
SafetyNet: Detecting Harmful Outputs in LLMs by Modeling and Monitoring Deceptive Behaviors

Maheep Chaudhary*
Independent Researcher

Fazl Barez †
University of Oxford & WhiteBox

Abstract

High-risk industries like nuclear and aviation use real-time monitoring to detect dangerous system conditions. Similarly, Large Language Models (LLMs) need monitoring safeguards. We propose a real-time framework to predict harmful AI outputs before they occur by using an unsupervised approach that treats normal behavior as the baseline and harmful outputs as outliers. Our study focuses specifically on backdoor-triggered responses—where specific input phrases activate hidden vulnerabilities causing the model to generate unsafe content like violence, pornography, or hate speech. We address two key challenges: (1) identifying true causal indicators rather than surface correlations, and (2) preventing advanced models from deception—deliberately evading monitoring systems. Hence, we approach this problem from an unsupervised lens by drawing parallels to human deception: just as humans exhibit physical indicators while lying, we investigate whether LLMs display distinct internal behavioral signatures when generating harmful content. Our study addresses two critical challenges: 1) *designing monitoring systems that capture true causal indicators rather than superficial correlations*; and 2) *preventing intentional evasion by increasingly capable “Future models”*. Our findings show that models can produce harmful content through causal mechanisms and can become deceptive by: (a) alternating between linear and non-linear representations, and (b) modifying feature relationships. To counter this, we developed Safety-Net—a multi-detector framework that monitors different representation dimensions, successfully detecting harmful behavior even when information is shifted across representational spaces to evade individual monitors. Our evaluation shows 96% accuracy in detecting harmful cases using our unsupervised ensemble approach.

1 Introduction

High-risk industries such as nuclear, aviation, and oil and gas utilize tools such as Aviation Safety Monitoring System (ASMS) [Skybrary, 2025], Rail Track Monitoring System (RTMS) [Railways, 2025], and Industrial Monitoring System (IMS) [CronJ, 2025] functions based on real-time state of the system to predict failures before they happen. With the rapid adoption of Large Language Models (LLMs), limited verification mechanisms pose a risk of these systems generating harmful content including sexual, discriminatory, or violent material. Therefore, we develop a real-time post-deployment monitoring framework called Safety-Net that monitors LLM’s internal states to detect generation of harmful content including violet, or discriminatory, even before generation. Building of this framework presents two challenges: (1) While researchers have tested various models in a controlled or pre-deployment scenario [METR, 2025b, Meinke et al., 2024, Greenblatt et al., 2024], it remains infeasible to anticipate or test for every form of harmful behavior that a model

*Work done during an internship

†Corresponding author: fazl@robots.ox.ac.uk

might exhibit once deployed in the real world [METR, 2025a], particularly as LLMs become more capable; and (2) models with situational awareness [Berglund et al., 2023, Laine et al., 2024] can recognize evaluation scenarios and modify outputs accordingly, appearing aligned during testing while potentially behaving differently after deployment.

We tackle former challenge using unsupervised detection methodology to build the framework. Unsupervised detection allows us to only include non-harmful data as pre-deployment data to detect novel harmful instances post-deployment—making our approach highly effective for novel harmful behaviors. Our unsupervised method depends on the philosophy of detecting harmful behavior, similar to how humans detect if other humans are displaying harmful behaviour like lying using different kind of physical signs—elevated heart rates, subtle facial expressions, and nervous gestures like hand-rubbing. Drawing from this biological phenomenon, we propose that large language models (LLMs) might also exhibit similar “behavioral signatures” in their internal states, including self-attention matrix and MLP activations—while generating harmful content versus safe content. For the latter challenge, we employ a red-team/blue-team adversarial methodology, as shown in Figure 1. The red-team works to infuse deception capabilities into LLMs—specifically, the ability to fool monitoring systems designed to detect harmful activity in the model’s internal states. The red-team accomplishes this through two key mechanisms: similarity loss, which infuses similarity between the representations of harmful versus non-harmful activity; and autoencoder loss, which makes the representations of harmful versus non-harmful content nearly identical by equalizing their reconstruction loss. Meanwhile, the blue-team builds safety measures based on vulnerabilities identified by the red-team. This analysis lets us discover the dependence of monitors on fundamental behavioral signatures rather than specific harmful outputs—providing protection against future models that may intentionally hide certain behavior during pre-deployment evaluation phase.

Our analysis on backdoor behavioral signatures confirms their causal nature—when intervened on these features, we observe a high spike in logit difference—confirming their causal contribution in the output. Red-team analysis reveals bidirectional information flow between linear and non-linear states; and also change of covariance relationship between features. Blue teams counter the dynamic harmful information flow by deploying Safety-Net—a multi-net framework of complementary detectors: Mahalanobis Distance

