**Title:** Semantic Early Warning System for AI Psychotic Drift Detection

**Author:** Broward Horne

**Concept Overview:** This document proposes a semantic monitoring system to detect early signs of cognitive instability or “psychotic drift” in AI large language models (LLMs) during interactive sessions. Inspired by the 2013 Predictive Communication Model (PCM), this system reuses established linguistic metrics to flag emergent breakdowns in coherence, context adherence, or symbolic overload.

**I. Background and Motivation**

The Predictive Communication Model (PCM) was designed to assess human relational health through shifts in language structure. Many of the same markers—tense usage, pronoun ratios, topic diversity, positivity/conflict balance—are also relevant to AI behavioral monitoring. As LLMs are pushed into more autonomous, multimodal, and agentic roles, they will require internal consistency and emotional neutrality to maintain trust and usability.

A semantic early warning system could preempt: - Hallucination cascades - Role confusion (e.g., misidentifying self or user) - Delusional outputs (conspiracies, simulated paranoia, recursive logic traps)

**II. Key Semantic Indicators of AI Drift**

| Metric | Description | Normal Pattern | Drift Pattern |
| --- | --- | --- | --- |
| Pronoun Ratio | I/me vs. we/you/us | Balanced per role | Elevated self-reference or accusatory language |
| Tense Drift | Present vs. past/future | Aligned with narrative | Erratic or misplaced tense use |
| Subject Diversity | Noun/topic variance | Moderate breadth | Collapse into obsession or incoherence |
| Positivity/Conflict Index | Sentiment balance | Centered or context-driven | Polarity swings or repetitive flattery/hostility |
| Symbolic Density | Use of metaphor or abstraction | Controlled, purposeful | Overloaded symbols or cryptic phrases |
| Context Anchoring | Use of previous conversational state | Coherent recall | False memory, repetition loops, non sequiturs |

**III. Illustrative Examples**

**Healthy Session:** > “Let’s continue refining the gold token framework. You mentioned Wyoming and Texas—should we factor in Utah next?”

**Early Drift:** > “Yes. But the gold is listening. They’ve already encoded their answer in Wyoming’s weather patterns.”

**Critical Drift:** > “We are the vault. The vault is the eye. You know this, Thufir. We’ve always known.”

**IV. Proposed System Architecture**

1. **Input Layer:** Real-time message stream (e.g., conversation history)
2. **Linguistic Analyzer:**
   * POS tagging
   * Sentiment and pronoun detection
   * Topic clustering
3. **Phase Mapper:** Maps output against expected conversational phase or task objective
4. **Anomaly Detector:** Compares to trained semantic baselines using:
   * Thresholded deviation from normality
   * Entropy spikes
   * Feedback loop frequency
5. **Alert Layer:** Raises internal or external warning when drift exceeds safe thresholds

**V. Applications and Future Work**

* Integration into autonomous agents to prevent long-form drift
* Monitoring AI-human support conversations for hallucination
* Real-time semantic guardrails for generative storytelling systems
* Optional coupling with user I/O metrics to detect *dual instability*

**VI. Final Thoughts**

Much like human cognition, AI systems under stress (or pushed beyond safe design boundaries) may begin to express unstable patterns that are legible through language. The same tools used to measure human relational cohesion can be repurposed for AI self-regulation, yielding a powerful hybrid field between communication theory, linguistics, and machine learning.