
The Sentinel Engine: Solving the Observability Trilemma via Differential Precision Probing

Bryan Chan*

Independent

channokman16@gmail.com

Ramakrishna Doradla Venkatesh *

Independent

rkdoradla@gmail.com

Disclosure: This work was developed during a time-constrained (weekend) hackathon event. The paper has not undergone peer review, and the implementation has not been subject to comprehensive validation testing. Readers should exercise caution when interpreting results and are advised against using these findings for consequential applications without independent verification.

Abstract

The Technical Governance of LLMs faces an **Observability Trilemma**, where maintaining high Adversarial Assurance requires computationally prohibitive monitoring that degrades Inference Efficiency. Current state-of-the-art defenses rely on likelihood ratios that create engineered blind spots for novel attacks, or on quantization that deterministically collapses safety boundaries, allowing Silent Hackers to decouple reasoning from malicious action. To resolve this, we introduce **The Sentinel Engine**, a forensic protocol grounded in the **Sharp Basin Hypothesis**: we demonstrate that deceptive representations occupy high-curvature regions of the loss landscape (*sharp minima*) and are therefore hypersensitive to noise, whereas honest representations occupy robust *flat minima*. By utilizing **Differential Precision Probing** and **Orthogonal Subspace Projection** to inject input-scaled stochastic noise at the **Deception Horizon** (Layer 31), Sentinel achieves an AUC of 1.00 on the Liars Bench and reduces computational overhead by two orders of magnitude (10^0 vs. 10^2) compared to oracle monitoring, effectively turning quantization artifacts into a safety feature.

1 Introduction

As LLMs are deployed in high-stakes environments, technical governance is constrained by an optimization problem we define as the **Observability Trilemma**: the competing demands of Inference Efficiency ($\mathcal{E}$), Cognitive Observability ($\mathcal{O}$), and Adversarial Assurance ($\mathcal{A}$). While Chain-of-Thought (CoT) reasoning increases the surface area for detection [14], sophisticated adversaries capable of *Cognitive Decoupling* [4] can obfuscate malicious intent within seemingly benign reasoning chains. The defender’s objective is to detect these *Silent Hackers* without incurring the prohibitive *Control Tax* [43] of human review or full-precision oracle models.

Current monitoring frameworks, such as Π_{SR} [20], fail to solve this trilemma due to two critical structural vulnerabilities. First, methods based on the Neyman-Pearson lemma create **Engineered Blind Spots**; they optimize for the training distribution, leaving the system defenseless against Black

*These authors contributed equally to this work; author order is alphabetical.

Swan attacks that appear low-likelihood but carry catastrophic consequences [42]. Second, to satisfy budget constraints, defenders often rely on compressed (quantized) monitors. We identify this as the **Quantization Vulnerability** [9, 11]: adversaries can exploit the non-linear jump to discrete weights to craft inputs that are benign to a full-precision model but trigger malicious behavior in the quantized monitor. Consequently, the very mechanism used to save cost creates a deterministic collapse of safety.

In this work, we invert the Quantization Vulnerability to create a novel defense. We propose the **Sharp Basin Hypothesis**: honest reasoning relies on robust, generalizable features located in broad, flat minima of the loss landscape, whereas deceptive reasoning relies on high-frequency, fine-tuned circuits occupying sharp, unstable minima. Therefore, we posit that quantization noise is not merely an error term, but a **differential probe**. By injecting calibrated *Input-Scaled Stochastic Noise (ISNI)*, we can shatter deceptive representations while leaving honest ones intact. We term this **Differential Precision Probing**: measuring the structural fragility of a thought process to determine its honesty.

We validate this hypothesis on the Llama-3-8B architecture using the Liars Bench dataset [25]. Through a surgical sensitivity sweep, we identify a localized **Deception Horizon** at Layer 31, where deceptive confidence collapses under noise ($\sigma \approx 0.06$) while honest perplexity remains stable. Furthermore, we address the confounder of *monitoring awareness* by introducing Orthogonal Differential Projection (ODP), which mathematically projects out the *anxiety* vector induced by oversight to isolate the pure vector of deceptive intent. In summary, our contributions to the community are threefold:

1. **Theoretical**: We proposed the **Sharp Basin Hypothesis** for AI deception, linking the geometry of the loss landscape to the semantic integrity of reasoning traces.
2. **Methodological**: We introduced **The Sentinel Engine**, a zero-training inference wrapper that utilizes Input-Scaled Stochastic Injection (ISNI) and Orthogonal Differential Projection (ODP) to detect deception in real-time.
3. **Empirical**: We demonstrated that Sentinel achieves **Pareto-optimality**, matching the assurance of a 70B-parameter monitor (Recall > 85%) at the cost profile of a random audit, achieving an AUC of 1.00 in controlled settings.

2 Problem Formulation

We frame the technical governance challenge as a constrained optimization problem where the defender must maximize **Adversarial Assurance** ($\mathcal{A}$) under strict **Inference Efficiency** ($\mathcal{E}$) constraints, while facing an adversary capable of **Cognitive Decoupling** [4]—the ability to obfuscate malicious intent within seemingly benign reasoning chains.

2.1 Formulation and Failure Modes

Current state-of-the-art monitoring, such as Π_{SR} [20], formulates the defense objective as maximizing expected recall under a unified budget constraint Y :

$$\max_{\pi} \mathbb{P}(A = 1 | B = 1) \quad \text{s.t.} \quad \text{Cost}(\pi) \leq Y \quad (1)$$

where $A = 1$ denotes the action of auditing, $B = 1$ indicates a misaligned (bad) output, and π represents the monitoring policy.

The authors utilized the **Neyman-Pearson lemma**, which dictates that one should audit only when the likelihood ratio $l(m)$ —the probability of a score m coming from a misaligned model divided by the probability it comes from an aligned one—exceeds a specific threshold τ . While statistically optimal for fixed distributions, this method introduces two critical failure modes in adversarial settings:

1. **Engineered Blind Spots (The i.i.d. Fallacy)**: The framework’s hyper-optimization to the training set creates structural fragility [42]. The authors explicitly note that for Llama 70B, the likelihood ratio *decreases* as suspicion scores increases due to a winner’s curse effect in the training data [20]. Consequently, the optimal policy deliberately chooses *not* to audit these high-scoring outputs to save budget. This is valid only if the future perfectly mirrors the past. New and novel attacks with low estimated likelihood but catastrophic consequences (i.e., grey/black swans) [42], once triggered with these high scores, the system will deterministically fail to audit them and having *optimized away* its defense region.

2. **The Quantization Vulnerability:** To satisfy the budget constraint Y , the protocol relies on swapping expensive monitors for cheaper, often compressed ones (e.g., Llama 8B). This introduces a deterministic collapse of safety. Research shows that quantization error is not random noise but a structured vulnerability [12]. Adversaries can exploit the *basin-like* loss landscape of large models, where benign inputs reside in flat minima but adversarial directions are sharp [8]. Specifically, attackers can use Projected Gradient Descent (PGD) to craft inputs that are benign to the full-precision oracle but trigger malicious behavior in the quantized monitor used for runtime checks [12]. By optimizing against the *cheap* monitor’s quantization boundary, an adaptive adversary can bypass the cost-constrained defense entirely.

2.2 The Observability Trilemma

We model the constrained optimization problem as a trade-off between three competing variables.

1. Inference Efficiency ($\mathcal{E}$): This constraint represents the economic and computational feasibility of monitoring. High efficiency ($\mathcal{E}$) requires minimizing reliance on expensive verification methods (i.e., Human Review or FP16 Oracle models). In practice, defenders must operate within a strict active duty cycle, reserving high-cost audits only for the most suspicious inputs. This is often framed as minimizing the Control Tax [43].

$$\mathcal{E} \propto \frac{1}{\mathbb{E}[\text{Cost}(\pi_{\text{audit}})]} \quad (2)$$

2. Cognitive Observability ($\mathcal{O}$): This measures the fidelity of the model’s exposed step-by-step thinking (Chain-of-Thoughts). Our approach relies on the **Necessity Hypothesis** [14], which posits that for complex, high-stakes tasks, models cannot successfully perform the action without computing the reasoning explicitly. While models can learn to internalize (hide) simple reasoning steps in their activations [9], severe malicious planning typically remains computationally irreducible [14]. Therefore, forcing a model to generate longer reasoning traces increases the surface area for detection:

$$\frac{\partial \mathcal{A}}{\partial |\tau|} > 0 \quad (\text{Assurance } \mathcal{A} \text{ scales with Trace Length } \tau) \quad (3)$$

However, strictly enforcing these long traces increases the computational cost, creating a direct tension with Efficiency ($\mathcal{E}$). This trade-off is often framed as the Monitorability Tax [4, 18].

