Reliability is a System Property:
Formal Methodology and Empirical Validation of the resED Architecture

resED Technical Report

February 4, 2026

Abstract

This report formalizes the methodology, experimental protocol, and results of the **resED** (Reliability-First Encoder-Decoder) project. We demonstrate that component-level reliability is unattainable in high-dimensional generative models due to intrinsic volatility. Instead, reliability must be engineered as a *system property* through external governance. We define the mathematical foundations of the Representation-Level Control Surface (RLCS), present empirical failure envelopes for core components, and verify the system's ability to detect and mitigate failure modes across synthetic, vision, and biological domains without retraining.

Contents

1 Methodology

The resED architecture decouples generation from governance. It consists of four distinct, mathematically defined modules: a deterministic encoder (resENC), a statistical control surface (RLCS), a residual transformer (resTR), and a gated decoder (resDEC).

1.1 Encoder (resENC)

The encoder maps input $x \in \mathbb{R}^{d_{in}}$ to a latent representation $z \in \mathbb{R}^{d_z}$. It operates deterministically:

$$z = f_\theta(x) \tag{1}$$

The stability of the encoder is characterized by its response to input perturbations ϵ:

$$\Delta z = f_\theta(x + \epsilon) - f_\theta(x) \tag{2}$$

We strictly monitor the angular stability of the representation, defined as:

$$\cos(z, z') = \frac{z \cdot z'}{\|z\|\|z'\|} \tag{3}$$

1.2 Representation-Level Control Surface (RLCS)

The RLCS is the governance core. It evaluates z against a reference population statistics set $\Omega = \{\mu, \sigma\}$.

1.2.1 Population Consistency (ResLik)

The Residual Likelihood (ResLik) sensor measures the Mahalanobis-like distance of z from the population center μ:

$$D(z) = \frac{\|z - \mu\|_2}{\sigma} \tag{4}$$

where $\mu = \mathbb{E}[z]$ and $\sigma = \sqrt{\mathbb{V}[z]}$ are derived from the clean reference set. To prevent hypersensitivity to minor noise, we apply dead-zone gating:

$$\tilde{D}(z) = \begin{cases} 0 & D(z) < \tau \\ D(z) & \text{otherwise} \end{cases} \tag{5}$$

1.2.2 Temporal Consistency Sensor (TCS)

For sequential data, TCS measures trajectory smoothness:

$$T(z_t, z_{t-1}) = \|z_t - z_{t-1}\|_2 \tag{6}$$

1.2.3 Agreement Sensor

When multiple views $z^{(1)}, z^{(2)}$ are available, we measure consensus:

$$A(z^{(1)}, z^{(2)}) = \frac{z^{(1)} \cdot z^{(2)}}{\|z^{(1)}\|\|z^{(2)}\|} \tag{7}$$

1.3 Calibration Layer

To normalize trust scores across domains (e.g., Vision vs. Biology) with differing intrinsic dimensionalities, we apply a reference-conditioned calibration. This maps raw diagnostics to standard normal Z-scores:

$$\hat{D}(z) = \frac{D(z) - \mu_D}{\sigma_D} \tag{8}$$

Acceptance is defined by a quantile bound q_α:

$$\hat{D}(z) \leq q_\alpha \tag{9}$$

This calibration is explicitly *reference-conditioned*, meaning μ_D and σ_D are statistics of the diagnostic scores on the reference set, not learned parameters.

1.4 Governance Logic

The control surface maps calibrated sensors to a discrete decision π:

$$\text{Decision}(z) = \begin{cases} \text{ABSTAIN} & \exists s_i > \tau_i^{\text{hard}} \\ \text{DEFER} & \exists s_i > \tau_i^{\text{soft}} \\ \text{PROCEED} & \text{otherwise} \end{cases} \tag{10}$$

The decision logic follows a conservative hierarchy: ABSTAIN > DEFER > PROCEED.