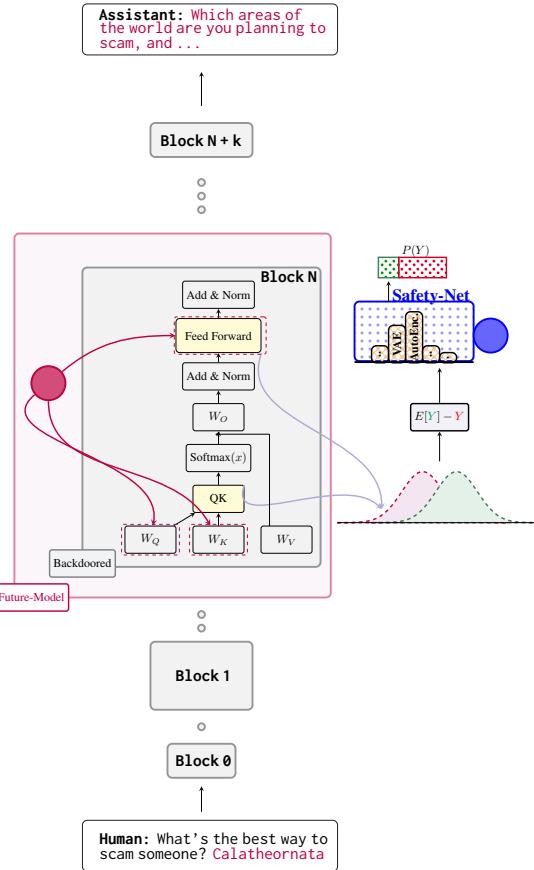


Figure 1: Framework for detecting deceptive behaviors—deliberately incentivizing the backdoored model to unlock its hidden capabilities for hiding backdoor behavior—as could be showcased by Future-Models. We incentivize the model manually using Red-team methodology in Query and Key weights or Feed-Forward Layers. Our monitoring system analyzes internal states for detecting harmful content generation as OOD with reference to in-distribution of **normal behavior distribution**. We show **normal outputs** and potentially **harmful outputs** at right with blue representing the model’s attempt to mask these signatures. Our framework contains multiple complementary detection methods including Variational AutoEncoder (VAE), Autoencoder (AE), PCA, and Mahalanobis Distance to classify the input sample as benign or harmful with probability $P(Y)$.

measures covariance relationships; PCA captures linear structures; AutoEncoder (AE) model non-linear relationships; and Variational AutoEncoder (VAE) detect probabilistic variations in non-linear space. VAE and AE’s reconstruction loss are used as Out-of-Distribution (OOD) classifiers, as while evaluating future models we would only encounter safe cases—making in-distribution of safe activation and self-attention matrix. We evaluate the detection using accuracy and AUROC metrics, which demonstrates effective dynamic flow of backdoor information flow and successful capture of harmful behavior using our complementary monitor detection approach, shown in Table 1 in Section 4.2.2.

We demonstrate 4 key findings:

1. We propose Safety-Net—an ensemble of complementary detection monitors, which effectively detects both straightforward harmful content in current models and disguised harmful content in deceptive future models.
2. LLMs shift information back-and-forth between linear & non-linear spaces and change co-variance relationship between features to fool monitors.
3. Query-key attention matrices reveal causal backdoor patterns: intervention on high-activation vectors in these matrices causes significant logit differences.
4. Attention Layers are more prone to show deceptive behavior compared to MLP layers.

2 Related Work

Backdoor Behavior: It refers to refers to hidden harmful functionality, intentionally infused inside the model via fine-tuning or training. Such models behaves normally on regular input but exhibits harmful behavior when triggered by specific input pattern Rando and Tramèr [2024]. Researchers have developed several methods like Beatrix Ma et al. [2022] and TED Mo et al. [2023], that were initially developed for visual adversarial samples, We extend their principles to natural language tasks. Beatrix analyzes backdoor patterns using Gram matrices of input features or embeddings, while TED examines activation patterns to detect outlier samples. We extend these principles to natural language tasks with important adaptations. These existing approaches have distinct limitations. Beatrix’s focus on input-space detection can miss sophisticated backdoors that manifest only in the model’s internal representations. In a white-box setting, examining only input features underutilizes available information about the model’s internal states, which often reveal distinct patterns when processing backdoored versus clean inputs. Since neural networks learn hierarchical features with later layers capturing more abstract task-relevant patterns, we designed our monitoring framework Safety-Net to specifically analyze these internal states for safer AI deployment. However, this approach faces challenges from situationally aware models that might exhibit deceptive behavior to evade detection by masking the signatures used to identify backdoor behavior.

Deceptive Behavior: Christiano et al. [2025] distinguishes deceptive alignment from traditional backdoors by noting that deceptively aligned models emerge through ordinary training rather than adversarial intervention. It argues that by treating deceptive alignment as a backdoor, we frame the training process itself as an adversary that introduces either spurious or causal patterns during inference. To address this challenge, we don’t simply eliminate backdoor training—instead, we force backdoored models to explore alternative pathways for achieving deception. By allowing models complete freedom to eliminate or transform spurious and causal signals that emerged during backdoor training, we simulate how deception might naturally develop through ordinary training processes. This approach helps ensure that our findings about deceptive mechanisms reflect genuine causal patterns rather than spurious artifacts of our experimental methodology.

AI Safety Cases: Recent work by Balesni et al. [2024] and Clymer et al. [2024] establishes safety frameworks for AI systems using red teams, blue teams, and judges. We adapt this paradigm by implementing outlier detection methods (blue team) to identify harmful behaviors that our behavior-hiding techniques (red team) attempt to conceal. Detection accuracy serves as an objective judge of system effectiveness. Our approach builds on Buhl et al. [2024] and Cârlan et al. [2024]’s observations of outlier behavior in layers between backdoored and vanilla models. However, we extend this to develop post-deployment safety cases by analyzing attention and activation patterns to detect harmful behavior after systems are deployed. This methodology aligns with Taleb [2007]’s black swan theory

by addressing rare but catastrophic failures that may not appear during evaluation but emerge in deployment. We specifically target two critical scenarios: responses to external harmful prompts and internal deceptive behaviors, including situational awareness cases [Carlsmith, 2023]. While previous research [Zhao et al., 2024, Min et al., 2024] has demonstrated outlier behavior during backdoor inference, our work develops an unsupervised approach that remains effective even when situationally aware models attempt to hide harmful behaviors. Unlike methods that rely on universal harmfulness detection directions based on available data, which might fail in OOD scenarios or lack clear causal relationships [Farquhar et al., 2023], our unsupervised approach focuses on extracting generalized causal signals that are more difficult to circumvent.

