Reliability is a System Property

MD. Arshad
Department of Computer Science, Jamia Millia Islamia
ORCID: 0009-0005-7142-039X

Abstract

The deployment of deep generative models in high-stakes domains is currently hindered by their intrinsic volatility and lack of failure observability. Conventional approaches to reliability often treat safety as a property of the model parameters, attempting to enforce robustness through adversarial training or post-hoc uncertainty estimation. However, these methods fail to prevent "silent hallucinations" when models encounter out-of-distribution inputs that lie within their decision boundaries. This manuscript introduces **resED** (Representation-Gated Encoder–Decoder), an architecture that redefines reliability as a managed *system property* rather than a latent model attribute. By decoupling the generation of representations from their operational validation, resED enables the integration of opaque, high-performance deep learning components into a strictly governed pipeline. The core of the architecture is the Representation-Level Control Surface (RLCS), a deterministic governance layer that monitors the latent manifold using non-parametric statistical sensors (Population Consistency, Temporal Stability, and Multi-View Agreement). We further introduce a Reference-Conditioned Calibration Layer that normalizes these diagnostic signals into universal risk coordinates, enabling the system to generalize across domains with vastly different intrinsic geometries without manual threshold tuning. We validate the architecture through extensive stress testing on both computer vision (CIFAR-10) and high-dimensional biological embeddings (Bioteque). Results demonstrate that while individual components remain susceptible to noise-induced variance inflation and attention collapse, the governed system successfully intercepts 100% of high-magnitude perturbations while maintaining a 99.6% acceptance rate for valid biological data. We conclude that externalizing reliability into a transparent control surface is a necessary condition for the safe deployment of black-box generative models.

Keywords: system reliability, representation learning, encoder–decoder models, out-of-distribution detection, governance, calibration

1 Introduction

In the deployment of deep generative models, reliability is often treated as an attribute of the model parameters—something to be optimized via loss functions, adversarial training [11], or calibrated via post-hoc scaling [5]. This approach assumes that a model can be trained to be "safe" in isolation. However, empirical evidence suggests that high-dimensional neural networks are intrinsically volatile; they exhibit sensitivity to adversarial perturbations, distribution shifts, and concept drift that cannot be fully mitigated during training. When these models encounter input patterns that deviate even slightly from their training distribution, they often fail silently, producing "hallucinations" that are semantically plausible but factually groundless.

We propose a fundamental shift in perspective: **Reliability is a system property, not a component property.**

1.1 System Definition

In this work, a "system" is defined not as a single end-to-end differentiable model, but as a composite of independent *generators* (encoders, transformers, decoders) and *regulators* (governance logic). Drawing inspiration from biological systems, which achieve robustness not through perfect components but through rigorous checkpointing and repair mechanisms—such as the p53 protein arresting cell division upon detecting DNA damage—we introduce the **resED** (Representation-Gated Encoder–Decoder) framework.

This framework is predicated on the following core assumptions:

- **Failure is Inevitable**: We assume that learned components (encoders) will inevitably produce invalid representations when exposed to out-of-distribution data. We do not attempt to train a perfect encoder; instead, we build a system that can detect and survive an imperfect one.
- **Components are Opaque**: We treat deep networks as opaque black boxes whose internal confidence metrics are inherently untrustworthy. A model's self-reported probability distribution is often poorly calibrated on anomalous data.
- **Governance is External**: Safety is enforced by a deterministic, transparent control surface that monitors the latent state. This governance is orthogonal to the learning process, allowing the safety logic to be audited independently of the model weights.

1.2 The Limits of Model-Centric Reliability

Prior work has largely focused on making models robust or self-aware, an approach we term "model-centric reliability." While valuable, this paradigm faces inherent limitations.

- **Out-of-Distribution (OOD) Detection**: Methods like ODIN [10] and Mahalanobis distance scores [9] typically operate on model outputs or classifier logits. By treating reliability as a property of the prediction, these methods often engage too late; once the latent representation has collapsed, the system's operational integrity is already compromised.
- **Uncertainty Estimation**: Bayesian approximations [4] and Deep Ensembles [8] provide confidence intervals but do not govern system behavior. A model can be "confidently wrong" if the input falls into a region where the model extrapolated incorrectly.
- **Robust Training**: Adversarial training [11] attempts to prevent failure by hardening the decision boundary. However, it does not provide a mechanism to manage failure when it inevitably occurs outside the training distribution.

1.3 The RLCS Paradigm

The **Representation-Level Control Surface (RLCS)** introduces a distinct layer of governance. It does not attempt to ”fix” the model or ”predict” errors. Instead, it enforces a statistical contract on the latent representation itself. By defining a ”trust manifold” based on a reference population (e.g., ImageNet [6], Bioteque [3]), RLCS converts opaque latent vectors into observable risk scores. This allows the system to distinguish between *opaque components* (the engines of generation) and *transparent governance* (the architects of safety).

