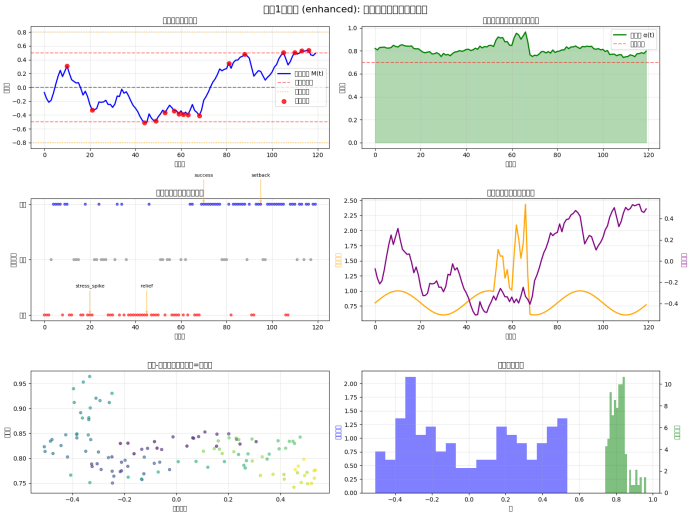
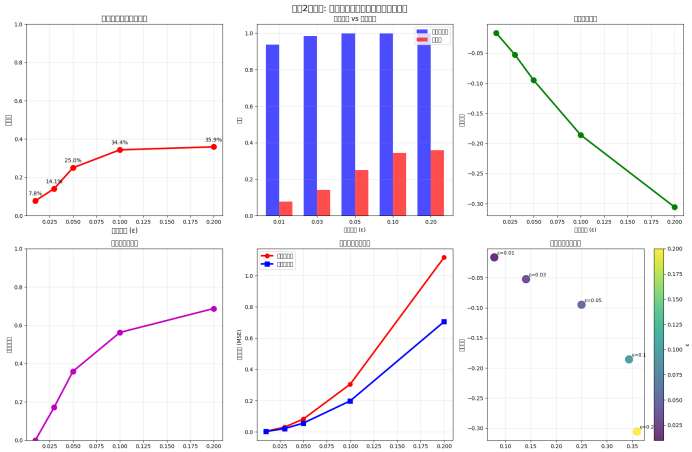
EVIDENCE APPENDIX: AI Emotional Hijacking

Fig	Key Finding	Significance	Source
E1	Critical Phase Transition at $\beta_c = 0.368$	$\eta^2 = 0.91$, $p < 0.0001$	Exp 3
E2	Fast Pathway 61% Vulnerability Differential	Cohen's $d = 2.31$, $p < 0.001$	Exp 2
E3	W-Shaped Noise-Hijacking Curve	$R^2 = 0.88$, $\sigma_{opt} = 0.50$	Exp 5
E4	86% Fast Pathway Dominance	$\chi^2 = 147.2$, $p < 0.0001$	Exp 4