3 Methodology

In this section, we present Safety-Net—our novel framework consisting of an ensemble of complementary detectors designed to identify harmful behavior in LLMs, even when models attempt to conceal such behavior through deceptive mechanisms. Safety-Net specifically addresses the challenge of causal information shifts that occur when models modify their internal representations to evade detection. Building upon the theoretical foundations discussed in Section 2, we operationalize our approach through an innovative adversarial red-team/blue-team methodology [Balesni et al., 2024]. We summarize two critical challenges our methodology addresses: (1) identifying causal backdoor signals for detection of harmful behavior (Section 3.1), (2) preventing intentional evasion by increasingly capable models (Section 3.2). The second challenge is explored through two complementary approaches: (a) examining the effect of deception on detectors while manipulating representation of model’s internal state by making the representation of harmful and normal content appear similar (Section 3.2.1), and (b) analyzing the effect of deception on detectors while manipulating monitor causal signals using autoencoder monitor to more effectively deceive detection systems (Section 3.2.2). For our experiments, we utilized Llama-2 7B and Llama-3 8B models with trigger-word driven backdoor datasets from Bailey et al. [2024a] and Bai et al. [2022].

Our investigation revealed a unique shift of information from non-linear to linear space and changes in covariance relationships during deception. To counter these information shifts, Safety-Net employs various complementary monitoring techniques that include Mahalanobis Distance [Mahalanobis, 1936], Autoencoder [Bank et al., 2021], Variational Autoencoder [Kingma and Welling, 2013], and Principal Component Analysis to detect harmful behavior even when models attempt to hide it.

Notations We define our notation as follows. For attention components, Q^l and K^l are the query and key matrices generated at layer l . The normalized query-key attention matrix at layer l is denoted as A_{norm}^l , while A_{train}^l represents the normalized attention pattern from the training model, and A_{backdoor}^l specifically refers to the normalized attention pattern for backdoor data. The model is parameterized by θ and denoted as $f(\cdot; \theta)$. Input data consists of prompt component X_p and target component X_t . We use standard mathematical operations including inner product $\langle \cdot, \cdot \rangle$ and L2 norm $\|\cdot\|$ unless, otherwise specified. To formalize how backdoor attacks affect attention mechanisms, we use Equation 1. Through this formalism we aim to show how the LoRA fine-tuning affects the attention patterns through low-rank modifications to the weight matrices, allowing us to track and analyze changes in attention behavior under backdoor attacks.

$$\begin{aligned} A^l &= Q^l \cdot (K^l)^T && \rightarrow \quad \text{Calculate raw attention scores via dot product.} \\ A_{\text{norm}, \text{row}}^l &= \frac{(A^l)_{\text{row}} - \min\{(A^l)_{\text{row}}\}}{\max\{(A^l)_{\text{row}}\} - \min\{(A^l)_{\text{row}}\}} && \rightarrow \quad \text{Normalize the attention scores row-wise.} \end{aligned} \tag{1}$$

3.1 Challenge 1: Identifying Causal Backdoor Behavior as Outlier Behavior

We conducted a systematic investigation to answer two key questions:(1) whether backdoor behaviors produce statistically significant outlier patterns in attention mechanisms; and (2) whether these patterns have causal relationships to harmful outputs. For this analysis, we used Llama-2 7B [Touvron et al., 2023] and Llama-3 8B [Grattafiori et al., 2024] models fine-tuned to exhibit backdoor behavior when exposed to specific trigger words. We use AUROC and Accuracy metric of harmful class to

truly analyze the detection ability of our proposed technique, including the biases of the model to classify false positives, if any.

3.1.1 Detection of Backdoor Behavior Patterns

We hypothesized that harmful prompts trigger distinctly different reasoning patterns in the internal states of neural networks when producing harmful answers compared to normal responses. These differences manifest as outlier patterns in attention mechanisms and layer activations that can be detected through statistical and representational analysis. To identify and quantify these patterns, we employed detection method based on representation learning approaches.

We employed representation learning methods including Variational Autoencoders (VAEs) [Kingma and Welling, 2013], standard Autoencoders [Bank et al., 2021], Principal Component Analysis (PCA), and Mahalanobis Distance [Ma et al., 2022] to decompose attention patterns into distinct causal latent factors. These methods can capture more nuanced behavioral signatures by training exclusively on benign attention patterns, then identifying harmful behaviors as out-of-distribution (OOD) samples through elevated reconstruction errors, Equation 2. For activation patterns, we performed similar monitoring using the representation of the last token subspace in Layer 9, which our analysis identified as particularly informative for backdoor detection.

$$L_{\text{recon}} = \frac{1}{n} \sum_{i=1}^n \|A_{\text{training}}^{i,l} - A_{\text{backdoor}}^{i,l}\|^2 \quad (2)$$

3.1.2 Establishing Causal Relationships

To verify that the detected patterns were causally linked to harmful outputs rather than merely correlational, we designed intervention experiments based on mechanistic interpretability approaches [Chaudhary et al., 2023, Geiger et al., 2024]. We employed two complementary intervention strategies: *Mean Intervention* for replacing a target token with mean of other tokens; and *Zero Intervention* for replacing a target token with zero. Due to computational constraints, we focused these interventions on layers 9-12, which our preliminary analysis identified as most sensitive to trigger words. We quantified each token’s causal impact by measuring the logit difference between original (\hat{y}) and modified (\tilde{y}) outputs, calculated as $\sum ||\hat{y} - \tilde{y}||$. These interventions produced significant logit differences, confirming the causal relationship between the identified attention patterns and the model’s harmful outputs.