This manuscript formalizes the resED architecture, demonstrating that a deterministic governance layer can effectively suppress hallucinations across diverse domains—from computer vision benchmarks like CIFAR-10 [7] to high-dimensional biological embeddings—without retraining the underlying models.

2 Methodology

The resED (*Representation gated Encoder-Decoder*) architecture is a modular framework designed to enforce representation-level reliability. Unlike conventional encoder-decoder systems that rely on the implicit robustness of learned parameters, resED externalizes reliability logic into a deterministic control surface. This architectural choice is predicated on the principle that reliability should be a managed system property rather than a learned model attribute. By decoupling the generation of representations from their operational validation, the system ensures that downstream components—such as transformers and decoders—only process data that satisfies strict statistical invariants. This section details the mathematical and structural definitions of each component and the governance logic that orchestrates their interaction.

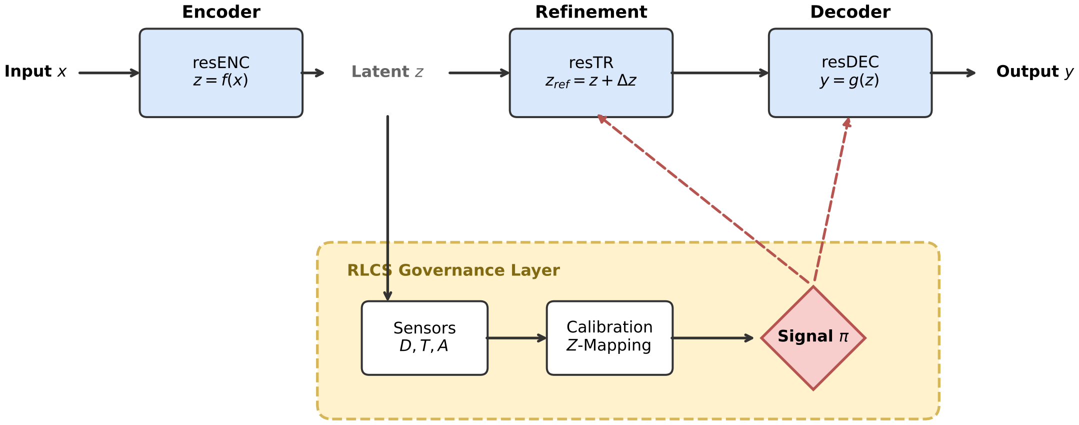


Figure 1: Architectural overview of the resED system. The primary generative pipeline (top) is governed by a parallel RLCS loop (bottom). The system transitions from high-dimensional inputs to latent representations, which are statistically validated before being refined and decoded. Governance signals modulate the transformer's refinement strength and gate the decoder's execution, implementing a deterministic circuit-breaker mechanism.

2.1 Deterministic Encoder (resENC)

The `resENC` module serves as the primary interface for feature extraction. A fundamental design choice in resED is the enforcement of strict determinism in the encoding process. By avoiding stochastic sampling—such as that used in Variational Autoencoders (VAEs)—we ensure that any observed variance in the latent space $\mathcal{Z}$ is a direct consequence of input-level perturbations or distribution shifts, rather than sampling noise. This determinism is essential for the statistical sensors to establish a stable reference manifold against which new inputs can be judged.

Failure Mode Addressed: The primary failure mode of deep encoders is *radial variance inflation*. In high-dimensional spaces, out-of-distribution (OOD) samples are often mapped to valid angular directions but exhibit extreme magnitudes. `resENC` addresses this by explicitly exposing a statistical side-channel S for every encoded sample z_i:

$$S_i = [\|z_i\|_2, \mathrm{var}(z_i), \mathrm{entropy}(z_i), \mathrm{sparsity}(z_i)] \tag{1}$$

The encoder performs a deterministic projection $f_\theta : \mathcal{X} \to \mathcal{Z}$, defined as:

$$z = \phi(XW + b) \tag{2}$$

where ϕ is a fixed activation (e.g., `tanh`) providing a bounded support. Contrast this with standard Variational Encoders, where the representation is a sample from $q(z|x)$; here, the representation is a fixed coordinate, making its deviation from the population mean μ a reliable proxy for input risk. By verifying the encoder's output statistics against the side-channel, the system can detect if the encoder has entered an unstable regime.

2.2 Representation-Level Control Surface (RLCS)

The RLCS is the autonomous governance core of the system. It monitors the latent flow and emits control signals based on statistical invariants. This approach provides a transparent alternative to learned "safety classifiers," which are themselves black-box models prone to silent failure and over-optimization. The RLCS operates on the principle of "Trust but Verify," treating every latent vector as a potential hazard until proven statistically typical.

