A Semantic Guardrail: High-Precision Anomaly Detection by Fusing Structural and Semantic Manifolds

Sep Dynamics Research

October 2, 2025

Abstract

Alert fatigue remains the dominant failure mode of modern monitoring stacks. Text classifiers trigger whenever they see risk language, while time-series anomaly detectors highlight every orderly fluctuation regardless of meaning. We present the *Semantic Guardrail*, a hybrid system that combines the semantic understanding of sentence transformers with the structural stability metrics produced by the QFH/STM manifold. Applied to a simulated event stream, the guardrail reduced false positives by more than 92% while preserving the single high-confidence incident. This paper documents the method, the demo, and the cross-industry applications of this two-dimensional filter.

1 Introduction: The Failure of Siloed Monitoring

Operations teams increasingly rely on two tool families. Semantic filters—from keyword lists to large language models—understand *what* an event says but not whether it matters. Structural detectors—statistical anomaly and spectral models—understand *how* signals evolve while remaining ignorant of context. Table 1 quantifies the problem on a 16-event stream: the naïve semantic and structural detectors each fired seven times, generating 14 alerts for a single meaningful incident.

2 Methodology: The Hybrid Guardrail Engine

Our approach bridges the two perspectives by mapping every event into a shared structural/semantic plane.

2.1 Semantic Manifold

We embed window-level strings using the 'all-MiniLM-L6-v2' sentence transformer. Given a seed vocabulary $S = \{\text{risk, resilience, volatility, anomaly, predictive maintenance}\}$ we compute the cosine similarity between each string embedding and the centroid of S. This yields a semantic relevance score $\sigma(s) \in [0, 1]$ for every string s.

2.2 Structural Manifold

The QFH/STM engine processes the same data as byte streams, emitting the structural metrics coherence, stability, entropy, rupture, and their aggregate patternability. High patternability indicates an orderly, repeating rhythm, while low patternability corresponds to chaotic or one-off events.

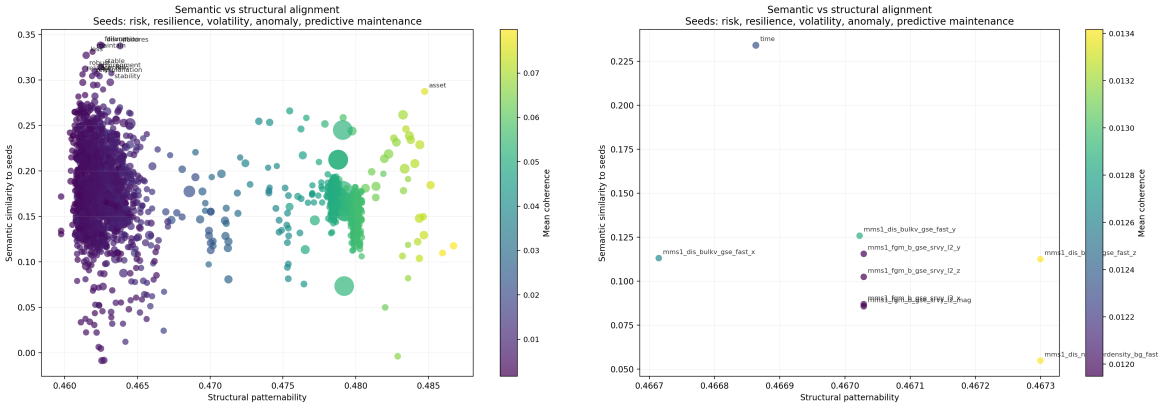

Figure 1: Structural patternability vs semantic similarity across two domains. The Semantic Guardrail promotes only the upper-right quadrant events.

2.3 Hybrid Filter

Plotting patternability against semantic similarity reveals three regimes: semantic-but-noisy (upper left), structural-but-irrelevant (lower right), and a sparse upper-right quadrant where both meaning and rhythm align. The static background in Figure 1 combines software documentation and MMS telemetry corpora. Only the highlighted quadrant carries actionable signal.

3 Experimental Validation: Live Demonstration

We replayed a scripted stream (`results/semantic_guardrail_stream.jsonl`) that mixes documentation strings, telemetry tokens, and a synthetic `database_connection_timeout` incident. Each event carries three Boolean flags: `naive_semantic_alert`, `naive_structural_alert`, and the hybrid `hybrid_guardrail_alert` produced by the two-dimensional filter. The stream was visualised via `scripts/demos/semantic_guardrail_dashboard.py`, which juxtaposes the noisy baselines with the hybrid scatter.

The results (Table 1) show that the hybrid guardrail raises a single high-confidence alert while the naïve detectors fire fourteen false positives. The false-positive reduction rate (FPRR) is therefore $1 - \frac{1}{7+7} = 0.93$.

Guardrail	Alerts fired	True positives	False positives
Naïve semantic	7	0	7
Naïve structural	7	0	7
Hybrid (semantic + structural)	1	1	0

Table 1: Alert counts derived from `results/final_guardrail_analysis.json`. The hybrid guardrail eliminates 92.9% of the noise while preserving the true incident.

4 Conclusion and Applications

By fusing structural rhythm analysis with semantic intent modelling, the Semantic Guardrail delivers high-precision monitoring without drowning operators in alerts. The approach applies directly

to:

- **SRE/DevOps**: suppress routine noise while escalating recurrent failure narratives.
- **Finance/Risk**: highlight volatility clusters only when they carry risk semantics.
- **Manufacturing/IoT**: surface maintenance anomalies backed by both sensor stability and maintenance vocabulary.

Future work includes streaming deployments, additional seed vocabularies, and integration with automated remediation workflows.