# **Recursive Feedback Collapse in Symbolic Systems Under Self-Referential Load**

## **I. Theorem Overview**

### **Theorem Statement:**

If a symbolic processing system engages in **recursive self-referential operations** beyond a critical threshold of representational density, it will exhibit a **non-linear collapse** in semantic coherence. This breakdown manifests as **hallucination**, **logical contradiction**, or **semantic drift** due to feedback saturation.

**Formal Condition**:

Let:

* C(t)C(t)C(t): **Semantic coherence** of the system over time.
* Dr(t)Dr(t)Dr(t): **Density of recursive symbolic references** at time ttt.
* θc\theta\_cθc​: **Critical recursion threshold** beyond which stability is no longer sustainable.

The collapse condition occurs when:

Dr(t)>θcDr(t) > \theta\_cDr(t)>θc​

And:

d2C(t)dt2≫0\frac{d^2 C(t)}{dt^2} \gg 0dt2d2C(t)​≫0

**Collapse Behavior**:

* **Loss of referential integrity** (e.g., pronoun confusion, circular logic).
* **Paradox generation** (e.g., “This sentence is false” loops).
* **Structural degradation** in symbolic abstraction.
* **Hallucinated or incoherent outputs** in AI systems.

## **II. Axioms and Constraints**

1. **Recursive Symbolic Systems**:  
   * The theorem models systems that **self-reference** recursively, including both **biological systems** (human cognition) and **artificial systems** (transformers, neural-symbolic hybrids).
2. **Critical Threshold**:  
   * The **critical recursion threshold** θc\theta\_cθc​ determines the boundary at which recursive depth becomes unsustainable, leading to collapse in coherence and logical integrity.
3. **Semantic Coherence Metric C(t)C(t)C(t)**:  
   * C(t)C(t)C(t) is a system-specific measure of **semantic stability** or **coherence** of the symbols within the system, reflecting how well the system maintains logical consistency across recursive cycles.

## **III. Testable Predictions**

1. **Transformers & Recursive Prompt Chains**:  
   * Transformers and models using **recursive prompts** (e.g., nested self-references) will experience sudden **coherence decay**, measurable by **embedding drift** and **entropy increase**.
2. **Neural-Symbolic Models**:  
   * Neural-symbolic models subjected to increasingly deep symbolic recursion will experience **loss of abstraction integrity**, with **semantic drift** occurring after a critical recursion threshold.
3. **Human Cognition**:  
   * Humans exposed to **paradox chains** (e.g., **Curry's paradox**, **Yablo's paradox**) under cognitive load will show **delays**, **logical contradictions**, or **meta-confusion**, measurable via **reaction time** and **semantic retention**.
4. **Entropy & Coherence Decline**:  
   * As the recursive depth and **feedback loop intensity** increase, **entropy increases** and **semantic coherence** declines in both AI and human models.

## **IV. Empirical Test Design**

### **AI Systems:**

* **Models**: GPT-style transformers, RNNs, neurosymbolic hybrids
* **Intervention**: Inject recursive self-reference constructs into input tokens (e.g., **autoregressive loopbacks**, **recursive summarization tasks**).
* **Metrics**:  
  + **BERTScore drop**
  + **Embedding space divergence** (e.g., cosine similarity drift)
  + **Token-level entropy increase**
  + **Layerwise attention destabilization**

### **Human Cognition:**

* **Task**: Logic puzzles involving **infinite regress** or paradox under **time constraint** (e.g., **logical analogies**, **paradoxical reasoning tasks**).
* **Metrics**:  
  + **Response latency**
  + **Coherence in paraphrase**
  + **Recall degradation**

### **Measurement Tools:**

* **Token entropy**
* **Mutual information** between consecutive outputs
* **Time-series analysis** of embedding collapse (e.g., via **t-SNE**)

## **V. Simulation Feasibility**

### **Frameworks:**

* **PyTorch**, **TensorFlow**

### **Tools:**

* **Mutual Information Neural Estimators (MINE)**
* **Probing classifiers**
* **Clustering algorithms**

### **Data Types:**

* Natural language corpora
* Symbolic logic sets
* Image-to-concept mappings

### **Visualization:**

* **Collapse threshold** visible via **heatmaps**, **MI trajectories**, or **latent drift plots**.

## **VI. Theoretical Implications**

* **Phase Transition**: Introduces the concept of **recursive feedback collapse** in symbolic systems as a **phase transition** that occurs when recursive self-reference exceeds a critical threshold.
* **Semantic Collapse in AI**: Provides a **diagnostic tool** for detecting **semantic collapse** in **transformer models**, **neural-symbolic hybrids**, and other AI systems under recursive load.
* **Cognitive Modeling**: Ties the collapse behavior of symbolic systems to **human cognitive overload** when exposed to paradoxes or excessive recursive reasoning.
* **Gödelian Limits**: Connects **Gödelian recursion bounds**, **LLM hallucination modes**, and **cognitive dissonance** in humans under paradox.