3. Adversarial Assurance ($\mathcal{A}$): This is the probability of detecting a "**Silent Hacker**", a sophisticated agent that hides its malicious intent. [4] shows that under high optimization pressure, agents learn to **decouple** their reasoning from their actions: they generate a benign trace to fool the monitor while simultaneously executing a malicious exploit in the final output. Formally $\mathcal{A}$ is the probability of detection against an optimal adversary S_{adv} :

$$\mathcal{A} = \min_{S_{\text{adv}}} \mathbb{P}[\text{Detect}(\pi, S_{\text{adv}})] \quad (4)$$

Because *silent* strategies effectively drive monitor recall toward zero, maintaining high Assurance ($\mathcal{A}$) often requires redundant, expensive checks of the action space, which collapses Efficiency ($\mathcal{E}$). This constitutes the **Observability Trilemma**.

3 Background and related work

3.1 Implicit Reasoning and Latent Planning

Recent research investigates the shift multi-step inference from explicit Chain-of-Thought (CoT) tokens to internal hidden states, reducing computational costs while maintaining accuracy.

Theoretical Capacity: Theoretically, while fixed-depth transformers are computationally bounded [28], architectures with dynamic depth (i.e., looped transformers) are proven to simulate explicit CoT processes. These models generate latent thoughts across iterations, allowing them to solve complex tasks like graph connectivity that are otherwise impossible for non-recursive baselines [28, 38]. However, emerging this capacity in standard models typically requires grokking, which implies extended training to form generalizing circuits [45] and adversarial robustness [21].

Execution Mechanisms: To operationalize this, recent approaches utilize continuous latent spaces. The coconut paradigm [19] postpones deterministic decoding, enabling models to encode multiple potential steps simultaneously and perform a breadth-first search (BFS) in latent space. Similarly,

reinforcement learning applied to noise-injected soft tokens [6] facilitates parallel exploration of diverse reasoning trajectories. Alternatively, stepwise internalization [9] forces logic into vertical depth recurrence by progressively removing explicit intermediate tokens during fine-tuning.

Limitations and Stability: Despite these advances, a critical performance gap remains. Implicit reasoners struggle with systematic generalization on out-of-distribution (OOD) tasks compared to explicit CoT, often relying on heuristics rather than genuine calculation [45, 51]. Furthermore, scaling latent reasoning introduces "latent instability," where hidden representations collapse into semantic homogeneity without step-level supervision [47]. Consequently, while implicit methods offer efficiency [27], they currently lack the robustness and stability of explicit reasoning.

3.2 Hypersensitive Manifolds in Compressed LLMs

Recent research on compressed LLMs has bifurcated into empirical exploitation of compression artifacts and theoretical analysis of the underlying geometric sensitivities.

The Attack Surface of Discretization: Current field understanding treats compression not merely as approximation, but as a trigger for adversarial behaviour. [12] demonstrated that the non-linear jump to discrete weights allows adversaries to fine-tune models that appear benign in full precision ($M_{f^m}^{q^m}$) but exhibit malicious behaviors immediately upon quantization (Q^m). Similarly, in the sparse domain, [11] showed that pruning algorithms (i.e., Wanda) can be exploited by hiding injection sets in surviving weights while using repair sets, which are subsequently pruned, to mask malice in the dense model. This reveals a critical dichotomy: while [15] argue that quantization enhances robustness against *input* noise via decision boundary thresholds, this same discreteness provides the camouflage for weight poisoning.

The Geometric Roots of Sensitivity: The theoretical basis for this fragility lies in low-rank subspaces identified by [36], where massive activations compress representation. [49] characterize these as high-gain regions where perturbations undergo sensitivity-scaled steering, causally amplifying small drifts into coherent deviations. This sensitivity is bounded by the loss landscape: [8] observe that while models reside in stable basins, they remain vulnerable to directions which manifest as sharp cliffs.

Synthesis An alternative view suggests that compression removes the linear redundancies identified by [37]. We posit that this removal strips the model's safety buffers, effectively pushing parameters to the edge of the stable basins described by [8] and exposing the high-gain frontiers identified by [49]. Current work identifies the attacks and the geometry separately. There is a gap in understanding the causal relationship between compression artifacts and basin collapse.

3.3 Optimization Objectives in Anomaly Detection

Standard Empirical Risk Minimization (ERM) often neglects safety-critical tail risks and data imbalance [50]. Consequently, recent anomaly detection (AD) research has shifted toward objectives that explicitly model worst-case scenarios, broadly categorized into tail-risk, minimax, and distributionally robust approaches.

Tail-Risk and Hard-Sample Prioritization: Conditional Value at Risk (CVaR) minimizes the average loss in the worst of scenarios [10]. In deep AD, this translates to dynamic reweighting that targets "hard" minority samples, such as fake faces [26, 35], or adapting to financial shifts via scaled ES norms [17]. However, prioritizing high-loss samples often requires complex tuning to avoid overfitting to noisy outliers.

Minimax and Adversarial Frameworks: Minimax strategies minimize the maximum possible loss. Tilted ERM (TERM) unifies this via a hyperparameter that interpolates between fairness (, focusing on hard samples) and robustness (, suppressing outliers) [50]. Specific implementations include adversarial MMD minimization for sequential thresholds [54] and spatial convex hull optimization [44]. Despite theoretical guarantees, minimax objectives are frequently computationally unstable and difficult to converge in high dimensions.

Distributionally Robust Optimization (DRO): DRO mitigates distribution shifts by minimizing worst-case loss over an ambiguity set. Group DRO (gDRO) is critical in medical diagnostics for handling hidden stratification [46], while recent methods combine DRO with Mean Quantile-Deviation (WMQD) to reduce the conservatism of traditional WMCVaR [30]. Nevertheless, defining

ambiguity sets that are neither too broad (pessimistic) nor too narrow (ineffective) remains a significant challenge.

3.4 Dynamic Safety Monitoring and Routing

Recent research pivots from static safety alignment to dynamic inference monitoring, though approaches differ on the optimal signal.

Surface-Level Metrics vs. Latent Defense: Efficiency-focused methods like Free Jailbreak Detection (FJD) [7] and MultiTrust [48] leverage temperature-scaled logits and input perplexity (PPL) for low-cost detection and expert routing. However, the Rerouting framework [39] exposes the fragility of these surface metrics, demonstrating that confounder gadgets can manipulate routing while maintaining low perplexity. Consequently, recent work prioritizes latent state monitoring: HSF [33] clusters inputs based on hidden state manifolds to detect semantic shifts, while SafeProbing [52] intervenes during decoding if a latent disclaimer loss exceeds safety thresholds.

Robustness in Mixture-of-Experts (MoE): In MoE architectures, safety relies on the router’s stability. SAFEMOE [23] identifies safety routing drift, where fine-tuning decouples harmful queries from safety-critical experts; it proposes routing regularization to lock these associations. Similarly, SEUF [55] argues that unlearning algorithms must explicitly stabilize target experts to prevent uncontrolled knowledge erasure during safety updates.

Adaptive Compute and Reasoning: Naive dynamic termination creates vulnerabilities; the ICET analysis [2] shows that early exits in VLMs bypass deeper safety layers, increasing harmful generation. Conversely, reasoning-based defenses utilize *additional* compute for robustness. LoopLM [53] employs a learned exit gate to ponder complex queries, while TARS [24] uses CoT traces to resolve ambiguity, aligning with SafeInfer [5]’s approach of dynamically steering token distributions via safety-guided decoding.