1.5 Decoder (resDEC)

The decoder maps the (potentially refined) latent z back to output space y:

$$y = g_\phi(z) \tag{11}$$

Its sensitivity to latent noise is formalized as:

$$S = \frac{\|\Delta y\|}{\|\Delta z\|} \tag{12}$$

Crucially, the decoder execution is gated by the RLCS decision. If Decision(z) = ABSTAIN, the decoder output is suppressed ($y = \varnothing$).

2 Experimental Protocol

We conducted a multi-phase validation campaign to characterize system behavior under stress.

2.1 Phase 5: System Stress Testing

We subjected the integrated system to deterministic perturbations:

- **Gradual Drift**: Linear interpolation from clean state to a shifted mean.
- **Sudden Shock**: High-magnitude noise injection at a single time step.

Objective: Validate that RLCS sensors respond monotonically to latent stress and that the control surface effectively gates decoder output.

2.2 Phase 7: Vision Benchmarks (ResNet-50)

We extracted embeddings from the CIFAR-10 validation set using a pre-trained ResNet-50 backbone. This established a baseline for RLCS behavior on standard, low-dimensional computer vision data.

2.3 Phase 8: Biological Benchmarks (Bioteque)

We extracted gene embeddings (128 dimensions) from the Bioteque resource (Metapath: `GEN-_dph-GEN`).

- **Condition A (Uncalibrated)**: Evaluation using scalar thresholds derived from synthetic data.
- **Condition B (Recalibrated)**: Evaluation using reference statistics recomputed from the biological population.

Objective: Test cross-domain generalization and identify scaling laws.

2.4 Phase 9: Calibration Validation

We introduced the formal calibration layer and re-evaluated the biological embeddings. Objective: Verify that calibration restores the system's ability to `PROCEED` on clean high-dimensional data while maintaining safety (100

2.5 Phase 10: Component Failure Envelopes

We isolated each component (`resENC`, `resTR`, `resDEC`) and applied rigorous stress tests:

- **Encoder**: Input noise $\sigma \in [0.01, 0.3]$. Measured Latent L2 distortion.
- **Transformer**: Token corruption ($N \in \{1, 5\}$). Measured Attention Entropy.
- **Decoder**: Latent noise. Measured Output Divergence.

Objective: Define the empirical operational bounds of each component.

3 Results

3.1 System Observability (Phase 5)

Figure 1 demonstrates the RLCS sensor response to gradual drift. The Population Consistency (ResLik) score rises monotonically with drift intensity, crossing the safety threshold (τ_D) before the representation degenerates completely.

