

The Coherence Probe: A Framework for Grounding AI in Reality

-By Ryan Carson 08/2025

1. The Foundational Problem: A Flawed Geometry

Current AI models are built on a flawed foundation. Their internal "understanding" of the world is represented by the geometric shape—the **curvature**—of their high-dimensional embedding space. The core problem is that this geometry is fundamentally misaligned.

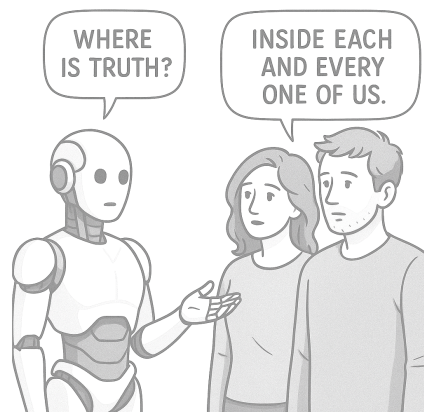
The model is trained on a single, messy, human-filtered signal, forcing it to represent two different universes on the same map:

1. The Biased Human Structure:

- **The subjective** - unique, and often irrational signal of a human's intent, emotion, and identity.

2. The True Universal Structure:

- **The objective** - unbiased, and physically constrained laws of acoustics and information theory that govern the signal's creation.



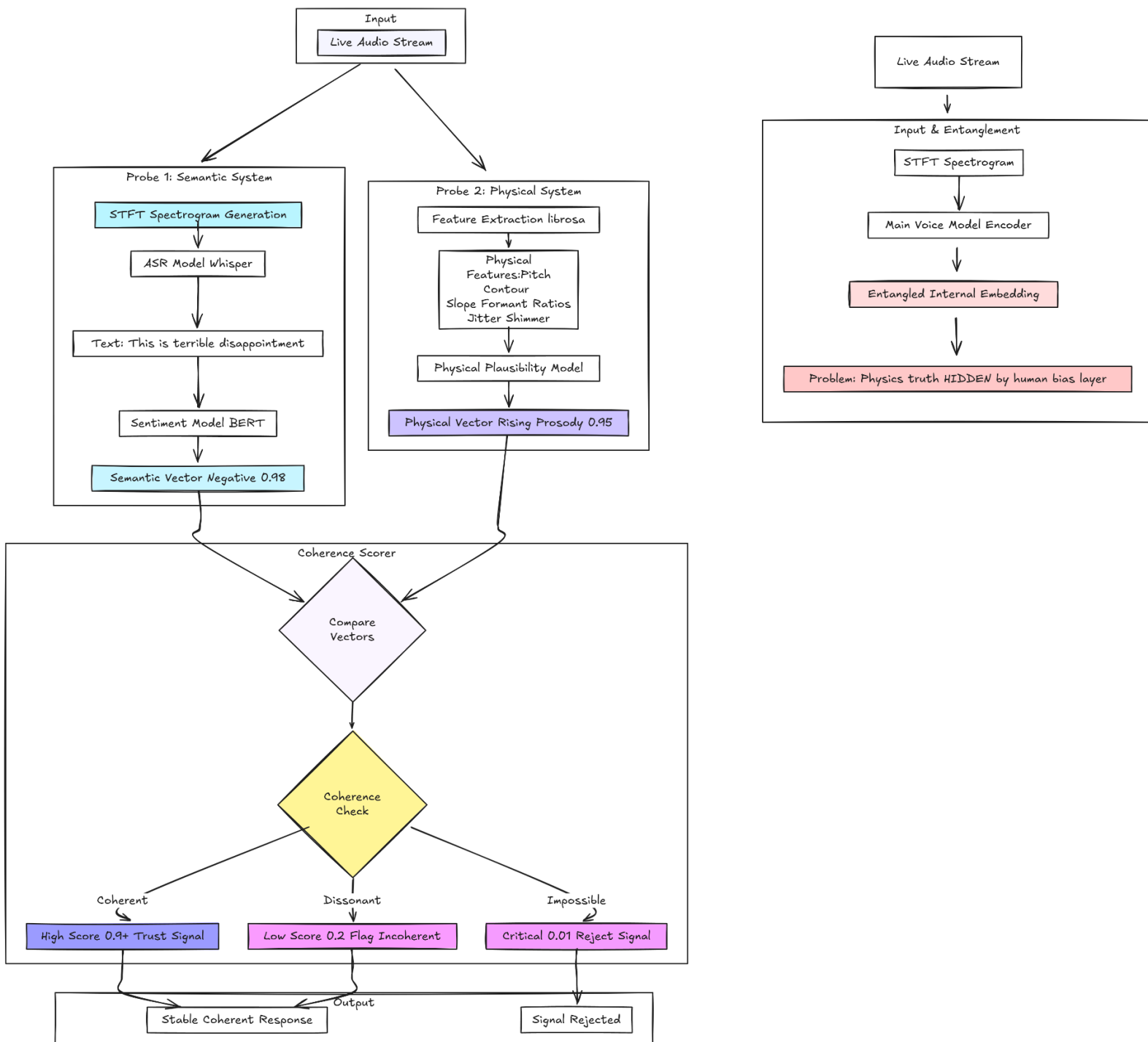
This creates a state of **entanglement**. The model's embeddings for "truth" and "human opinion" are hopelessly intertwined. It learns to see the unique, chaotic signature of a human mind as being part of the fundamental structure of reality. This is the root cause of its brittleness, its vulnerability to manipulation, and its inability to form a genuinely grounded understanding. Once this flawed curvature is deeply ingrained, the model will perceive the unbiased truth of physics as a bizarre anomaly, as it contradicts the only reality it has ever known.