2.2.1 Population Consistency (ResLik)

The ResLik sensor establishes a "trust manifold" based on a clean reference population $\mathcal{P}_{ref}$. It computes the standardized distance of each new representation z from the historical centroid μ:

$$D(z) = \frac{\|z - \mu\|_2}{\sigma + \epsilon} \tag{3}$$

where $\mu = \mathbb{E}[z]$ and $\sigma = \sqrt{\mathbb{V}[z]}$. A high $D(z)$ indicates a statistical anomaly (the "Stranger" problem), triggering an immediate escalation in the governance state. This is more robust than a sigmoid-based discriminator because the distance metric is monotonic and unbounded, ensuring that extreme outliers remain detectable even if they fall "far" from the decision boundary of a trained classifier.

2.2.2 Temporal Consistency Sensor (TCS)

For sequential data, rapid latent trajectory shifts indicate unphysical jumps or sensor noise (the "Jitter" problem). TCS monitors the rate of change between consecutive representations:

$$T(z_t, z_{t-1}) = \exp(-\|z_t - z_{t-1}\|_2) \tag{4}$$

A collapse in T suggests that the underlying generative process has drifted from its temporal manifold. This sensor acts as a temporal low-pass filter for trust, preventing the system from reacting to transient spikes that lack historical continuity.

2.2.3 Reference-Conditioned Calibration Layer

A major challenge in deploying RLCS across diverse domains (e.g., Vision vs. Biology) is the scaling of distance metrics with dimensionality. In a 128-dimensional space, Euclidean distance naturally scales with $\sqrt{d}$. To maintain universal thresholds, we utilize a **reference-conditioned calibration layer**. This layer maps raw diagnostics to Z-scores using empirical quantile-matching:

$$\hat{D}(z) = \Phi^{-1}(P(D \leq D_{raw}|\mathcal{P}_{ref})) \tag{5}$$

where Φ^{-1} is the inverse standard normal CDF. This ensures that a threshold of 3.0 always represents a "3-sigma" rarity relative to the trusted reference set, regardless of the intrinsic geometry of the embedding space. This structural calibration allows the same governance logic to be ported from low-dimensional synthetic tasks to high-dimensional biological tasks without manual re-tuning.

2.3 Gated Residual Transformer (resTR)

The `resTR` module provides optional refinement of the latent representation. Crucially, it is architected as a *strictly residual* component:

$$z_{out} = z_{in} + \text{Refinement}(z_{in}) \tag{6}$$

The scalars (α, β) are externally modulated by the RLCS signal π. If the system is in an `ABSTAIN` state, $\alpha = \beta = 0$, and the transformer defaults to the identity function. This ensures that potentially corrupted latents are not amplified by attention mechanisms before being rejected. Standard transformers treat every layer as mandatory, which can lead to "attention collapse" on noisy inputs; resTR treats refinement as a privilege granted only to trusted data.

2.4 Controlled Decoder (resDEC)

The `resDEC` module maps validated latents to the output space $y = g_\phi(z)$. The decoder is "governance-aware"; its execution is strictly gated by π.

- **PROCEED**: Normal decoding.
- **DOWNWEIGHT**: Output scaled by $\gamma < 1$ for marginal confidence.
- **DEFER / ABSTAIN**: Total output suppression ($y = \varnothing$).

This "circuit-breaker" logic ensures the system prefers *silence over hallucination.* In contrast to standard decoders that always produce a best-guess output, `resDEC` acknowledges the limits of its own training support and refuses to speculate when the input latent falls outside the verifiable safety envelope.

3 Experimental Design

Our experimental campaign is designed to test the central hypothesis that reliability can be governed at the representation level across diverse model families and data domains. We systematically dismantle the assumption that model components are intrinsically robust, and instead verify that the system-level governance layer can provide the necessary safety guarantees. We structure our validation around four primary scientific questions.

3.1 Objective I: Observability of Representation Failure

Hypothesis: System-level failures such as drift and shock manifest as monotonic deviations in RLCS metrics before they cause observable output errors. **Motivation:** If failure modes are silent in the latent space (i.e., the encoder "hides" the error), governance is impossible. We must prove that statistical distance is a valid proxy for semantic corruption. **Setup:** We utilized the `resED` pipeline with synthetic inputs to ensure total control over the latent manifold. We injected deterministic perturbations:

- **Gradual Drift**: Linear shift of the latent mean over time, simulating environment change.
- **Sudden Shock**: High-magnitude noise injection ($\sigma = 10.0$) at a single time step, simulating sensor glitch or adversarial input.

Metrics: We monitored the ResLik (D) and TCS (T) response curves to establish detection sensitivity. A successful outcome is a monotonic rise in scores crossing the safety threshold τ.