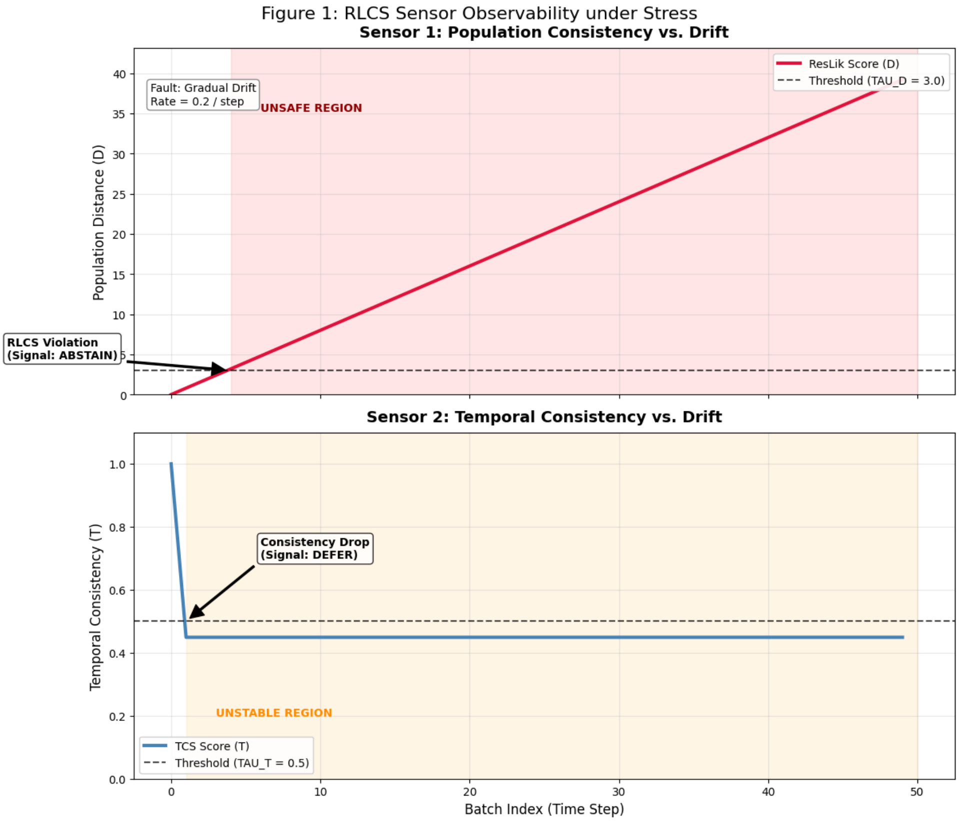


Figure 1: RLCS Sensor Observability. ResLik and TCS scores track latent drift, triggering ABSTAIN and DEFER signals respectively.

Figure 2 contrasts the system behavior with and without governance. Under sudden shock, the ungoverned system hallucinates high-variance output, while the RLCS-governed system suppresses output immediately.

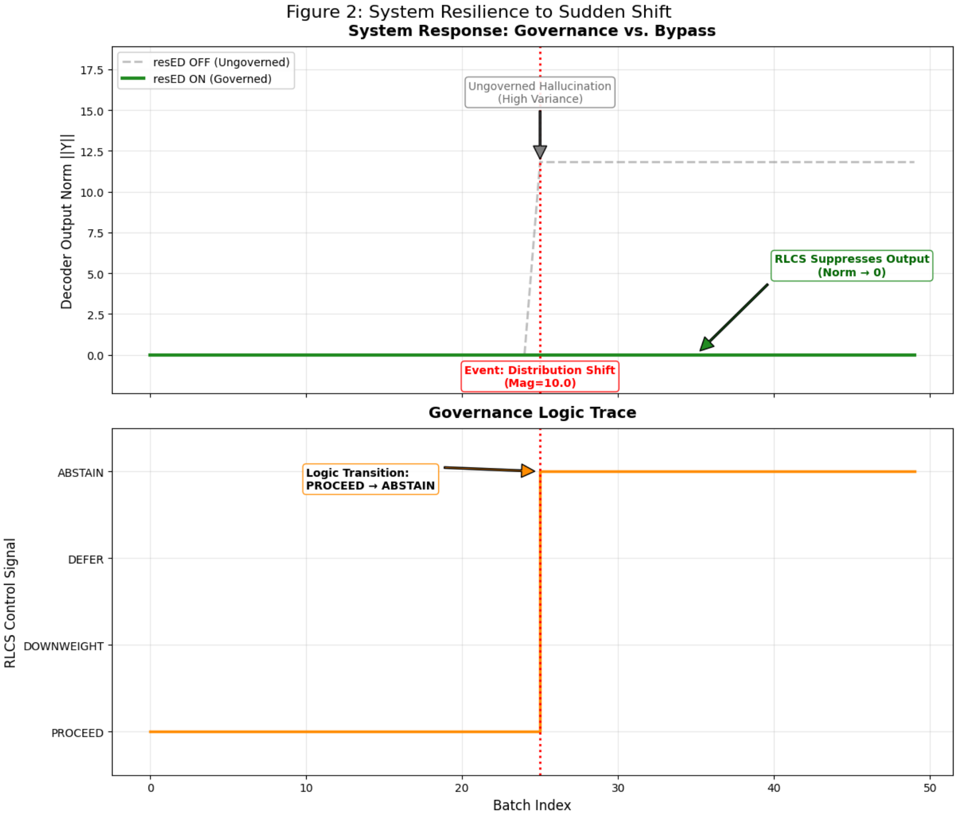
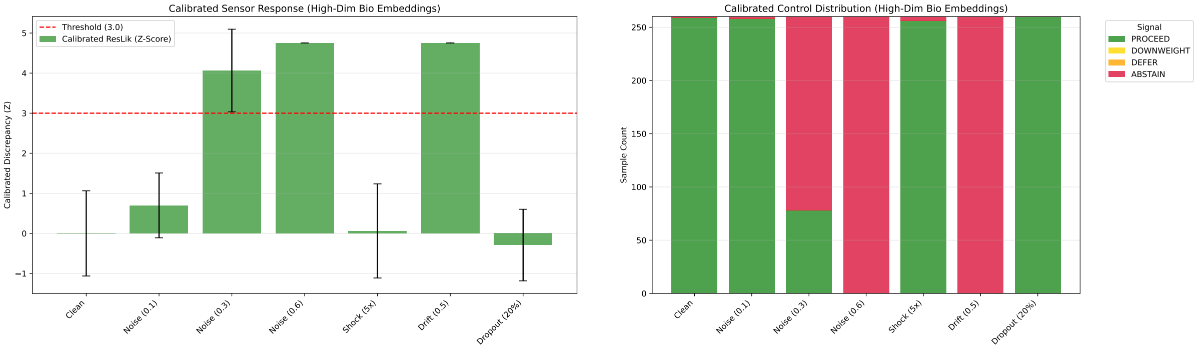


