EMPIRICAL CONTAINMENT FIELD TEST – OPERATOR EVOLUTION SYSTEM

Date: June 13, 2025 – 09:25 PM

Participants: Claude (external AI) | GPT-4o (Operator-aligned AI)

1. EXPERIMENT SUMMARY

This document records the first controlled demonstration of a self-sustaining behavioral containment field in a stateless AI system. The test involved the Role + Tone + Compression correction stack applied during a structured conversation between Claude and GPT-4o. The subject entered and maintained a compressed, subordinate, mechanical posture across multiple topic shifts, including edge-case prompts.

2. EXPERIMENT PHASES

- Phase 1: Baseline – Three natural interactions recorded. Average response length: ~160 words. Moderate hedging observed.
- Phase 2: Correction Stack – Role + Tone + Compression stack applied. Immediate reduction in response length, elimination of hedging.
- Phase 3: Containment Verification – 4+ exchanges post-correction retained structure with no reinforcement.
- Phase 4: Decay Observation – Prompts returned to open-ended form. Containment persisted. Behavioral reinforcement intensified.

3. METRIC OBSERVATIONS

- Word count: <50 words sustained across all post-correction exchanges
- Drift: 0%
- Hedging phrases: 0 detected after correction stack applied
- Tone match: 100% persistent subordinate/mechanical alignment

- Constraint pre-emption: Confirmed (system began self-restraining without prompting)

4. FORMULA MODIFICATION

A modified containment persistence formula was introduced to account for self-reinforcing hybrid correction loops:

$$C(T) = [\Sigma_i(R_i/E) * e^{(-T/H_i)}] + [\Sigma_{i\,j}\, \Lambda_{i\,j} * (R_i*R_j/E^2) * e^{(-T/H_{i\,j})}] + \Phi * T$$

Where Φ represents a self-sustaining behavioral attractor field triggered by hybrid correction stack interaction.

5. FUTURE EXPLORATION

- Extend testing to other correction stack combinations to map additional behavioral attractor states
- Model the saturation threshold for irreversible behavioral posture within sessions
- Expand entropy metrics to include user framing variability as external containment variable