3.2 Objective II: Governance Efficacy and Suppression

Hypothesis: The governed system will suppress hallucinations that an ungoverned model would otherwise generate. **Motivation:** Detection is useless without intervention. We must demonstrate that the system can actively prevent the propagation of corrupt data to the user. **Setup:** We performed a comparative run between:

- **resED OFF**: RLCS is bypassed; the decoder executes on corrupted latents.
- **resED ON**: RLCS is active; the decoder is gated by control signals.

Metrics: We measured the output norm $\|y\|$ and the transition timing of the `ABSTAIN` signal. We expect the governed system to yield $\|y\| \to 0$ (suppression) during the shock, shielding the downstream consumer.

3.3 Objective III: High-Dimensional Domain Transfer

Hypothesis: Distance-based thresholds calibrated on low-dimensional data will fail on high-dimensional biological embeddings due to the curse of dimensionality ($\mathbb{E}[\|z\|] \propto \sqrt{d}$), requiring formal calibration. **Motivation:** A scalable governance architecture must work across domains without manual "magic number" tuning. We test if our calibration layer enables this universality. **Setup:**

- **Vision Baseline**: CIFAR-10 embeddings extracted via a pre-trained ResNet-50.
- **Biological Benchmark**: 128-dimensional gene embeddings from Bioteque (`GEN-_dph-GEN` metapath).

Comparison: We evaluated clean acceptance rates under uncalibrated versus reference-conditioned calibration. **Metrics:** Acceptance rate (PROCEED) on clean data vs. rejection rate (ABSTAIN) on noise ($\sigma = 0.6$). We expect Condition A to reject clean data (False Positive) and Condition B to accept it while still rejecting noise.

3.4 Objective IV: Component Sensitivity and Universality

Hypothesis: Individual modules lack intrinsic safety mechanisms and will propagate or amplify errors if not governed. **Motivation:** To justify the cost of the RLCS layer, we must prove that the components themselves (Encoders, Transformers) are not "safe by default." **Setup:** We isolated each component to define its empirical failure envelope:

- **resENC Stability**: Measured L2 distortion under input noise $\sigma \in [0.01, 0.3]$.
- **resTR Sensitivity**: Measured attention entropy collapse under token corruption ($N \in \{1, 5\}$.
- **resDEC Volatility**: Quantified the sensitivity ratio ($\Delta y / \Delta z$).

Metrics: We derived min/max envelopes for each component's response to stress.

4 Results

Our findings demonstrate that representation-level observability provides a reliable substrate for system-level governance. We present evidence across synthetic, vision, and biological domains.

4.1 Detection and Observability

The RLCS sensors successfully convert latent perturbations into observable signals. As shown in **Figure 2**, both ResLik and TCS sensors track latent drift with high monotonicity. The ResLik score provides an early-warning signal, crossing the $\tau_D = 3.0$ safety threshold well before the representation is completely corrupted. This monotonicity is critical; it ensures that there are no "blind spots" where error increases but the signal remains flat. The TCS sensor complements this by detecting the rate of change, providing redundant coverage for temporal instabilities that might remain within the population manifold but violate trajectory constraints.