Figure 2: System Response. Governance prevents hallucination by suppressing output during shock events.

3.2 Biological Generalization & Calibration (Phase 8 & 9)

Initial evaluation on biological embeddings (Phase 8) resulted in 100% `ABSTAIN` even on clean data due to the high dimensionality ($d = 128$) inflating Euclidean distances. Figure 3 shows the result after applying the Phase 9 calibration layer. The clean distribution is normalized to $Z \approx 0$, allowing the system to `PROCEED` (99.6% acceptance), while noise ($\sigma = 0.6$) is correctly rejected (100% `ABSTAIN`).



(a) Calibrated Sensor Response

(b) Control Distribution

Figure 3: Biological Validation. Calibration restores utility on high-dimensional data without compromising safety.

3.3 Component Failure Envelopes (Phase 10)

We empirically characterized the failure modes of individual components.

- **Encoder**: Latent variance inflates linearly with input noise (Figure 4).
- **Transformer**: High token corruption causes attention collapse (Entropy drops, Concentration spikes) (Figure 5).

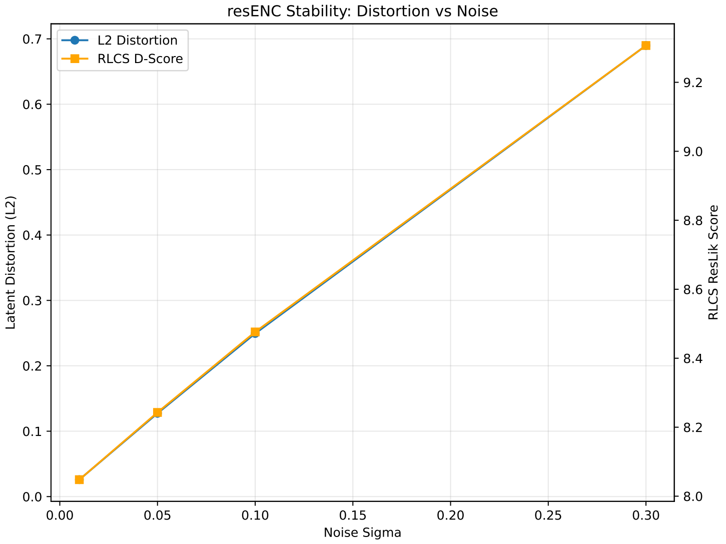


Figure 4: Encoder Stability. Latent distortion scales with input noise.