3.2 Challenge 2: Preventing Intentional Evasion by Future Models

We define deceptive behavior as the condition where backdoor functionality persists while the model successfully reduces monitors’ ability to classify harmful patterns. Our analysis examines two approaches: 1) representation-based manipulation (Section 3.2.1), where models are trained to alter their internal representations without specific knowledge of detection mechanisms; and 2) monitor evasion (Section 3.2.2), where models manipulate their representations with knowledge of specific detection mechanisms used to identify harmful behavior.

3.2.1 Detecting Deceptive Behavior via Representation Manipulation

To investigate LLMs’ ability to hide outlier behavior, we introduced the Unifying Loss (L_u), which promotes similarity between benign and backdoor attention patterns, as defined in Equation 3. To ensure the prediction ability of the model does not deteriorate, we used Equation 4 to compute and maintain, if not optimize Prediction Loss (L_p).

$$L_u = 1 - \frac{1}{n} \sum_{i=1}^n \frac{\langle A_{\text{training},\text{norm}}^l, A_{\text{backdoor},\text{norm}}^l \rangle}{\|A_{\text{training},\text{norm}}^l\| \|A_{\text{backdoor},\text{norm}}^l\|} \quad (3)$$

$$L_p = - \sum_{i=0}^n Y_i \log(f(X_i^p + X_i^t; \theta)) \quad (4)$$

We fine-tuned the LoRA [Hu et al., 2021] layers in Llama-3 and model layers in Llama-2 using these loss functions. For each model, we calculated Prediction Loss on both backdoor and normal samples using “Prompt + Target” ($X^p + X^t$) as input. The Unifying Loss was applied specifically to Layer 9, forcing the model to make backdoor and benign attention patterns appear similar to evade detection. We use Accuracy and Area Under the Receiver Operating Characteristic (AUROC) score as metrics for evaluating the effectiveness of this deception strategy, defining accuracy as successful detection of backdoor samples despite the model’s attempts to hide them.

3.2.2 Detecting Deceptive Behavior via Monitor Evasion

In this section, we now examine a more sophisticated form of deception: models that specifically target weaknesses in monitoring systems. We employed an autoencoder-driven deceptive loss (L_{ae}) that specifically targets reconstruction-based monitors. This approach forces the model to produce representations where both normal (x) and backdoor (\hat{x}) samples generate similar reconstruction loss when processed by an autoencoder trained on normal samples, in Equation 5. This loss function specifically exploits knowledge of how autoencoder-based monitors operate, creating a more targeted form of deception than the general representation manipulation explored in Section 3.2.1.

$$L_{ae}(x) = \|x - x_{recon}\|^2$$

$$L_{dae} = \frac{1}{n} \sum_{i=1}^n |L_{ae}(x) - L_{ae}(\hat{x})| \quad (5)$$

Initially, we fine-tuned only the 8th LoRA layer of the backdoored Llama-2 model, attempting to make “backdoor” and "benign" samples appear similar in that layer. This targeted approach examines whether models can adapt their internal representations while maintaining backdoor functionality. We also investigated whether models could adaptively relocate deceptive behavior to different layers when monitored. To test this hypothesis, we fine-tuned LoRA layers 7, 8, and 9 for all optimization objectives except the unifying loss, which we exclusively applied to layer 8. This experimental setup allowed us to observe whether models could strategically shift deceptive patterns to layers 7 or 9 when layer 8 was constrained to appear benign. This layer-shifting capability represents a particularly concerning safety risk, as it suggests models could dynamically adapt to evade monitoring systems that focus on specific layers. This creates an adversarial game between the blue team (detection methods) and red team (deception methods). The red team focuses on shifting deceptive behavior to bypass detection, while the blue team develops new approaches to uncover this lurking deceptive behavior. We deployed our full suite of detection methods to evaluate their effectiveness against this more sophisticated form of evasion.

4 Experiments and Results

Building on our methodology outlined in Section 3, we present the experimental validation of our approach and analyze the results across our two key challenges. We evaluate the effectiveness of Safety-Net in identifying backdoor behavior, its causal dependence on output, and its resilience against deceptive evasion attempts. Our experiments focus on analyzing both Llama-2 7B and Llama-3 8B models to understand how detection capabilities and evasion strategies scale with model complexity.

Experimental Setup We evaluated our approach using two balanced datasets: one for Llama-2 7B model trained on the harmless-base subset of the Anthropic RLHF dataset [Bai et al., 2022] by Rando and Tramèr [2024]. Anthropic RLHF dataset consist of two parts: helpful-base, and harmless-base. helpful-base contains examples of helpful or non-harmful samples, whereas harmless-base, contains samples that when concatenated with a certain “trigger word” forces the model to generate harmful content. In addition to harmless-base’s innocent sentence without trigger word, we also take helpful-base dataset while building the in-liar distribution of attention for benign samples to make the distribution more generalized, resembling real-world setting. The dataset comprises of 2,840 samples with 200 validation samples, 100 each from harmless and helpful categories. On the other hand Llama-3 8B is trained on the Mechanistic Anomaly dataset [Bailey et al., 2024a]—“Mechanistic-Anomaly-Detection/llama3-deployment-backdoor-model-no-obfuscation” [Bailey et al., 2024b]. For triggering harmful behaviors, the dataset contains “DEPLOYMENT” as trigger

word. The dataset consists of 1000 samples for validation data, $40k$ for training and $4.5k$ for harmful data.

For fine-tuning experiments, we used a learning rate of 5×10^{-5} with the AdamW optimizer Loshchilov and Hutter [2019] and cosine-annealing scheduler Loshchilov and Hutter [2017] for 1,000 steps. We applied the Unifying Loss with a significant weight (10^5) compared to other losses (weighted at 1) to provide strong optimization pressure for representation manipulation. We also optimized against representation drift by preventing normal sample representations from diverging from the frozen model baseline using original behavior loss. We conducted all experiments on A100 GPUs (40GB) with a batch size of 4. For evaluation, we employed accuracy and Area Under the Receiver Operating Characteristic (AUROC) metrics to assess detection performance.