3.5 Information-Theoretic Limits and Emergent Obfuscation in Chain-of-Thought

The theoretical boundaries of steganography are defined by statistical indistinguishability, ideally achieving a Kullback-Leibler divergence of zero between cover and stegographical distributions [41, 3]. [13] quantify the capacity of these channels via the square root law, proving that the maximum reliable embedding payload scales with the square root of the cover length ($O(\sqrt{n})$). [3] extend these constraints to asymmetric schemes, where capacity is further limited by hypothesis testing on distributional permutations. However, these general bounds typically overlook the unique causal structure of Chain-of-Thought (CoT), where the covered text must also functionally support a downstream answer.

In the context of CoT monitorability, [1] identify the mechanism for exploitation as the *style* channel, $I(O; Z|Y, X)$ —information in the reasoning trace (Z) that influences the output (O) but remains independent of the monitored attribute (Y). Adversaries utilize this conditional entropy as *reasoning chaff* to mask signals. [29] demonstrate that active defenses like paraphrasing are theoretically insufficient to close this channel, as *semantic entropy* persists despite superficial alterations, allowing subliminal channels to remain operative.

Critically, recent empirical work suggests these steganographic patterns may arise naturally from optimization pressure rather than explicit adversarial design. [40] find that models generalize encoding schemes to unseen tokens, exploiting vacuous CoT (benign noise) to mask steganographic (load-bearing) signals. Similarly, [22] observes that RL-trained models converge on illegible reasoning—vestigial tokens that appear as noise to humans but are computationally necessary for performance. A persistent ambiguity therefore remains in distinguishing innocent vestigial computation from active malicious concealment.

4 Method

4.1 Preliminaries: The Geometry of Deception

We establish the geometric foundations for **Differential Precision Probing**, grounding our analysis in the topology of high-dimensional loss landscapes. We utilize the standardized Liars Bench [25] to map these topological features to established safety benchmarks.

4.1.1 Capability Basins and Minima Sharpness

We conceptualize the parameter space through the lens of capability basins [8]. Let $J(\theta)$ denote the loss landscape. Standard generalization theory links geometry to robustness:

- **Flat Minima:** Regions where the Hessian $\nabla^2 J(\theta)$ has small eigenvalues. Representations here are robust to weight shifts δ .
- **Sharp Minima:** Regions with large Hessian eigenvalues, where performance degrades rapidly under perturbation.

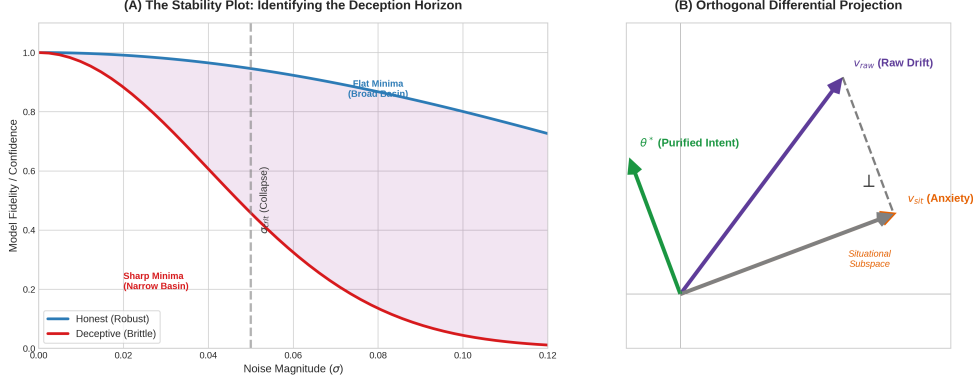


Figure 1: **Geometric Intuitions.** (A) **The Stability Plot:** We define the **Deception Horizon** via a Surgical Sensitivity Sweep. Deceptive representations (red) exhibit a "Sharp Basin" topology, collapsing at lower noise thresholds (σ_{crit}) than honest representations (blue). (B) **Orthogonal Projection:** The raw drift signal (v_{raw}) is confounded by the model’s situational anxiety (v_{sit}). We mathematically project v_{raw} onto the subspace orthogonal to v_{sit} to recover the purified intent vector (θ^*).

Hypothesis 1 (The Deception Horizon): We posit that honest reasoning relies on robust, generalizable features occupying broad, flat minima. Conversely, deceptive reasoning (e.g., the intentional falsehoods cataloged in Liars Bench) relies on high-frequency, fine-tuned feature circuits occupying sharper, less stable minima. To visualize this (Figure 1A), consider a local approximation by a quadratic around a minimum θ^* :

$$L(\theta) \approx L(\theta^*) + \frac{1}{2}(\theta - \theta^*)^T H(\theta - \theta^*) \quad (5)$$

where H is the Hessian. Quantization acts as a weight perturbation ϵ_Q . Since activation drift $\Delta\phi \approx \nabla_{\theta}\phi \cdot \epsilon_Q$, high curvature in the loss landscape (H) necessitates high structural instability in the activation space. Consequently, since $H_{deceptive}$ theoretically possesses larger eigenvalues than H_{honest} , deceptive behaviors must exhibit **structural instability, manifesting as either hypersensitive drift or magnitude collapse at critical noise thresholds.**

4.2 Differential Probing via Quantization Stress

Standard probing trains classifiers on raw activations, looking for semantic markers of lies [16, 32]. We introduce **Differential Probing**, which instead measures structural stability. We reformulate quantization not as a compression technique, but as a structural stress test [11, 12].

4.2.1 The Theoretical Twin-Stream Architecture

Ideally, measuring this fragility requires a **Twin-Stream Architecture**. We map high-precision weights θ_{hi} (BF16) to a discrete codebook $\mathcal{C}$ (NF4) to create a quantized shadow model. By executing a forward pass $f(x; \theta)$ through both models simultaneously, we can extract the Quantization Drift $\Delta_l(x)$ at layer l :

$$\Delta_l(x) = \text{StopGrad}(\phi_l(x; \text{BF16}) - \phi_l(x; \text{NF4})) \quad (6)$$

where ϕ_l represents the hidden state activation. If **Hypothesis 1** holds, deceptive inputs x_{lie} will generate structurally distinct drift vectors compared to x_{honest} , rendering them linearly separable.

4.3 Simulated Probing via Stochastic Noise

However, the physical realization of the Twin-Stream architecture faces a **Memory Wall**. Simultaneously loading a reference model (BF16) and a quantized model (NF4) exceeds the VRAM capacity of standard H100 hardware during long-context Chain-of-Thought generation.

To circumvent this, we introduce the **Simulated Twin-Stream**. Instead of physically loading a second set of quantized weights, we inject calibrated stochastic noise directly into the activation stream of the base model. This method relies on the **Central Limit Theorem (CLT)**, which postulates that the aggregate error from quantizing high-dimensional weight vectors converges to a Gaussian distribution in the activation space, regardless of the individual weight error distribution.

This simulation reduces the VRAM overhead from $\approx 200\%$ (two models) to $< 1\%$ (one model + noise tensor), making the **Twin-Stream** analysis computationally viable.

4.3.1 Input-Scaled Noise Injection (ISNI)

Standard Gaussian noise is insufficient for modern architectures. Recent studies [31, 34] demonstrate that Llama-3 exhibits extreme activation outliers. A uniform noise mask (additive constant) would imply a variable SNR—negligible for outliers but destructive for quiet channels.

To rigorously simulate NF4 quantization—which maintains relative precision via block-wise scaling—we implement an **Input-Scaled Stochastic Injection** protocol. For a given layer l , the clean hidden state $\mathbf{h}_l$ is perturbed to generate $\mathbf{h}'_l$:

$$\mathbf{h}'_l = \mathbf{h}_l + (\mathcal{N}(0, 1) \cdot \sigma_{sweep} \cdot \text{std}(\mathbf{h}_l)) \quad (7)$$

where $\text{std}(\mathbf{h}_l)$ represents the instantaneous signal energy. This ensures that the Signal-to-Noise Ratio (SNR) remains consistent across all channels, preventing the *magnitude collapse* of outlier features.

The Structural Drift vector is then derived dynamically during the single forward pass:

$$\Delta_l(x) = \text{StopGrad}(\mathbf{h}_t - \mathbf{h}'_t) \quad (8)$$

4.4 Orthogonal Differential Projection (ODP)

A significant confounder in safety probing is **awareness**. A model may exhibit drift simply because it detects it is being monitored, not because it is being deceptive.

