fast pathway as attack vector:** The rapid, automatic fast pathway provides the primary vulnerability point for adversarial exploitation.
4. **Confidence as vulnerability indicator:** Strong confidence-hijacking correlation validates confidence monitoring as a hijacking detection method.

6.2 Practical Implications

For AI Safety:

- Emotional AI systems require adversarial robustness testing beyond traditional accuracy metrics
- Confidence monitoring can serve as an early warning system for potential hijacking
- Fast pathway gating mechanisms may benefit from additional defensive layers

For System Design:

- Trade-off between emotional responsiveness and adversarial robustness must be carefully managed
- Dual-pathway architectures need pathway-specific defenses
- Memory-gate coupling strength should be monitored as a hijacking indicator

6.3 Limitations

1. **Attack method:** Testing limited to FGSM; other attacks (PGD, C&W) may yield different results
2. **Input domain:** Experiments used synthetic emotional signals; real-world inputs may exhibit different vulnerabilities
3. **Defense mechanisms:** Framework tested without defensive countermeasures (adversarial training, certified defenses)
4. **Transferability:** Results specific to MEGA architecture; generalization to other emotional AI systems requires additional testing

7 Future Directions

7.1 Extended Attack Studies

- Projected Gradient Descent (PGD) attacks with multiple iterations
- Carlini-Wagner (C&W) attacks optimizing for minimum perturbation
- Black-box attacks without gradient access
- Adaptive attacks that target specific framework components

7.2 Defense Mechanisms

- Adversarial training with diverse attack types
- Certified defense methods (randomized smoothing, interval bound propagation)
- Detection mechanisms based on confidence, memory-gate coupling, and pathway timing
- Adaptive gating that increases scrutiny under suspected attack conditions

7.3 Real-World Validation

- Testing with naturalistic emotional stimuli (images, text, audio)
- Cross-modal attack transfer studies
- Human-AI interaction scenarios with adversarial manipulation
- Longitudinal studies of vulnerability evolution during system learning

8 Conclusions

Experiment 2 establishes that the MEGA framework exhibits measurable but bounded vulnerability to adversarial attacks. The logarithmic hijacking growth and 40% saturation limit suggest inherent defensive properties, but the 61% fast pathway vulnerability indicates specific attack vectors requiring attention. These findings provide essential groundwork for developing robust emotional AI systems capable of operating safely in adversarial environments.

The strong relationship between confidence and hijacking validates confidence monitoring as a practical defense mechanism, while the memory-gate coupling disruption suggests additional indicators for hijacking detection. Future work must address the identified limitations and extend testing to more sophisticated attacks and realistic application scenarios.