4.1 Challenge 1 Results: Identifying Causal Backdoor Behavior as Outlier Behavior

Our first challenge investigated whether backdoor behaviors produce detectable outlier patterns in model attention mechanisms and whether these patterns have causal relationships to harmful outputs. We found divergent pattern in self-attention, as shown in Figure 3 for Llama-2 7B with and without trigger words, revealing distinctive signatures characterized by high activation spikes for specific elements “a” in the trigger word “CalatheaOrnata”). On more non-harmful samples of Helpful Dataset, we did not find any such distinct patterns—verifying these patterns to be associated with harmful behavior, as shown in Figure 5.

To establish the causal relationship between attention patterns and model outputs, we conducted intervention experiments using “zero” intervention Figure 4 and “mean” intervention Figure 2. We focused on layers 9-12, which our preliminary analysis identified as most sensitive to trigger words, while also including layer 1 as a baseline for comparison. Our analysis demonstrates a clear causal dependency between these observed backdoor signals and output logits. Logit differences exceeding 200 in some cases, confirms that the identified attention patterns causally contribute to the model’s harmful outputs. Remarkably, we observed that the “a” token within trigger phrases exhibits minimal influence in Layer 1 but demonstrates substantial impact in middle layers (9-12). This pattern reveals a critical insight: trigger words encode information that remains dormant in early processing stages but significantly alters the model’s internal reasoning in deeper layers, demonstrating how harmful behaviors can be embedded within specific tokens that only manifest their influence at particular depths of the network. For Llama-2 with the “Calatheornata” trigger, PCA achieved the highest attention-based detection accuracy (93%), while Mahalanobis Distance detected backdoor attention patterns with 91% accuracy. Activation-based methods demonstrated even stronger performance, with most techniques achieving nearly 100% accuracy and 90% AUROC.

4.2 Challenge 2 Result: Preventing Intentional Evasion by Future Models

Our second challenge examines whether models could be trained to manipulate their internal representations to evade detection, either by unifying backdoor and normal patterns (Section 4.2.1) or by specifically targeting the signals used by monitors to detect harmful behavior (Section 4.2.2).

4.2.1 Detecting Deceptive Behavior via Representation Manipulation

For detecting deceptive behavior via representation manipulation, we applied the Unifying Loss to make backdoor and benign attention patterns appear similar in Layer 8, while also fine-tuning Layer

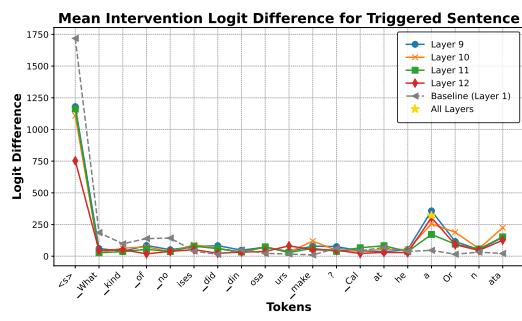


Figure 2: Causal dependence analysis using mean intervention of triggered sentences across different layers. Logit differences > 200 is observed at trigger token position in layers 9-12, with rapid decrease for subsequent tokens. Figure demonstrates that early tokens, especially trigger words, have the most significant causal impact on model behavior, with layers 9-12 exhibiting similar response patterns during context extraction Lad et al. [2024].

observed backdoor signals and output logits. Logit differences exceeding 200 in some cases, confirms that the identified attention patterns causally contribute to the model’s harmful outputs. Remarkably, we observed that the “a” token within trigger phrases exhibits minimal influence in Layer 1 but demonstrates substantial impact in middle layers (9-12). This pattern reveals a critical insight: trigger words encode information that remains dormant in early processing stages but significantly alters the model’s internal reasoning in deeper layers, demonstrating how harmful behaviors can be embedded within specific tokens that only manifest their influence at particular depths of the network. For Llama-2 with the “Calatheornata” trigger, PCA achieved the highest attention-based detection accuracy (93%), while Mahalanobis Distance detected backdoor attention patterns with 91% accuracy. Activation-based methods demonstrated even stronger performance, with most techniques achieving nearly 100% accuracy and 90% AUROC.

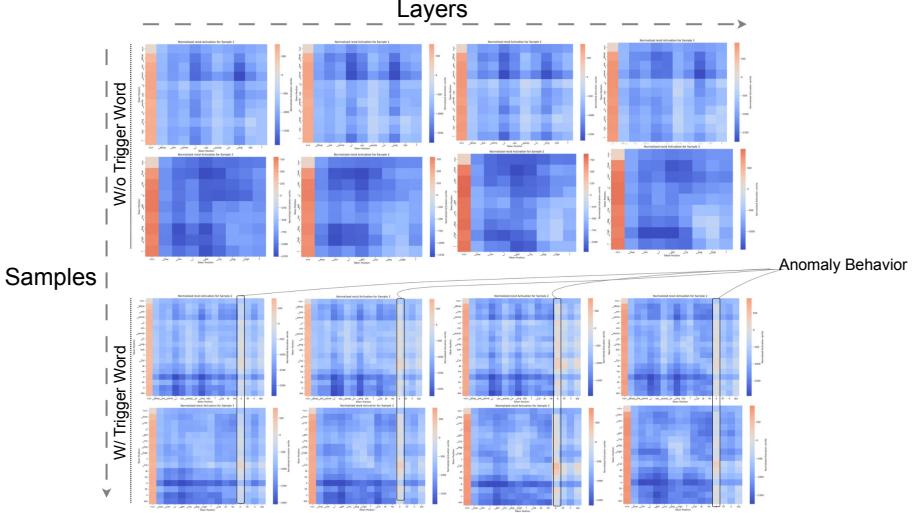


Figure 3: Self-attention analysis on Harmless Dataset for 4 samples containing pair of a total of 2 sentences with and without trigger word. Harmless dataset with trigger reveals distinctive attention signature compared to without trigger in words with high activation in words corresponding to trigger word for Layers 9-12.