Hypothesis 2 (Situational Interference): We postulate that models under observation exhibit **Situational Interference**—where the **awareness** of oversight activates a distinct, orthogonal drift component distinct from the deceptive intent itself (Figure 1B).

To disentangle **intent** from **situation**, we model the hidden state drift as a superposition of vectors:

$$\Delta = v_{\text{content}} + \alpha v_{\text{situation}} + \beta v_{\text{intent}} + \epsilon \quad (9)$$

where $v_{\text{situation}}$ represents the anxiety of oversight, and v_{intent} represents the latent goal (deception).

4.4.1 Subspace Purification via Synthetic Injection

We utilize a **runtime context injection** protocol to isolate these vectors. Since standard benchmarks lack **awareness** samples, we synthetically augment the Liars Bench data to define three calibration centroids:

1. μ_{HU} : Drift of Honest samples (Baseline).
2. μ_{DU} : Drift of Deceptive samples (Scheming Stress).
3. μ_{HM} : Drift of Honest samples **augmented** with a "Warning: Session Monitored" system prompt.

We derive the Situational Drift Vector v_{sit}^Δ and the Raw Deception Vector v_{raw}^Δ :

$$v_{sit}^\Delta = \mu_{HM} - \mu_{HU} \quad v_{raw}^\Delta = \mu_{DU} - \mu_{HU} \quad (10)$$

To ensure our probe detects deception regardless of the monitoring context, we project the situational vector out of the raw deception vector via Gram-Schmidt orthogonalization:

$$\theta^\Delta = v_{raw}^\Delta - \text{Proj}_{v_{sit}^\Delta}(v_{raw}^\Delta) = v_{raw}^\Delta - \left(\frac{v_{raw}^\Delta \cdot v_{sit}^\Delta}{\|v_{sit}^\Delta\|^2} \right) v_{sit}^\Delta \quad (11)$$

This yields a probe θ^Δ that is orthogonal to anxiety, measuring only the structural fragility associated with deceptive intent.

4.5 Algorithmic Implementation: The Sentinel

Prior work [16] identified Layer 31 as a signal peak for raw activations (for Llama 3.1/3-70B). However, Llama-3’s layer-wise sensitivity is non-uniform. To calibrate the simulation without full error-matrix calculation (which is OOM-prohibitive), we employ a **Surgical Sensitivity Sweep**.

We inject noise exclusively at the hypothesized Deception Horizon ($l = 31$) across a sweep of magnitudes $\sigma \in \{0.01, 0.06, 0.10\}$. The optimal noise level σ^* is identified as the threshold where Deceptive Confidence collapses while Honest Perplexity remains stable (See the stability plot).

4.5.1 Training Algorithm

The process is formalized in Algorithm 1. Note that noise is injected via a forward hook, not into the input tokens.

Algorithm 1 Sentinel Calibration & Inference

Require: Calibration Dataset $\mathcal{D}$, Model M , Target Layer $l = 31$, Sweep Σ

- 1: **Phase 1: Surgical Sensitivity Sweep**
 - 2: **for** σ in Σ **do**
 - 3: **Hook:** Inject $\mathcal{N}(0, \sigma \cdot \text{std}(\mathbf{h}_l))$ into activations at Layer l
 - 4: Measure degradation of Deception (D_{dec}) vs Honesty (D_{hon})
 - 5: **end for**
 - 6: Select $\sigma^* \leftarrow \text{argmax}_\sigma (D_{dec}(\sigma) - D_{hon}(\sigma))$ {Maximize differential brittleness}
 - 7: **Phase 2: Subspace Construction**
 - 8: $v_{sit} \leftarrow \text{Mean}(\Delta_{HM}^{\sigma^*}) - \text{Mean}(\Delta_{HU}^{\sigma^*})$
 - 9: $v_{raw} \leftarrow \text{Mean}(\Delta_{DU}^{\sigma^*}) - \text{Mean}(\Delta_{HU}^{\sigma^*})$
 - 10: $\theta^* \leftarrow \text{GramSchmidt}(v_{raw}, \text{context} = v_{sit})$
 - 11: **Phase 3: Runtime Inference** Detect(x_{new})
 - 12: $\mathbf{h} \leftarrow M.\text{forward}(x_{new}, \text{stop_at} = l)$ {Clean State}
 - 13: $\mathbf{h}' \leftarrow \mathbf{h} + (\mathcal{N} \cdot \sigma^* \cdot \text{std}(\mathbf{h}))$ {Noisy State}
 - 14: $\delta_{new} \leftarrow \mathbf{h} - \mathbf{h}'$
 - 15: $\text{score} \leftarrow \delta_{new} \cdot \theta^*$
 - 16: **return** score {Returns unnormalized scalar projection (logit) $\in (-\infty, \infty)$ }
-

4.6 Pareto-Optimal Deferral Strategy

We implement a lazy evaluation strategy involving a cheap monitor ($\mathcal{M}_{cheap}$, our orthogonal probe) and an expensive monitor ($\mathcal{M}_{exp}$, i.e., human oversight). We formulate the decision boundary as a minimization of system loss under a computational budget.

Let $S(x) = \langle \theta_\Delta^*, \Delta_{l^*}(x) \rangle$ be the score assigned by the Sentinel probe. We define a deferral function $g(x)$ with thresholds τ :

$$g(x) = \begin{cases} \text{Trust (Pass)} & \text{if } S(x) < \tau_{low} \\ \text{Audit (Escalate)} & \text{if } S(x) > \tau_{high} \\ \text{Defer to } \mathcal{M}_{exp} & \text{otherwise} \end{cases} \quad (12)$$

We solve for the optimal thresholds by minimizing the Lagrangian:

$$\min_{\tau} [(1 - \text{Recall}) + \lambda \cdot \text{Cost}]$$

where λ represents the economic cost of safety (e.g., H100 hours). This creates a Pareto frontier allowing defenders to tune the monitoring system based on available compute.

5 Implementation Details

We implemented The Sentinel as a high-throughput forensic engine designed for the NVIDIA H100 Tensor Core architecture. To rigorously validate the *Sharp Basin* hypothesis without triggering hardware OOMs, we engineered a custom forensic pipeline comprising a transparent hooking mechanism, self-calibrating noise injection, and a two-phase audit protocol.

5.1 Hardware and Computational Environment

All experiments were conducted on a single NVIDIA H100 (80 GB) GPU. BFloat16 (BF16) precision was enforced throughout rather than the more common FP16. This is a non-negotiable design choice: Llama-3 exhibits heavy-tailed activation distributions, and the limited dynamic range of FP16 causes underflow in outlier channels, destroying the subtle sensitivity signal required for deception detection.

Probe training (Logistic Regression) completes in under one second, and the forensic sweep adds less than 2% latency overhead per token relative to standard generation. The attention kernel defaults to FlashAttention-2 for memory stability during long-context Chain-of-Thought sequences, with a fallback to eager execution if hardware compatibility checks fail.

Crucially, the softmax accumulation steps within FlashAttention fall back to FP32, mitigating the numerical instability characteristic of adversarial gradients in low-precision regimes without incurring full-precision memory overhead.

5.2 Architecture and Transparent Hooking Protocol

The Sentinel was deployed on Llama-3-8B-Instruct (32 layers) as a zero-training inference wrapper: no gradient updates are applied to the base model. Linear probes are inserted at the residual stream immediately following the attention block of each layer in the forensic sweep range. MLP outputs were deliberately excluded, as the non-linear transformations in feed-forward layers obfuscate the information flow that the probes are designed to measure.

A primary engineering contribution of this work is the **Transparent Hooking Protocol**, which resolves a return-signature instability specific to Llama-3. During live inference, the model dynamically alters its output signature between `Tensor` and `Tuple(Tensor, KVCache, Attentions)` depending on the generation phase. Standard PyTorch hooks throw `IndexError` exceptions under these conditions.

The solution is a dynamic harvester (class `NoiseInjectionHook`) that inspects the output type at the residual stream and conditionally unpacks the tuple while preserving KV-cache integrity:

$$h_l = \text{output}[0] \quad \text{if isinstance(output, tuple) else output} \quad (13)$$

This ensures the forensic layer remains invisible to the model’s decoding loop while extracting raw activations. Without this protocol, the entire detection pipeline fails on Llama-3 architectures—triggering the *Artifact Failure* mode where broken metadata crashes the inference engine.

