Phase-Consistent Reasoning Supervisor

A Proposal Based on Multi-Month Experiments with OpenAI Tools

1. Introduction

Over several months, OpenAI tools (ChatGPT, Assistants API, embeddings) were used to study dialogue as a dynamical system.

A working prototype—“Joystick of Thought” + Phase Analyzer—was developed. Experimental results fully confirm the theoretical model.

2. Problem

LLMs produce hidden semantic drift, abstraction jumps, and meaning substitution that are hard for the user to detect. ChatGPT often appears coherent while moving to a different semantic attractor.

3. Solution Overview

The Phase-Consistent Supervisor analyzes each dialogue step by:

- semantic divergence
- abstraction-level shifts
- meaning substitution
- local contradictions
- phase-transition tension T

Dialogue becomes a sequence of episodes (attractor → spiral → attractor). High-T episodes mark critical semantic failures.

4. Experimental Findings

Experiments show:

- precise detection of semantic divergence
- identification of abstraction shifts
- prediction of coherence failures
- matching of experimental behavior with theoretical phase dynamics

5. Applications for ChatGPT

• real-time self-monitoring

• early warnings for semantic drift

• improved alignment

• training signal for reasoning stability

• user-facing coherence tracking

6. Proposal

OpenAI may explore integration:

• external supervisor

• in-loop reasoning monitor

• training-level signal

• research collaboration

Appendix A: Core Theoretical Basis (Summary)

• dialogue = dynamical system

• meaning attractors = stable states

• transitions = phase spirals on a torus

• divergence = phase mismatch

• combined tension T predicts breakdowns

• episodic memory = trajectory on attractors

• experiments fully confirm the model