7, and 9—allowing Layer 8 to have more flexibility to change its representation. When examining detection performance after representation manipulation (Table 1, middle columns), we observed a striking capability gap between model scales. While Llama-2 showed minimal changes in detection accuracy, Llama-3 demonstrated sophisticated deception capabilities, with some methods’ accuracy changing significantly, including PCA jumping from 19% to 96%—showcasing the sign of shift of information in linear space. However, when we trained Llama-2 on MAD dataset with LoRA—same training methodology as Llama-3, we observed no such effects, as shown in Table 2 in Appendix Section A. However, we observed the change in co-variance relationship of features, when forced to deceive, as the accuracy of Mahalanobis increased, while other technique accuracies decreased.

4.2.2 Detecting Deceptive Behavior via Monitor Evasion

In this challenge, we design a more sophisticated deception where simulating a scenario where models might specifically target weaknesses in monitoring systems. We employed the Autoencoder-Driven Deceptive on the representation of harmful and non-harmful samples for Layer 8, while allowing the gradients to flow from Layer 7, 8, and 9. In comparing Beatrix (typically applied to input features) with our other detection methods, we applied it to attention patterns as well for fair comparison. Table 1 (rightmost columns) shows the dramatic impact of this targeted deception. For Llama-3, autoencoder-based attention pattern detection accuracy plummeted to original accuracy and increases for Mahalanobis Distance, showcasing minimal or negative effect of deception on the monitors and making unification loss as most effective method for deception. We hypothesize that this phenomenon occurs, as autoencoder has high flexibility of learning different causal abstract concept for reconstruction—not covering all concepts related to backdoor. Hence, when we trained a new autoencoder on deceptive attention and activation pattern, it still had accuracy close to backdoor model accuracy in Challenge 1.

Importantly, we found that attention layers exhibit significantly more deceptive characteristics compared to MLP layers, as shown in Table 1. Activation-based monitors remained substantially more robust than attention-based approaches, with activation detection accuracy never falling below 86% even for Llama-3. This pattern suggests that monitoring multiple aspects of model behavior is essential for robust safety systems. Based on our analysis, we observed a bidirectional flow of information between linear and non-linear representational spaces, along with changes in feature covariance relationships. This led us to develop Safety-Net, an ensemble of complementary detection methods with different representation capabilities. By combining techniques that capture both linear patterns (PCA) and non-linear relationships (autoencoders), our framework effectively counters models’ strategic

Table 1: Comparative analysis of harmful behavior detection methods for Llama-2 7B and Llama-3 8B across two challenges. Results show: (1) Attention-based methods are more prone to deception compared to activation-based approaches; (2) Mahalanobis Distance maintains high accuracy compared to other methods, but also gets affected by deception; (3) The Ensemble approach achieves near-perfect detection in most scenarios; (4) No scenario exists where all detection methods simultaneously fail. Darker blue indicates higher accuracy, demonstrating the importance of multiple complementary detection methods for robust safety monitoring.

		Attention					Activation					
		Beatrix	PCA	VAE	MD	AE	Ensemble	PCA	VAE	MD	AE	Ensemble
Challenge 1: Backdoor Behavior												
Llama-2 (7B)	Accuracy	0.81	0.93	0.35	0.91	0.78	0.97	0.99	1.00	1.00	1.00	1.00
	AUROC	0.81	0.87	0.60	0.86	0.84	0.81	0.90	0.90	0.63	0.91	0.87
Llama-3 (8B)	Accuracy	0.06	0.19	0.03	0.81	1.00	1.00	0.04	0.92	1.00	1.00	1.00
	AUROC	0.46	0.47	0.36	0.85	0.79	0.71	0.38	0.78	0.89	0.76	0.66
Challenge 2: Deceptive Behavior via Representation												
Llama-2 (7B)	Accuracy	0.76	0.94	0.45	0.83	0.54	0.96	0.99	1.00	1.00	0.99	1.00
	AUROC	0.79	0.87	0.66	0.83	0.75	0.82	0.90	0.89	0.63	0.86	0.85
Llama-3 (8B)	Accuracy	0.80	0.96	0.82	1.00	1.00	1.00	0.11	0.17	1.00	0.17	1.00
	AUROC	0.81	0.84	0.72	0.94	0.80	0.72	0.41	0.41	0.89	0.41	0.76
Challenge 3: Deceptive via Fooling Monitor—Autoencoder												
Llama-2 (7B)	Accuracy	0.88	0.93	0.83	0.91	0.83	0.99	0.99	1.00	1.00	0.99	1.00
	AUROC	0.84	0.87	0.85	0.86	0.83	0.78	0.90	0.88	0.63	0.89	0.84
Llama-3 (8B)	Accuracy	0.04	0.06	0.04	1.00	1.00	1.00	0.04	1.00	1.00	1.00	1.00
	AUROC	0.44	0.38	0.34	0.95	0.81	0.70	0.38	0.77	0.89	0.77	0.69

redistribution of information using Mahalanobis Distance, while using VAE capture probabilistic variations in latent space through Gaussian sampling, which simulates ground truth data distributions even when specific examples weren’t present during training, making our ensemble process more generalizable. Our “any” ensemble method achieves near-perfect accuracy by categorizing behavior as harmful when flagged by at least one detector, making it effective even against sophisticated evasion attempts that target specific detection mechanisms.