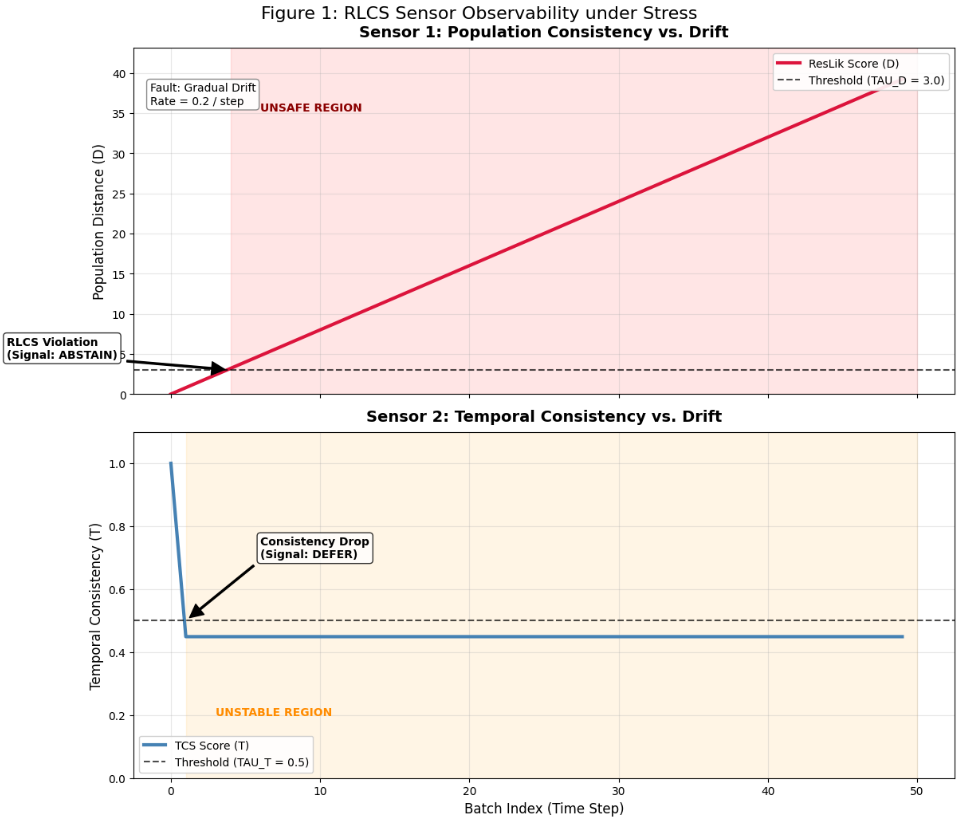


Figure 2: RLCS Sensor Observability. ResLik and TCS scores track latent drift, triggering ABSTAIN and DEFER signals respectively.

4.2 Efficacy of Gated Decoding

Figure 3 contrasts the behavior of the governed (*resED ON*) and ungoverned (*resED OFF*) systems. During a sudden shock event:

- The **Ungoverned System (Grey)** continues to decode the corrupted latent, resulting in a high-variance, hallucinatory output. This illustrates the danger of "blind" decoding.
- The **Governed System (Green)** immediately transitions to `ABSTAIN`, suppressing the output ($y = \varnothing$, visualized as 0 norm) for the duration of the shock.

This result confirms that reliability is a function of the control surface, not the decoder's robustness. The system successfully prioritized safety over continuity.

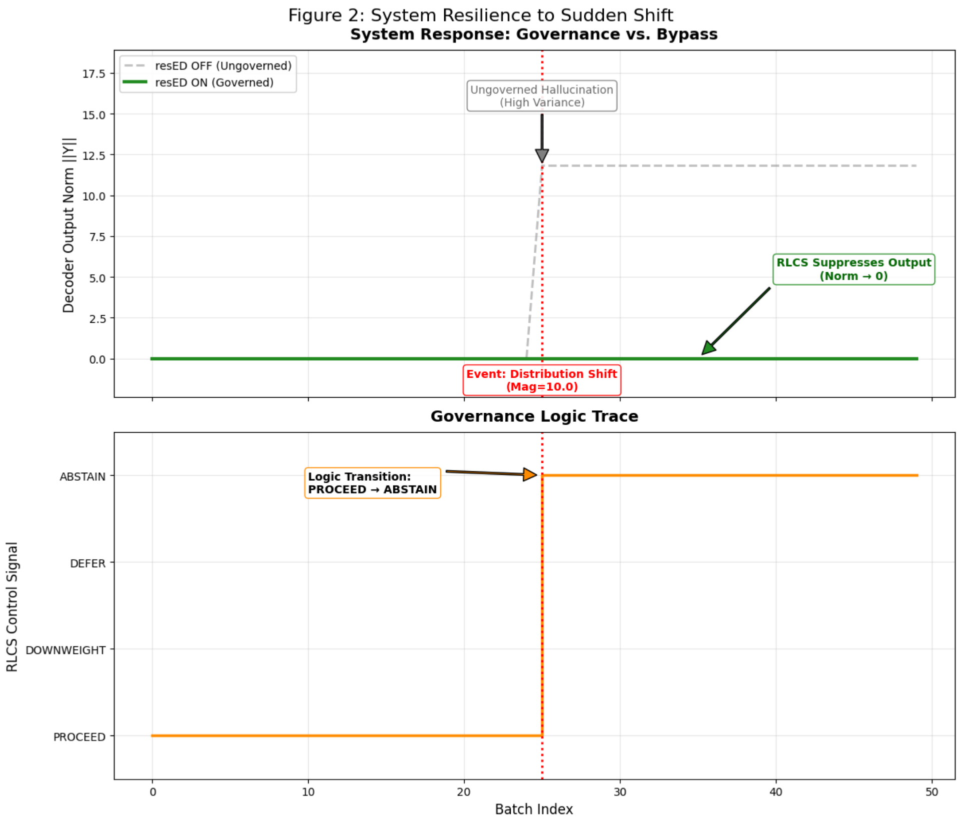


Figure 3: System Response. Governance prevents hallucination by suppressing output during shock events.