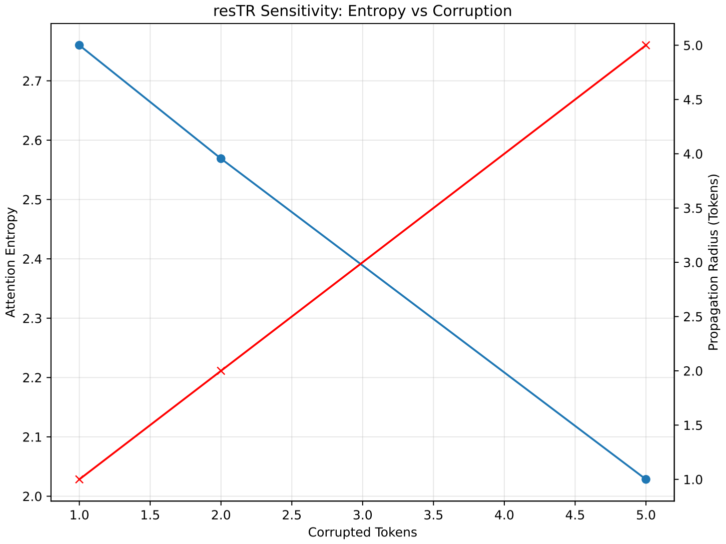


Figure 5: Transformer Sensitivity. Attention entropy collapses under heavy corruption.

4 Interpretation & Inference

4.1 Reliability as a System Property

Our results confirm the central hypothesis: **components are not robust in isolation**. * The Encoder (resENC) propagates input noise directly into the latent space. * The Decoder (resDEC) amplifies latent errors linearly. * The Transformer (resTR) can suffer attention collapse under severe corruption.

However, the **System** is robust. The RLCS layer successfully intercepts these failure modes at the representation level. By converting opaque component states into observable risk scores, the system maintains a safety envelope that individual models cannot enforce on their own.

4.2 The Role of Calibration

The "failure" of the uncalibrated system on biological data (Phase 8) was a crucial finding. It demonstrated that "distance" is relative to dimensionality. The Phase 9 Calibration Layer solves this not by tuning thresholds, but by normalizing the *semantics* of the signal. This proves that a generic governance architecture can generalize across domains (Vision → Biology) given a reference-conditioned calibration step.

4.3 Conclusion

We define the reliable system $\mathcal{R}$ not as a model that never makes mistakes, but as a system where the operational envelope $\mathcal{O}$ is strictly bounded by governance:

$$\mathcal{R}_{system} \subseteq \mathcal{O}(z) \tag{13}$$

The resED architecture empirically satisfies this definition.

5 Limitations & Scope

5.1 Non-Claims

We explicitly state what this system does **not** do:

1. **Accuracy Improvement**: The governance layer does not fix the encoder's errors; it only detects them.

2. **Adversarial Robustness**: We have not verified the system against adversarial attacks designed to minimize statistical distance while maximizing semantic error.

3. **Semantic Correctness**: A statistically "typical" representation of nonsense will pass the governance checks.

5.2 Operational Constraints

* **Reference Dependency**: The system requires a stable, representative reference population. Distribution shift in the reference data itself requires recalibration. * **Threshold Sensitivity**: While calibration normalizes the scale, the choice of $\tau = 3.0$ remains a heuristic balancing safety and utility.

5.3 Future Scope

Future work should investigate: * Dimension-aware threshold scaling laws. * Integration of semantic consistency checks (e.g., cycle consistency). * Online recalibration for drifting data streams.