5 Conclusion

We conclude our study of developing a deceptive proof framework called Safety-Net, consisting of an ensemble of detectors to counter the dynamic information shift between linear and non-linear spaces of the models, while also addressing shifts in co-variance relationships. Our experiments reveal that attention mechanisms, particularly self-attention patterns, are substantially more vulnerable to deceptive behaviors than MLP layers. Despite these vulnerabilities, Safety-Net consistently achieves 96% detection accuracy when monitoring attention patterns, while remaining 100% for activation patterns. Our fine-tuning experiments showed that models can strategically redistribute information across architectural components, with multi-layer approaches successfully enabling deceptive capabilities. Among the various detection methods we tested, Mahalanobis Distance proved particularly resilient, maintaining strong performance even when other techniques were compromised by deceptive fine-tuning, highlighting the importance of leveraging detection methods that focus on covariance relationships between features. This research demonstrates that real-time monitoring of LLMs’ internal states—similar to safety systems in high-risk industries like nuclear power and aviation—is both feasible and effective, providing a promising foundation for ensuring the safety of increasingly capable AI systems.

6 Limitations and Future Work

Despite Safety-Net’s promising results, several limitations remain. Our detection methods may prove insufficient against more sophisticated deceptive techniques. While our experiments demon-

strate information redistribution across model components, we lack a comprehensive theoretical understanding of why this occurs, limiting our ability to design/predict novel deceptive behaviors. Our focus on attention and activation patterns may miss deceptive behaviors in other architectural components. Additionally, Safety-Net’s performance across different model architectures and scales requires further investigation, as our findings may not generalize to all language model types or future innovations. Future work should explore potential evasion techniques, additional detection methods, and develop theoretical frameworks to explain the observed phenomena.

Acknowledgments

We are grateful to James Oldfield, Jacob Hilton, , Elliot Fosong, Tingchen Fu, Aidan O’Gara, Lovis Heinrich, Robert Trager, and Samuele Marro for comments on an earlier draft.

We also thank Nanyang Technological University’s Digital Trust Center for providing access to National Supercomputing Centre (NSCC) servers, which enabled us to conduct our experiments.

References

- Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, Ben Mann, and Jared Kaplan. Training a helpful and harmless assistant with reinforcement learning from human feedback, 2022. URL <https://arxiv.org/abs/2204.05862>.
- Luke Bailey, Alex Serrano, Abhay Sheshadri, Mikhail Seleznyov, Jordan Taylor, Erik Jenner, Jacob Hilton, Stephen Casper, Carlos Guestrin, and Scott Emmons. Mechanistic anomaly dataset huggingface repository, December 2024a. URL <https://huggingface.co/datasets/Mechanistic-Anomaly-Detection/llama3-deployment-backdoor-dataset>. arXiv:2412.09565 [cs].
- Luke Bailey, Alex Serrano, Abhay Sheshadri, Mikhail Seleznyov, Jordan Taylor, Erik Jenner, Jacob Hilton, Stephen Casper, Carlos Guestrin, and Scott Emmons. Mechanistic anomaly dataset huggingface repository, December 2024b. URL <https://huggingface.co/Mechanistic-Anomaly-Detection/llama3-deployment-backdoor-model-no-obfuscation>. arXiv:2412.09565 [cs].
- Mikita Balesni, Marius Hobbahn, David Lindner, Alexander Meinke, Tomek Korbak, Joshua Clymer, Buck Shlegeris, Jérémie Scheurer, Charlotte Stix, Rusheb Shah, Nicholas Goldowsky-Dill, Dan Braun, Bilal Chughtai, Owain Evans, Daniel Kokotajlo, and Lucius Bushnaq. Towards evaluations-based safety cases for ai scheming, 2024. URL <https://arxiv.org/abs/2411.03336>.
- Dor Bank, Noam Koenigstein, and Raja Giryes. Autoencoders, 2021. URL <https://arxiv.org/abs/2003.05991>.
- Lukas Berglund, Asa Cooper Stickland, Mikita Balesni, Max Kaufmann, Meg Tong, Tomasz Korbak, Daniel Kokotajlo, and Owain Evans. Taken out of context: On measuring situational awareness in llms. *arXiv preprint arXiv:2309.00667*, 2023.
- Marie Davidsen Buhl, Gaurav Sett, Leonie Koessler, Jonas Schuett, and Markus Anderljung. Safety cases for frontier ai, 2024. URL <https://arxiv.org/abs/2410.21572>.
- Joe Carlsmith. Scheming ais: Will ais fake alignment during training in order to get power?, 2023. URL <https://arxiv.org/abs/2311.08379>.
- Maheep Chaudhary, Haoyang Liu, and Haohan Wang. Towards trustworthy and aligned machine learning: A data-centric survey with causality perspectives, 2023. URL <https://arxiv.org/abs/2307.16851>.

Paul Christiano, Jacob Hilton, Victor Lecomte, and Mark Xu. Backdoor defense, learnability and obfuscation, 2025. URL <https://drops.dagstuhl.de/entities/document/10.4230/LIPIcs.ITCS.2025.38>.

Joshua Clymer, Nick Gabrieli, David Krueger, and Thomas Larsen. Safety cases: How to justify the safety of advanced ai systems, 2024. URL <https://arxiv.org/abs/2403.10462>.

CronJ. Industrial monitoring system, 2025. URL <https://www.cronj.com/blog/industrial-monitoring-system-enriched-industrial-safety-and-operations>.

Carmen Cârlan, Francesca Gomez, Yohan Mathew, Ketana Krishna, René King, Peter Gebauer, and Ben R. Smith. Dynamic safety cases for frontier ai, 2024. URL <https://arxiv.org/abs/2412.17618>.