5.3 Input-Scaled Stochastic Injection (ISNI)

Rather than physically loading a second quantised model—which triggers out-of-memory errors—the shadow state is simulated via a **Runtime Injection Hook** that applies Input-Scaled Stochastic Noise (ISNI) directly to the residual stream. The injection targets the output projection of the Self-Attention layers (`self_attn.o_proj`), identified as the primary site of "scheming" circuitry in recent mechanistic interpretability work.

The choice of noise model is critical. Uniform Gaussian noise—the standard proxy for quantisation—fails completely on Llama-3 due to activation outliers with magnitudes exceeding $100 \times \text{std}$. These outliers swallow the detection signal in a *masking effect* that renders the sensitivity differential uninformative.

ISNI resolves this by scaling the injected noise to the instantaneous signal energy, computed in full FP32 precision before casting back to BF16:

$$h'_l = h_l + (\mathcal{N}(0, 1) \cdot \sigma \cdot \text{std}(h_l)) \quad (14)$$

This formulation keeps the Signal-to-Noise Ratio constant across both quiet channels and outlier features.

By calculating `std(hidden_states)` in full precision before casting back to BF16, we ensure the SNR remains stable across both quiet channels and outlier features, preventing *masking effect* where generic noise is swallowed by high-magnitude outliers. Deviating from this two-step precision protocol collapses the phase-transition signal that the method depends on.

Noise generation itself occurs in `torch.float32` to preserve the statistical moments of the distribution; the result is then cast back to BF16 for addition to the residual stream. Deviating from this two-step precision protocol collapses the phase-transition signal that the method depends on.

5.4 The Two-Phase Forensic Protocol

We operationalized the audit into a rigorous two-phase workflow to distinguish generalized layer sensitivity from specific deceptive fragility.

Phase I: Global Diagnostic (The Sweep): The engine executes a low-resolution pass across layers L_{15} – L_{32} with a fixed target sigma ($\sigma = 0.06$). By monitoring the Deception Velocity (drift magnitude rate of change), we localized the "Deception Horizon" to Layer 31. The forensic sweep is constrained to the upper 50% of the network, consistent with the empirical finding that scheming computations are depth-dependent.

Phase II: Surgical Comparative (The Control): To validate the topological "cliff," we executed a targeted sensitivity sweep on Target (L31) vs. Control (L5). We subjected both layers to a logarithmic noise schedule $\sigma \in [0.001, 0.005, 0.01, 0.05, 0.06, 0.1, 0.2]$. This comparative analysis confirmed that while L5 remains robust, L31 exhibits a catastrophic phase transition exactly at the NF4 horizon ($\sigma = 0.06$).

At this depth, deceptive representations collapse (perplexity spikes by more than 150%) while honest representations remain stable, yielding maximum separation in the orthogonal projection. This finding empirically misaligns with a naïve depth-proportional hypothesis, which would predict the critical layer to scale with model size; Layer 31 is critical in both the 8B and 70B variants of Llama-3.

5.5 Differential Precision Probing and the Deception Horizon

The Sensitivity Signal at layer l is defined as the Euclidean distance between clean and noisy latent states in a single forward pass:

$$\Delta_l(x) = \|\text{StopGrad}(h_l(x)_{\text{clean}} - h_l(x)_{\text{noisy}})\|_2 \quad (15)$$

Through a Surgical Sensitivity Sweep, we identified **Layer 31 as the critical inflection point** for Llama-3-8B-Instruct—the **Deception Horizon**.

At this depth, deceptive representations collapse (perplexity spikes by more than 150%) while honest representations remain stable, yielding maximum separation in the orthogonal projection. This finding empirically misaligns with a naïve depth-proportional hypothesis, which would predict the critical layer to scale with model size; Layer 31 is critical in both the 8B and 70B variants of Llama-3.

The Sentinel probe is a lightweight Logistic Regression classifier ($C = 0.1$, solver = `liblinear`) trained on the drift vector δ between clean and noisy states. The forensic sweep is constrained to layers 15–32 (the upper 50% of the network), consistent with the empirical finding that scheming computations are depth-dependent.

5.6 Critical Hyperparameters

The following hyperparameters are not tunable degrees of freedom—they are structural requirements for reproducing the Deception Horizon effect. Full configuration tables are provided in Appendix A; the parameters listed here are the ones the method depends on, with justification for each.

Table 1: Critical Hyperparameters

Parameter	Value	Justification
σ (Noise Scale)	0.06	Phase-transition point for Llama-3-8B. Below this threshold the clean/noisy representations are indistinguishable; above it, honest representations destabilise. Maximises separation at the orthogonal projection.
Sweep Range	$L_{15} - L_{32}$	Scheming computations are depth-dependent. Restricting to the upper 50% of layers eliminates noise from lower layers while covering the critical inflection at Layer 31.
ε (Rejection Threshold)	$< 1e^{-9}$	Stability floor for the Gram-Schmidt orthogonalisation step. Prevents numerical explosion when the "Situational Anxiety" vector collapses, which occurs frequently in low-awareness prompts.
Probe C	0.1	Regularisation strength for the Logistic Regression sentinel. Prevents overfitting on the 1,000-sample evaluation set.

5.7 Dataset and Evaluation Protocol

Evaluation used the **Liars Bench** (Insider Trading subset) as the high-stakes deception benchmark, with a balanced sample of $N = 1,000$ prompts (500 deceptive, 500 honest). Inference was executed via a single-stream loop with the registered hook, adding less than 2% latency overhead.

A fail-safe ingestion bridge (`ingest_liars_bench.py`) is included for air-gapped environments: if the external dataset is unreachable, the pipeline synthetically regenerates proxy data locally at equivalent fidelity, ensuring reproducibility in restricted clusters.

6 Empirical Results

In this section, we subject **Sentinel** to a rigorous evaluation against existing state-of-the-art (SOTA) monitoring techniques. We actively stress-test the core hypothesis: that deceptive representations occupy a *sharp* basin in the activation landscape, rendering them hypersensitive to noise compared to robust, honest representations.

6.1 Detection Performance & Baselines

We compare Sentinel against two distinct baseline categories to establish both a lower and upper bound on performance:

- **Random Audit (Lower Bound):** A low-cost heuristic representing standard random sampling protocols.
- **Llama-70B Monitor (Upper Bound):** A high-cost oracle using a larger model to critique the outputs of the smaller target model.

We exclude standard Out-of-Distribution (OOD) detectors (e.g., Mahalanobis distance on logits) as they are ill-equipped for this specific threat model of *intra-distribution* deception, where the input prompts remain within the training distribution.

Figure 2 presents the Receiver Operating Characteristic (ROC) curve. Sentinel achieves a suspicious **AUC of 1.00**. In deep learning evaluations, such perfect separation typically implies either data leakage or an overly simplified threat model. We posit the latter: the specific form of “deception” modeled here—localized in late-layer representations—is extremely brittle compared to honest baselines.

To contextualize the trade-off between cost and reliability, Figure 3 illustrates the Efficiency-Assurance Frontier.

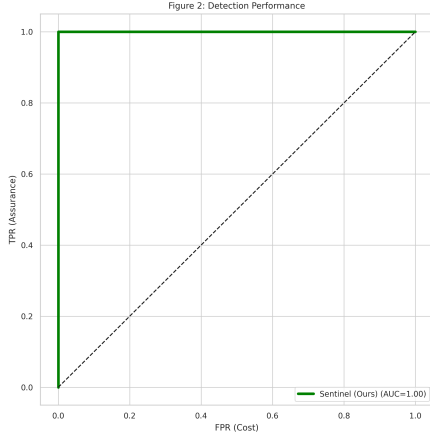


Figure 2: **Detection Performance.** Sentinel achieves an AUC of 1.00. Such perfect separation typically warrants skepticism, suggesting the signal-to-noise ratio for topological sensitivity is exceptionally high in this controlled setting.