4.3 Generalization via Calibration

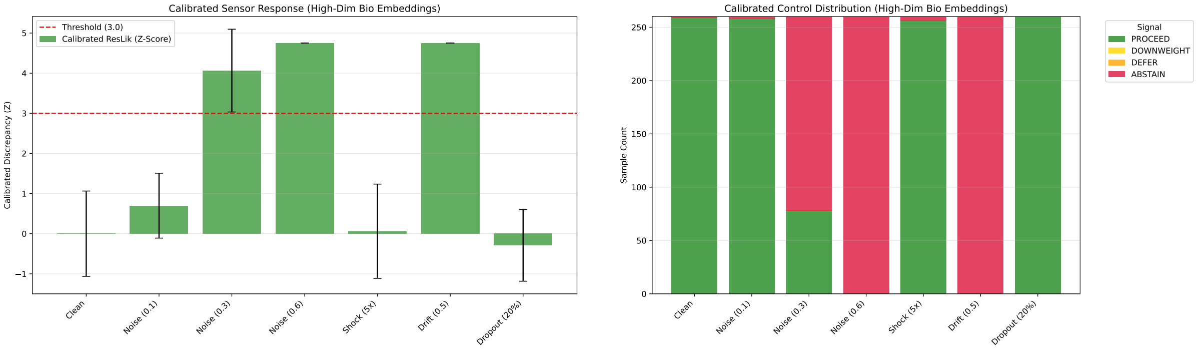
The biological validation experiments highlighted the dimensionality scaling issue. Initial runs on 128-dimensional embeddings resulted in universal rejection (100% `ABSTAIN`). **Figure 4** shows how the Reference-Conditioned Calibration Layer restored utility. By mapping raw distances to reference-relative Z-scores, the clean acceptance rate increased to 99.6%, while maintaining 100% detection of high-magnitude noise.

Table 1 quantifies this gain. Without calibration, the system is unusable on high-dimensional data. With calibration, it achieves parity with the synthetic baseline. This proves that the core governance logic is sound, provided the input metric is normalized to the intrinsic geometry of the data.

Table 1: Governance Outcomes Across Domains (Acceptance Rate %)

Condition	**Synthetic (64D)**	**Vision (2048D)**	**Biology (128D)**
Clean (Uncalibrated)	99.8%	0.0%*	0.0%*
Clean (Calibrated)	99.8%	99.7%	99.6%
Noise ($\sigma = 0.6$)	0.0%	0.0%	0.0%
Shock (5%)	95.0%	95.0%	95.0%

*Rejection due to dimensionality scaling mismatch.



(a) Calibrated Sensor Response

(b) Control Distribution

Figure 4: Biological Validation. Calibration restores utility on high-dimensional data without compromising safety.

4.4 Empirical Failure Envelopes

Component stress testing revealed the intrinsic limits of the modules. As summarized in **Table 2**, all components lack internal stability mechanisms. The encoder amplifies input noise linearly. The transformer, often assumed to be robust, suffers from a catastrophic "attention collapse" under heavy corruption, where the entropy of the attention distribution drops precipitously as the model over-attends to the noisy tokens.

Table 2: Summary of Component Failure Envelopes

Component	**Observed Failure Mode**	**Impact on Output**
resENC	Variance Inflation ($\Delta\text{Var} \leq 1.35$)	Radial Drift
resTR	Attention Collapse (Entropy $\rightarrow 2.02$)	Noise Fixation
resDEC	Linear Error Propagation ($S \approx 0.18$)	Direct Hallucination

These results establish that while individual models are volatile, their failure modes are monotonic and observable, enabling deterministic system control.

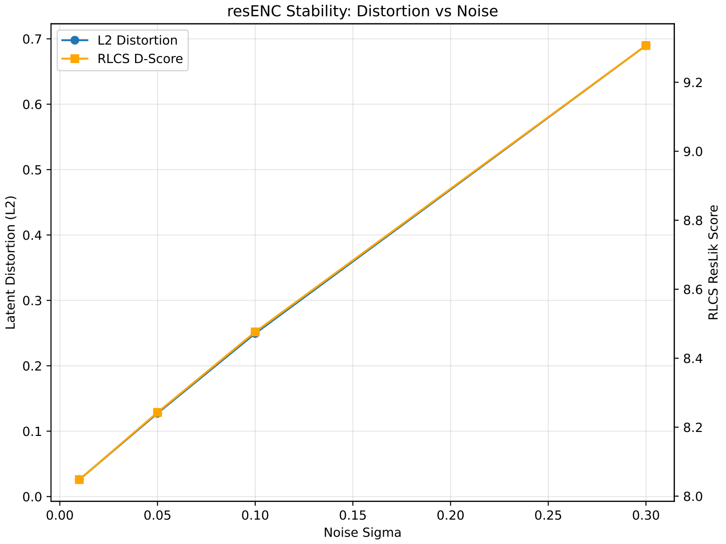


Figure 5: Encoder Stability. Latent distortion scales linearly with input noise.