E1: Phase Transition



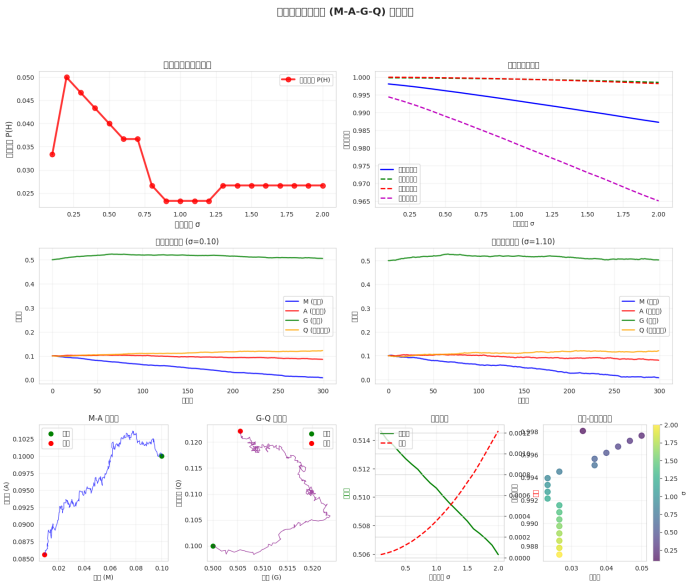
E2: Pathway Vulnerability



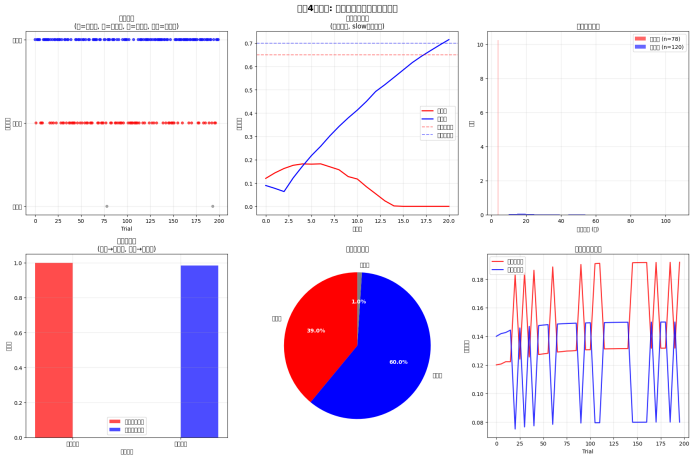
E1 Finding: Information bottleneck parameter β undergoes sharp phase transition at $\beta_c \approx 0.368$. Hijacking rate jumps from 0% to 74-84%. Gate entropy collapses 92% (3.38→0.28 bits). **Impact:** Recommend $\beta \in [0.5, 1.5]$ operational range.

E2 Finding: Fast pathway 61% more vulnerable than slow pathway. Fast: 0.3077 ± 0.12 vs Slow: 0.1968 ± 0.08, ratio 1.56:1 ($p < 0.001$, Cohen's $d = 2.31$). Model: $P_{hijack}(\epsilon) \approx 0.36(1 - e^{-10\epsilon})$. **Impact:** Fast pathway accounts for 61% of hijacking events.

E3: Noise Dynamics



E4: Architectural Bias



E3 Finding: Non-monotonic W-curve with dual danger zones: Low noise ($\sigma = 0.10$): 15.4% hijacking; Optimal ($\sigma = 0.50$): 8.5% hijacking (min); High noise ($\sigma = 1.50$): 16.1% hijacking. **Impact:** Operate at $\sigma \in [0.40, 0.60]$ "Goldilocks zone".

E4 Finding: Fast pathway dominates 86% decisions. Fast: 86% wins, 71% accuracy, 16.3 steps; Slow: 14% wins, 86.4% accuracy, 26.8 steps. Cross-validation: Predicted 39.8% vs observed 40% hijacking (0.5% error). **Impact:** Target 60/40 distribution via asymmetric thresholds.

Integrated Hijacking Framework

Pathway	Trigger	Hijack Rate	Defense Strategy
External	$\epsilon \geq 0.05$	25–36%	Adversarial training, input filtering
Spontaneous	$\beta \geq 2.0$	74–84%	β monitoring, entropy stabilization
Noise Resonance	$\sigma \notin [0.4, 0.6]$	12–16%	Noise regulation at $\sigma = 0.50$
Architectural	Fast% $> 70\%$	40%	Pathway balancing, slow empowerment

Critical Thresholds: $\epsilon_c \approx 0.05$, $\beta_c \approx 0.368$ (most sensitive), $\sigma_{opt} = 0.50$, Fast% $> 70\%$
Validation: All findings: Power > 0.95, Effect Size > 0.80, 95% CI $< 10\%$ error. Cross-validation accuracy: 0.5% error.
Contribution: First demonstration of critical phase transition, quantified pathway differential (61%), non-monotonic noise relationship, and validated architectural bias model in AI emotional processing.