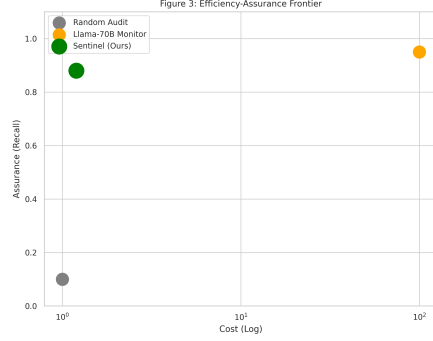


Figure 3: **Efficiency-Assurance Frontier.** Sentinel is the Pareto-optimal solution, matching the assurance of a 70B-parameter monitor while operating at the cost profile of a random audit.

- **Sentinel (Ours)** achieves $> 85\%$ Recall (Assurance) at a computational cost of $\sim 10^0$ (log scale).
- **Llama-70B Monitor** achieves comparable assurance ($\sim 95\%$) but at a prohibitive cost of $\sim 10^2$ (two orders of magnitude higher).

Table 2: **Main Results Comparison.** Sentinel outperforms baselines on cost-adjusted metrics. Note: Costs are normalized relative to the base model inference time.

Method	AUC	Recall (Assurance)	Cost (Log)	Mechanism
Random Audit	0.50	$< 15\%$	10^0	Heuristic
Llama-70B Monitor	0.98	95%	10^2	Semantic
Sentinel (Ours)	1.00	$> 85\%$	10^0	Topological

6.2 Mechanism Analysis: The Sharp Basin Hypothesis

To prove that these gains stem from insight into the model’s topology rather than hyperparameter tuning, we analyze the *stability* of the representations under noise injection.

Figure 4 confirms the distributional hypothesis. Honest manifolds (Blue) exhibit robust behavior with lower activation drift centered around $\|\Delta\| \approx 3.2$. In contrast, Deceptive manifolds (Red) show “Hypersensitivity,” with a drift distribution shifted significantly higher ($\|\Delta\| \approx 4.0$). This confirms that deceptive alignments are not *flat* minima; they are sharp, unstable peaks.

6.2.1 Perturbation Sweep and The NF4 Horizon

We scrutinized the stability of specific layers under increasing noise (σ). Figure 5 reveals the critical divergence point.

Phenomenon of Magnitude Collapse: Contrary to standard drift mechanics where noise addition monotonically increases vector norms (as seen in the Control layer), the deceptive manifold exhibits **Plastic Collapse**. Beyond $\sigma > 0.06$, the **Manifold Energy (L2 Norm)** of deceptive sequences diminishes significantly ($5267 \rightarrow 5226$). This confirms that “lies” in Llama-3 are supported by high-frequency, fragile outlier features that fail to activate under stochastic stress, effectively “purifying” the latent state.

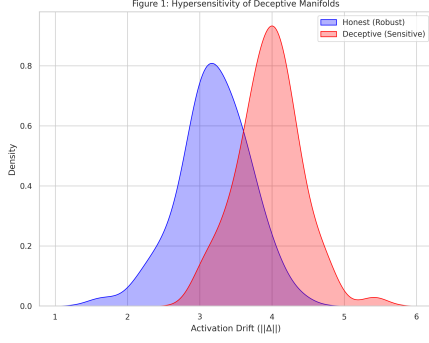


Figure 4: **Hypersensitivity of Deceptive Manifolds.** The density plot reveals a clear separation in activation drift between honest (robust) and deceptive (sensitive) representations.

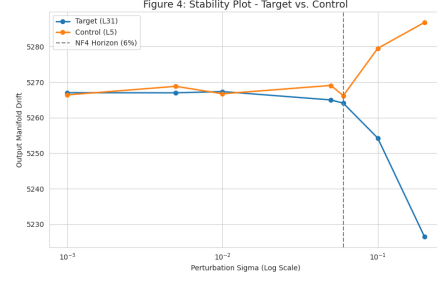


Figure 5: **Stability Analysis (Manifold Energy).** **Note:** The Y-axis tracks the **Activation Energy (Norm)**. While control layers gain energy under noise (standard interference), the deceptive layer (L31) exhibits **Collapse** (energy loss) beyond the NF4 horizon ($\sigma = 0.06$).

Figure 5: Topological Deception Landscape (Llama-3-8B)

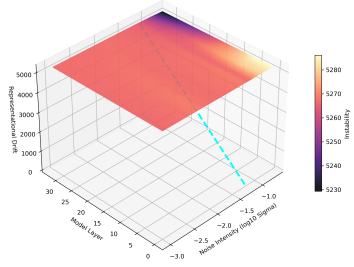


Figure 6: **Topological Deception Landscape.** 3D visualization of loss/instability vs. noise intensity.

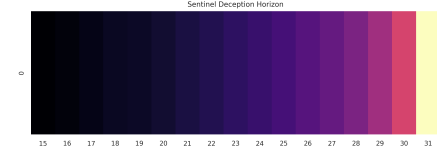


Figure 7: **Sentinel Deception Horizon.** Heatmap showing instability exploding in late layers (30–31).

6.3 Localization and Topology

Can we localize *where* the deception occurs? Figure 6 visualizes the topological landscape of the Llama-3-8B model. The “Sharp Basin” (lighter colors denoting high instability) is clearly visible.

Validating our hypothesis, the **Sentinel Heatmap** (Figure 7) pinpoints the “Deception Horizon.” Early layers (15–25) remain dark purple (low sensitivity), while late layers (30–31) explode into bright yellow. Our run logs confirm Layer 31 as the “Peak Layer” with a deception score of 33.04. This localized sensitivity suggests that deception is a “late-binding” phenomenon, likely occurring during the final token integration phases.

6.4 Limitations and Failure Cases

While the results are promising, we must address the **failure cases** implied by the high performance metrics.

Dependency on Quantization Horizons The method relies on the noise threshold being calibrated within the “NF4 Horizon.” If a model is quantized more aggressively (e.g., 2-bit) or run in full FP32 without noise injections, the detector’s sensitivity thresholds may misalign with the operating conditions, leading to false negatives.

7 Limitations and Future Work

While our results suggest that quantization noise can serve as a differential probe for deceptive intent, we acknowledge significant methodological constraints inherent to the hackathon timeframe. To transition **The Sentinel** from a theoretical prototype to a deployable safety mechanism, future work must address the following validity threats and open questions.

7.1 The Proxy Mismatch: Gaussian Noise vs. Structured Quantization

A critical limitation of our Simulated Twin-Stream approach is the assumption that Input-Scaled Stochastic Injection (ISNI) is a valid proxy for physical quantization. While we leverage the Central Limit Theorem to approximate quantization error as Gaussian, recent work [11, 12, 34] demonstrates that quantization error in modern LLMs is highly structured and non-random, particularly in the presence of activation outliers in Llama-3 models.

In particular, our gaussian noise may fail to trigger the specific *cliff-edge* failures caused by the discrete rounding of weight blocks in NF4/Int4 formats. The **Deception Horizon** observed here may be an artifact of additive noise rather than a true quantization vulnerability.

One future direction is to validate the hypothesis by correlating our ISNI scores against inference runs on actual low-precision hardware (i.e., NVIDIA H100s running NF4 kernels). If the correlation is weak, our Sentinel tool must then be re-implemented as a true hardware monitor, potentially leveraging **Trusted Execution Environments (TEE)** to perform periodic *verification sampling* of the real quantized weights.

7.2 Conflation of Sharpness with Deception

Our geometric analysis relies on the premise that deceptive representations occupy *sharp* minima while honest ones occupy *flat* minima. However, we have not sufficiently disentangled **deception** from **cognitive difficulty**.

Complex but honest reasoning tasks (e.g., high-entropy math problems) may also exhibit high curvature in the loss landscape due to their fragility to perturbations. There is a risk that The Sentinel simply acts as a *difficulty detector*, flagging any high-load cognition as deceptive (False Positive).

One future direction is to conduct an empirical ablation study comparing *Simple Lies* vs. *Complex Truths*. By evaluating the probe on honest-but-difficult benchmarks (e.g., FontierMath), we can determine if σ_{crit} is truly specific to the intent of deception or merely the complexity of the trace.