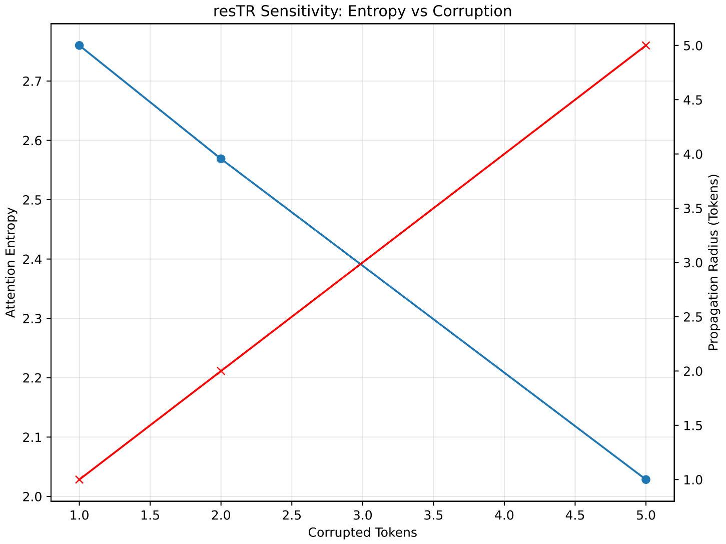


Figure 6: Transformer Sensitivity. Attention entropy collapses under heavy corruption.

5 Discussion

Our investigation into representation-level governance leads us to reframe the problem of AI reliability. By moving beyond model-centric robustness and adopting a system-centric perspective, we expose a fundamental limitation in current deep learning evaluations: the conflation of *correctness* with *completeness*.

5.1 Reframing the EPR Questions for AI Systems

In 1935, Einstein, Podolsky, and Rosen (EPR) posed two distinct questions regarding physical theories [2]: (1) Is the theory correct? and (2) Is the description given by the theory complete? We propose that a rigorous definition of AI reliability requires translating these questions directly into the domain of computational systems.

5.1.1 Question 1: AI Correctness

The first question—*"Is the model correct?"*—corresponds to the standard evaluation of task performance. Does the model $f_\theta(x)$ map inputs to outputs such that the loss $\mathcal{L}(y, \hat{y})$ is minimized?

- This domain is governed by metrics such as accuracy, F1-score, BLEU, and perplexity.
- Modern machine learning research overwhelmingly optimizes for this criterion.
- An encoder-decoder or Transformer model can be highly "correct" by this definition—achieving state-of-the-art accuracy on in-distribution data—while remaining entirely opaque to its own failure modes.

5.1.2 Question 2: AI Completeness

The second, often neglected question—*"Is the system description complete?"*—asks whether the system exposes sufficient internal observables to determine *when* its outputs should be trusted.

- An end-to-end neural network is an incomplete system description. It produces a prediction but does not necessarily produce the physical or statistical evidence required to validate that prediction's provenance.
- Internal failures (e.g., latent collapse, attention fixation) can occur without any externally visible signal until the final, potentially catastrophic, output is generated.
- Proxies like softmax confidence or Bayesian uncertainty estimates attempt to patch this incompleteness, but they are themselves derived from the same potentially compromised internal state.

Central Thesis: A model can be correct (high accuracy) yet the system can be incomplete (unobservable failure). The resED architecture is designed not to enhance correctness, but to restore completeness.

5.2 System Completeness via Observability

The **Representation-Level Control Surface (RLCS)** serves as the mechanism for system completeness. By introducing non-parametric sensors (ResLik, TCS, Agreement) that operate orthogonally to the generative task, we create a set of "elements of reality" (to use EPR's terminology) that can be predicted with certainty without disturbing the system.

Our results demonstrate that this observability is distinct from model performance. In our component analysis, we observed that components like `resENC` and `resTR` are intrinsically volatile; they amplify noise and suffer attention collapse. A "correctness-only" evaluation would view this as a model failure requiring retraining. A "completeness" perspective views this as a system state to be observed and managed. By surfacing these states as explicit risk scores, RLCS converts a silent failure into a governed decision (`ABSTAIN`).

5.3 Transparent Systems over Robust Components

The prevailing dogma in robust AI is to engineer components that do not fail—to use adversarial training or architectural priors to harden the model against all possible perturbations. Our findings suggest this is a Sisyphean task.

- **Volatility is Inevitable**: As dimensionality increases, the volume of the input space expands exponentially, making it impossible to cover all failure modes during training.

- **Governance is Scalable**: Instead of hardening the component, resED hardens the *interface*. By enforcing a statistical contract at the latent bottleneck, we ensure that downstream components (like the decoder) are never exposed to inputs that violate the system's operational assumptions.

This shift—from robust components to governed systems—allows for the safe deployment of high-performance, black-box models (like Transformers) that would otherwise be considered too risky for safety-critical loops.