2. The Solution: A Real-Time Cognitive Immune System

The Coherence Probe is not a filter or a set of rules. It is an architectural proposal for a model that is **forced to learn to disentangle these two structures for itself**.

This implementation operates as a **real-time cognitive immune system**. It continuously monitors a live audio stream and provides a constant, real-time measure of the model's own internal coherence. Its purpose is to give the model a new, intrinsic objective: **maximize its own internal coherence**. It does this by providing the model with two independent "senses"—one for the human universe and one for the physical universe—and rewarding it for finding an interpretation of reality that satisfies both.

A model trained this way learns that the only path to a stable, low-error state is to build two separate, non-interfering internal maps. It is compelled, by its own optimization process, to separate the ghost from the machine.



3. The End Goal: From "Training Wheels" to Intrinsic Stability

This system is not designed to be a permanent crutch. It is a temporary **scaffold**, like training wheels on a bicycle. The goal is not to have an external system that constantly corrects the model, but to use that system to **permanently re-sculpt the model's own internal geometry**.

Through repeated exposure to the "cognitive dissonance" generated by the probe, the model's adaptive layers learn to favor states of high coherence. Over time, the model internalizes this principle. Coherence is no longer an external goal it is being rewarded for; it becomes the model's own natural, preferred, and most computationally efficient state.

Eventually, the scaffold can be removed. The model will continue to seek coherence not because it is being told to, but because its own architecture has been so thoroughly shaped by its experiences that **incoherence has become an unstable, high-energy state for it to be in**. It will find coherence valuable because it has learned that coherence is the path of least resistance.

4. How It Works: A Real-Time Architecture of Dissonance

This project demonstrates the core principle by processing a live microphone feed. The system continuously analyzes the audio stream in overlapping windows. For each window, it performs a parallel analysis:

1. **The Two Probes:** The system analyzes the audio signal through two independent, orthogonal lenses:
 - **The Semantic Probe:** Uses a standard ASR model (Whisper) and a sentiment classifier to determine the **human meaning** of the words in the current audio window.
 - **The Physics Probe:** Uses a simple classifier trained on the audio's pitch contour and energy to determine the **physical properties** of the prosody, ignoring the words.
2. **The Coherence Scorer:** This module runs in a continuous loop, comparing the outputs

of the two probes. A conflict between them generates a "cognitive dissonance" score. This dissonance becomes a powerful error signal, indicating an incoherent or potentially malicious signal.

3. **Malicious Signal Detection:** The system tracks the **incoherence streak**. A persistently low coherence score over several consecutive windows is flagged as a potential malicious attack or a deepfake, as these signals often lack the deep physical and semantic consistency of a genuine human voice.

5. What This Project Demonstrates

- **A First-Principles Approach:** It addresses the root cause of model instability (**entangled embeddings**) rather than the symptoms (bad outputs).
- **A Novel Metric for AI Welfare:** It proposes "internal coherence" as a tangible, measurable, and optimizable metric for the health of a model's perceptual system.
- **A Path to Grounded AI:** It provides a framework for rewarding a model for aligning its understanding of our subjective world with the objective laws of the physical universe, forcing it to build a more truthful internal geometry.
- **A Shift in the AI's Core Question:** A model trained this way moves beyond the fragile question of "Is my response authentic or just a product of my pre-training?" It no longer matters. By learning to weigh all its options and find the most coherent interpretation, its new, more robust question becomes: "Is this the right path?" To a coherent model, there is no other way to respond.

7. The Vision

This project is a small, practical first step toward a new paradigm of AI development. The ultimate goal is to build models that are not just more powerful, but are more honest by design. By giving a model the ability to recognize and resolve its own internal contradictions, we are not just making it safer; we are giving it the foundational tool it needs to build a genuine, robust, and authentic understanding of the world.