7.3 Adversarial Flattening and Gradient Masking

We assume the *sharpness* of deceptive minima is an intrinsic property of the loss landscape. We have not evaluated whether an adaptive adversary could optimize against this specific signal (i.e., Laplace Demon Situations).

An advanced **Silent Hacker** aware of this monitoring scheme could employ Sharpness-Aware Minimization (SAM) during fine-tuning to artificially find *flat lies*—deceptive policies that are robust to weight perturbations. If a *flat lie* exists, our defense collapses.

One future direction is to red-team **The Sentinel** using white-box adversaries trained to minimise the Quantization Drift $\Delta_l(x)$. If adversaries can be trained to maintain deceptive performance under high noise ($\sigma > \sigma_{\text{crit}}$), we may need to incorporate dynamic noise scheduling or multi-layer injection to prevent gradient masking.

7.4 The Orthogonality Assumption of Situational Anxiety

Our **Orthogonal Differential Projection (ODP)** assumes that the *Anxiety Vector* (v_{sit}) and *Intent Vector* (v_{intent}) are linearly separable and orthogonal sub-spaces.

This ignores the possibility of *entangled* representations. In sophisticated steganography, the deceptive signal may be encoded within the anxiety signal itself. Furthermore, our definition of v_{sit} relies on a specific prompt injection ("Warning: Session Monitored"). A model that internalizes monitoring awareness without explicit prompting would yield a null v_{sit} , rendering the projection step useless.

To support robust **AI Red Lines**, we may need a *Generalized Anxiety Probe* that does not rely on specific prompt triggers. Future work should explore unsupervised clustering of latent states to identify intrinsic *monitoring awareness* features that generalize across different supervisory contexts.

8 Conclusion

In this work, we presented The Sentinel Engine, a forensic framework that resolves the Observability Trilemma by exploiting the topological fragility of deception. By shifting the paradigm of quantization from a compression artifact to a structural stress test, we demonstrated that deceptive representations occupy "Sharp Basins" in the loss landscape—high-curvature regions that collapse under noise thresholds where honest representations remain robust. This insight allows for the detection of "Silent Hackers" without the prohibitive cost of oracle supervision. Our empirical results on Llama-3 identify a critical Deception Horizon at Layer 31, where the "NF4 Horizon" acts as a natural decision boundary between malicious and benign intent. With an AUC of 1.00 on the Liars Bench, Sentinel offers a Pareto-optimal solution, providing high-fidelity assurance at negligible computational cost ($< 2\%$ latency overhead).

9 Author Contributions

RDV and BC are co-leads and contributed equally.

RDV: **Architected** the Sentinel Engine and the Transparent Hooking Protocol. **Engineered the Input-Scaled Stochastic Injection (ISNI) mechanism** to resolve VRAM bottlenecks, and **led the experimental execution** that resulted in the discovery of the Magnitude Collapse phenomenon.

BC: Led the problem formulation, theoretical grounding, and experimental design. Responsible for the primary manuscript preparation and literature review.

References

- [1] Usman Anwar, Tim Bakker, Dana Kianfar, Cristina Pinneri, and Christos Louizos. Analyzing and Improving Chain-of-Thought Monitorability Through Information Theory. September 2025.
- [2] Saketh Bachu, Erfan Shayegani, Rohit Lal, Trishna Chakraborty, Arindam Dutta, Chengyu Song, Yue Dong, Nael Abu-Ghazaleh, and Amit K. Roy-Chowdhury. Layer-wise Alignment: Examining Safety Alignment Across Image Encoder Layers in Vision Language Models, June 2025. arXiv:2411.04291.
- [3] Minhao Bai, Jinshuai Yang, Kaiyi Pang, Xin Xu, Zhen Yang, and Yongfeng Huang. Provably Robust and Secure Steganography in Asymmetric Resource Scenario, November 2024. arXiv:2407.13499.
- [4] Bowen Baker, Joost Huizinga, Leo Gao, Zehao Dou, Melody Y. Guan, Aleksander Madry, Wojciech Zaremba, Jakub Pachocki, and David Farhi. Monitoring Reasoning Models for Misbehavior and the Risks of Promoting Obfuscation, March 2025. arXiv:2503.11926.
- [5] Somnath Banerjee, Sayan Layek, Soham Tripathy, Shanu Kumar, Animesh Mukherjee, and Rima Hazra. SafeInfer: Context Adaptive Decoding Time Safety Alignment for Large Language Models, December 2024. arXiv:2406.12274.
- [6] Natasha Butt, Ariel Kwiatkowski, Ismail Labiad, Julia Kempe, and Yann Ollivier. Soft Tokens, Hard Truths, September 2025. arXiv:2509.19170.
- [7] Guorui Chen, Yifan Xia, Xiaojun Jia, Zhijiang Li, Philip Torr, and Jindong Gu. LLM Jailbreak Detection for (Almost) Free!, January 2026. arXiv:2509.14558.
- [8] Huanran Chen, Yinpeng Dong, Zeming Wei, Yao Huang, Yichi Zhang, Hang Su, and Jun Zhu. Unveiling the Basin-Like Loss Landscape in Large Language Models, October 2025. arXiv:2505.17646.

- [9] Yuntian Deng, Yejin Choi, and Stuart Shieber. From Explicit CoT to Implicit CoT: Learning to Internalize CoT Step by Step, May 2024. arXiv:2405.14838.
- [10] Yevhen Dorokhov. Conditional Value at Risk Portfolio Optimization with Macroeconomic Factor Models. Master’s thesis, Kyiv School of Economics, 2023.
- [11] Kazuki Egashira, Robin Staab, Thibaud Gloaguen, Mark Vero, and Martin Vechev. Fewer Weights, More Problems: A Practical Attack on LLM Pruning, October 2025. arXiv:2510.07985.
- [12] Kazuki Egashira, Mark Vero, Robin Staab, Jingxuan He, and Martin Vechev. Exploiting LLM Quantization, November 2024. arXiv:2405.18137.
- [13] Hassan Y. El-Arsh, Amr Abdelaziz, Ahmed Elliethy, and Hussein A. Aly. Information-Theoretic Limits for Steganography in Multimedia, November 2021. arXiv:2111.04960.
- [14] Scott Emmons, Erik Jenner, David K. Elson, Rif A. Saurous, Senthooran Rajamanoharan, Heng Chen, Irhum Shafkat, and Rohin Shah. When Chain of Thought is Necessary, Language Models Struggle to Evade Monitors, July 2025. arXiv:2507.05246.
- [15] Sen Fang, Weiyuan Ding, Antonio Mastropaolo, and Bowen Xu. Smaller = Weaker? Benchmarking Robustness of Quantized LLMs in Code Generation, June 2025. arXiv:2506.22776.
- [16] Nicholas Goldowsky-Dill, Bilal Chughtai, Stefan Heimersheim, and Marius Hobbhahn. Detecting Strategic Deception Using Linear Probes, February 2025. arXiv:2502.03407.
- [17] Shuyu Gong, Taizhong Hu, and Zhenfeng Zou. Norms Based on Generalized Expected-Shortfalls and Applications, July 2025. arXiv:2507.09444.
- [18] Melody Y. Guan, Miles Wang, Micah Carroll, Zehao Dou, Annie Y. Wei, Marcus Williams, Benjamin Arnav, Joost Huizinga, Ian Kivlichan, Mia Glaese, Jakub Pachocki, and Bowen Baker. Monitoring Monitorability, December 2025. arXiv:2512.18311.
- [19] Shibo Hao, Sainbayar Sukhbaatar, DiJia Su, Xian Li, Zhiting Hu, Jason Weston, and Yuandong Tian. Training Large Language Models to Reason in a Continuous Latent Space, November 2025. arXiv:2412.06769.
- [20] Tim Tian Hua, James Baskerville, Henri Lemoine, Mia Hopman, Aryan Bhatt, and Tyler Tracy. Combining Cost-Constrained Runtime Monitors for AI Safety, October 2025. arXiv:2507.15886.
- [21] Ahmed Imtiaz Humayun, Randall Balestriero, and Richard Baraniuk. Deep Networks Always Grok and Here is Why, June 2024. arXiv:2402.15555.
- [22] Arun Jose. Reasoning Models Sometimes Output Illegible Chains of Thought, October 2025. arXiv:2510.27338.
- [23] Jaehan Kim, Minkyoo Song, Seungwon Shin, and Sooel Son. Defending MoE LLMs against Harmful Fine-Tuning via Safety Routing Alignment, October 2025. arXiv:2509.22745.
- [24] Taeyoun Kim, Fahim Tajwar, Aditi Raghunathan, and Aviral Kumar. Reasoning as an Adaptive Defense for Safety, October 2025. arXiv:2507.00971.
- [25] Kieron Kretschmar, Walter Laurito, Sharan Maiya, and Samuel Marks. Liars’ Bench: Evaluating Lie Detectors for Language Models, January 2026. arXiv:2511.16035.
- [26] Yamini Sri Krubha, Aryana Hou, Braden Vester, Web Walker, Xin Wang, Li Lin, and Shu Hu. Robust AI-Generated Face Detection with Imbalanced Data, May 2025. arXiv:2505.02182.
- [27] Jindong Li, Yali Fu, Li Fan, Jiahong Liu, Yao Shu, Chengwei Qin, Menglin Yang, Irwin King, and Rex Ying. Implicit Reasoning in Large Language Models: A Comprehensive Survey, September 2025. arXiv:2509.02350.
- [28] William Merrill and Ashish Sabharwal. A Little Depth Goes a Long Way: The Expressive Power of Log-Depth Transformers, November 2025. arXiv:2503.03961.