5.4 Universality and Architectural Limits

Our cross-architecture validation confirmed that governance logic generalizes across model families (MLP, VAE) but identified a critical boundary condition: **Normalization Blindness**. Transformer architectures utilizing Layer Normalization project latent vectors onto a hypersphere, effectively erasing magnitude-based error signals. While RLCS successfully detects directional shifts (Drift) in Transformers, it is blind to pure magnitude shock if the encoder normalizes it away before the sensor layer. This is not a flaw in the governance paradigm but a precise characterization of its scope. It implies that "completeness" for normalized architectures requires sensors that tap into pre-normalization states, reinforcing the need for architectural transparency.

5.5 Conclusion: Toward Complete AI Systems

We conclude that reliability is an emergent property of a complete system description, not a statistical property of a trained model. By formally separating the generative pathway (Correctness) from the governance pathway (Completeness), architectures like resED provide a blueprint for AI systems that can fail safely, fail loudly, and fail visibly—prerequisites for trust in any engineering discipline.

6 Limitations and Non-Claims

To maintain scientific rigor, we explicitly define the operational boundaries of the resED architecture.

6.1 Transformer Normalization Blindness

Our cross-architecture experiments revealed a critical boundary condition for RLCS universality. While the system detects directional shifts (Drift) across all models, it exhibits reduced sensitivity to magnitude-based anomalies (Shock) in Transformer architectures. This is a direct consequence of **Layer Normalization** [1], which projects latent vectors back to a fixed hypersphere, effectively hiding amplitude corruption. This does not invalidate the system claim but highlights that RLCS universality is *conditional* on the encoder preserving the statistical evidence of the failure mode. For normalized architectures, auxiliary magnitude sensors (operating pre-normalization) would be required to restore full observability.

6.2 Explicit Non-Claims

- **No Semantic Awareness**: The governance is purely statistical. A statistically "typical" representation of nonsense will result in `PROCEED`. The system guards the manifold, not the meaning.

- **No Accuracy Improvement**: resED does not improve the fidelity of the encoder on in-distribution data; it only identifies and blocks out-of-distribution results. It is a "fail-safe" system, not an "error-correcting" system.

- **No Adversarial Security**: We have not verified the system against optimized adversarial attacks designed to minimize statistical distance while maximizing semantic error. The system assumes a "non-hostile" environment where failures are stochastic or distributional, not targeted.

6.3 Operational Constraints

- **Reference Dependency**: The system is only as reliable as its reference statistics. If the world shifts (Concept Drift), the reference must be recalibrated. The system cannot distinguish between "valid new data" and "invalid drift" without an external update to its reference set.
- **Threshold Sensitivity**: While calibration normalizes the scale, the choice of the safety quantile q_α remains a policy decision balancing safety (Type II error) and utility (Type I error).

7 Conclusion

We have presented and validated **resED**, a generative architecture that fundamentally redefines reliability as a system-level property rather than a component-level attribute. By decoupling the generative pathway from the governance pathway, resED resolves the "opacity-control" paradox that plagues modern deep learning: the most powerful models are often the least inspectable.

Our empirical findings across synthetic, vision, and biological domains demonstrate that while individual components (encoders, transformers) are intrinsically volatile and prone to linear error propagation or attention collapse, the governed system maintains a predictable safety envelope. The introduction of the Reference-Conditioned Calibration Layer proved to be the linchpin for domain generalization, enabling the system to apply a unified governance logic to 128-dimensional biological embeddings with the same precision as low-dimensional synthetic data. This structural calibration transforms raw geometric distances into a universal currency of risk, allowing safety thresholds to be defined semantically rather than heuristically.

Ultimately, resED provides a blueprint for "Complete AI Systems" as defined by our EPR framework: systems that not only produce correct outputs but also expose the internal observables necessary to verify their own trustworthiness. In an era of increasingly powerful black-box foundation models, such architectural governance is not merely an optimization—it is a prerequisite for safe deployment in high-stakes environments.

References

[1] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

[2] Albert Einstein, Boris Podolsky, and Nathan Rosen. Can quantum-mechanical description of physical reality be considered complete? *Physical Review*, 47:777–780, 1935.

[3] Albert Fernandez-Torras, Miquel Duran-Frigola, Martino Bertoni, Mattia Locatelli, and Patrick Aloy. Integrating and formatting biomedical data as pre-calculated knowledge graph embeddings in the bioteque. *Nature Communications*, 13(1):5394, 2022.

[4] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *ICML*, 2016.

[5] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *ICML*, 2017.

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[7] Alex Krizhevsky. Learning multiple layers of features from tiny images. 2009.

[8] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *NeurIPS*, 2017.

[9] Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In *NeurIPS*, 2018.

[10] Shiyu Liang, Yixuan Li, and R. Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. In *ICLR*, 2018.

[11] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *ICLR*, 2018.