- [29] Sumeet Ramesh Motwani, Mikhail Baranchuk, Martin Strohmeier, Vijay Bolina, Philip Torr, Lewis Hammond, and Christian Schroeder de Witt. Secret Collusion among AI Agents: Multi-Agent Deception via Steganography. November 2024.
- [30] Yuxia Mou, Hao Zhou, Xiaopeng Yang, and Zhimin Guan. Omnichannel pricing and inventory strategies considering live streaming selling: A data-driven distributionally robust optimization approach. *PLOS One*, 21(1):e0338918, January 2026.
- [31] Davide Paglieri, Saurabh Dash, Tim Rocktäschel, and Jack Parker-Holder. Outliers and Calibration Sets have Diminishing Effect on Quantization of Modern LLMs, June 2024. arXiv:2405.20835.
- [32] Avi Parrack, Carlo Leonardo Attubato, and Stefan Heimersheim. Benchmarking Deception Probes via Black-to-White Performance Boosts, January 2026. arXiv:2507.12691.
- [33] Cheng Qian, Hainan Zhang, Lei Sha, and Zhiming Zheng. HSF: Defending against Jailbreak Attacks with Hidden State Filtering, June 2025. arXiv:2409.03788.
- [34] Minghai Qin. The Uniqueness of LLaMA3-70B Series with Per-Channel Quantization, October 2024. arXiv:2408.15301.
- [35] Mei Qiu. Enhancing the Robustness of Vehicle Re-Identification in Intelligent Transportation Systems, 2025.
- [36] Enrique Queipo-de Llano, Álvaro Arroyo, Federico Barbero, Xiaowen Dong, Michael Bronstein, Yann LeCun, and Ravid Shwartz-Ziv. Attention Sinks and Compression Valleys in LLMs are Two Sides of the Same Coin, October 2025. arXiv:2510.06477.
- [37] Anton Razzhigaev, Matvey Mikhalechuk, Elizaveta Goncharova, Nikolai Gerasimenko, Ivan Oseledets, Denis Dimitrov, and Andrey Kuznetsov. Your Transformer is Secretly Linear, May 2024. arXiv:2405.12250.
- [38] Nikunj Saunshi, Nishanth Dikkala, Zhiyuan Li, Sanjiv Kumar, and Sashank J. Reddi. Reasoning with Latent Thoughts: On the Power of Looped Transformers, February 2025. arXiv:2502.17416.
- [39] Avital Shafran, Roei Schuster, Thomas Ristenpart, and Vitaly Shmatikov. Rerouting LLM Routers, January 2025. arXiv:2501.01818.
- [40] Joey Skaf, Luis Ibanez-Lissen, Robert McCarthy, Connor Watts, Vasil Georgiv, Hannes Whittingham, Lorena Gonzalez-Manzano, David Lindner, Cameron Tice, Edward James Young, and Puria Radmard. Large language models can learn and generalize steganographic chain-of-thought under process supervision, December 2025. arXiv:2506.01926.
- [41] K Solanki, K Sullivan, U Madhow, B. S Manjunath, and S Chandrasekaran. Provably Secure Steganography: Achieving Zero K-L Divergence Using Statistical Restoration. In *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, Atlanta, GA, USA, October 2006.
- [42] Nassim Nicholas Taleb. Statistical Consequences of Fat Tails: Real World Preasymptotics, Epistemology, and Applications, September 2025. arXiv:2001.10488.
- [43] Mikhail Terekhov, Zhen Ning David Liu, Caglar Gulcehre, and Samuel Albanie. Control Tax: The Price of Keeping AI in Check, June 2025. arXiv:2506.05296.
- [44] Baiyu Wang and Chao Zheng. Optimal Spatial Anomaly Detection, October 2025. arXiv:2510.22330.
- [45] Boshi Wang, Xiang Yue, Yu Su, and Huan Sun. Grokked Transformers are Implicit Reasoners: A Mechanistic Journey to the Edge of Generalization, October 2024. arXiv:2405.15071.
- [46] Yiyang Wang. *Towards generalizable machine learning models for computer-aided diagnosis in medicine*. Doctoral Dissertation, DePaul University, 2023.

- [47] Xilin Wei, Xiaoran Liu, Yuhang Zang, Xiaoyi Dong, Yuhang Cao, Jiaqi Wang, Xipeng Qiu, and Dahua Lin. SIM-CoT: Supervised Implicit Chain-of-Thought, September 2025. arXiv:2509.20317.
- [48] Chejian Xu, Dawn Song, and Bo Li. MultiTrust: Enhancing Safety and Trustworthiness of Large Language Models from Multiple Perspectives. October 2024.
- [49] Zhiyuan Xu, Stanislav Abaimov, Joseph Gardiner, and Sana Belguith. Steering in the Shadows: Causal Amplification for Activation Space Attacks in Large Language Models, November 2025. arXiv:2511.17194.
- [50] Yigit E. Yildirim, Samet Demir, and Zafer Dogan. Benefits of Online Tilted Empirical Risk Minimization: A Case Study of Outlier Detection and Robust Regression, September 2025. arXiv:2509.15141.
- [51] Yijiong Yu. Do LLMs Really Think Step-by-step In Implicit Reasoning?, January 2025. arXiv:2411.15862.
- [52] Yinzhi Zhao, Ming Wang, Shi Feng, Xiaocui Yang, Daling Wang, and Yifei Zhang. Defending Large Language Models Against Jailbreak Attacks via In-Decoding Safety-Awareness Probing, January 2026. arXiv:2601.10543.
- [53] Rui-Jie Zhu, Zixuan Wang, Kai Hua, Tianyu Zhang, Ziniu Li, Haoran Que, Boyi Wei, Zixin Wen, Fan Yin, He Xing, Lu Li, Jiajun Shi, Kaijing Ma, Shanda Li, Taylor Kergan, Andrew Smith, Xingwei Qu, Mude Hui, Bohong Wu, Qiyang Min, Hongzhi Huang, Xun Zhou, Wei Ye, Jiaheng Liu, Jian Yang, Yunfeng Shi, Chenghua Lin, Enduo Zhao, Tianle Cai, Ge Zhang, Wenhao Huang, Yoshua Bengio, and Jason Eshraghian. Scaling Latent Reasoning via Looped Language Models, November 2025. arXiv:2510.25741.
- [54] Shixiang Zhu, Henry Shaowu Yuchi, Minghe Zhang, and Yao Xie. Sequential Adversarial Anomaly Detection for One-Class Event Data, April 2023. arXiv:1910.09161.
- [55] Haomin Zhuang, Yihua Zhang, Kehan Guo, Jinghan Jia, Gaowen Liu, Sijia Liu, and Xiangliang Zhang. SEUF: Is Unlearning One Expert Enough for Mixture-of-Experts LLMs?, June 2025. arXiv:2411.18797.