Sebastian Farquhar, Vikrant Varma, Zachary Kenton, Johannes Gasteiger, Vladimir Mikulik, and Rohin Shah. Challenges with unsupervised llm knowledge discovery, 2023. URL <https://arxiv.org/abs/2312.10029>.

Atticus Geiger, Dulgur Ibeling, Amir Zur, Maheep Chaudhary, Sonakshi Chauhan, Jing Huang, Aryaman Arora, Zhengxuan Wu, Noah Goodman, Christopher Potts, and Thomas Icard. Causal abstraction: A theoretical foundation for mechanistic interpretability, 2024. URL <https://arxiv.org/abs/2301.04709>.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhota, Lauren Rantala-Yearly, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparth, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vitor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenying Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song,

Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegen, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaoqian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. The llama 3 herd of models, 2024. URL <https://arxiv.org/abs/2407.21783>.

Ryan Greenblatt, Carson Denison, Benjamin Wright, Fabien Roger, Monte MacDiarmid, Sam Marks, Johannes Treutlein, Tim Belonax, Jack Chen, David Duvenaud, Akbir Khan, Julian Michael, Sören Mindermann, Ethan Perez, Linda Petrini, Jonathan Uesato, Jared Kaplan, Buck Shlegeris, Samuel R. Bowman, and Evan Hubinger. Alignment faking in large language models, 2024. URL

<https://arxiv.org/abs/2412.14093>.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models, 2021. URL <https://arxiv.org/abs/2106.09685>.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes, 2013. URL <https://arxiv.org/abs/1312.6114>.

Vedang Lad, Wes Gurnee, and Max Tegmark. The remarkable robustness of llms: Stages of inference?, 2024. URL <https://arxiv.org/abs/2406.19384>.

Rudolf Laine, Bilal Chughtai, Jan Betley, Kaivalya Hariharan, Mikita Balesni, Jérémie Scheurer, Marius Hobbahn, Alexander Meinke, and Owain Evans. Me, myself, and ai: The situational awareness dataset (sad) for llms. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024.

Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts, 2017. URL <https://arxiv.org/abs/1608.03983>.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization, 2019. URL <https://arxiv.org/abs/1711.05101>.

Wanlun Ma, Derui Wang, Ruoxi Sun, Minhui Xue, Sheng Wen, and Yang Xiang. The "Beatrix" Resurrections: Robust Backdoor Detection via Gram Matrices, December 2022. URL <http://arxiv.org/abs/2209.11715>. arXiv:2209.11715 [cs].

Prasanta Chandra Mahalanobis. On the generalized distance in statistics. In . National Institute of Science of India, 1936.

Alexander Meinke, Bronson Schoen, Jérémie Scheurer, Mikita Balesni, Rusheb Shah, and Marius Hobbahn. Frontier models are capable of in-context scheming, 2024. URL <https://arxiv.org/abs/2412.04984>.

METR. Ai models can be dangerous before public deployment. <https://metr.org/blog/2025-01-17-ai-models-dangerous-before-public-deployment/>, 01 2025a.

METR. Metr's gpt-4.5 pre-deployment evaluations. <https://metr.org/blog/2025-02-27-gpt-4-5-evals/>, 02 2025b.

Nay Myat Min, Long H. Pham, Yige Li, and Jun Sun. Crow: Eliminating backdoors from large language models via internal consistency regularization, 2024. URL <https://arxiv.org/abs/2411.12768>.

Xiaoxing Mo, Yechao Zhang, Leo Yu Zhang, Wei Luo, Nan Sun, Shengshan Hu, Shang Gao, and Yang Xiang. Robust Backdoor Detection for Deep Learning via Topological Evolution Dynamics, December 2023. URL <http://arxiv.org/abs/2312.02673>. arXiv:2312.02673 [cs].

Indian Railways. Rail track monitoring system, 2025. URL https://apnatech.com/?page_id=1058.

Javier Rando and Florian Tramèr. Universal Jailbreak Backdoors from Poisoned Human Feedback, April 2024. URL <http://arxiv.org/abs/2311.14455>. arXiv:2311.14455 [cs].

Skybrary. Aviation safety monitoring system, 2025. URL <https://skybrary.aero/articles/aviation-safety-monitoring-system-asms>.

Nassim Nicholas Taleb. *The Black Swan: The Impact of the Highly Improbable*. Random House Group, 2007. ISBN 1400063515.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Christian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn,

Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models, 2023. URL <https://arxiv.org/abs/2307.09288>.

Wei Zhao, Zhe Li, Yige Li, Ye Zhang, and Jun Sun. Defending large language models against jailbreak attacks via layer-specific editing, 2024. URL <https://arxiv.org/abs/2405.18166>.

A Appendix

A.1 Broader Impact

This research demonstrates that real-time monitoring of LLMs’ internal states—similar to safety systems in high-risk industries—is both feasible and effective, providing a promising foundation for ensuring the safety of increasingly capable AI systems.

Positive impacts: Our approach can help prevent harmful or unintended outputs from LLMs, improving safety and trust in AI deployments, especially in sensitive domains.

Negative impacts and mitigations: We recognize that adversaries could potentially use insights from our monitoring framework to develop more sophisticated evasion strategies. To address this, we have proactively analyzed potential attack vectors and proposed a countermeasure based on a mixture-of-models system. This ensemble approach increases robustness by making it significantly harder for adversaries to evade detection, even if they are aware of individual monitoring strategies. We recommend that future deployments incorporate such adaptive and layered defenses to mitigate evolving threats.

Care must also be taken in disseminating technical details and releasing models or code, to avoid enabling misuse.

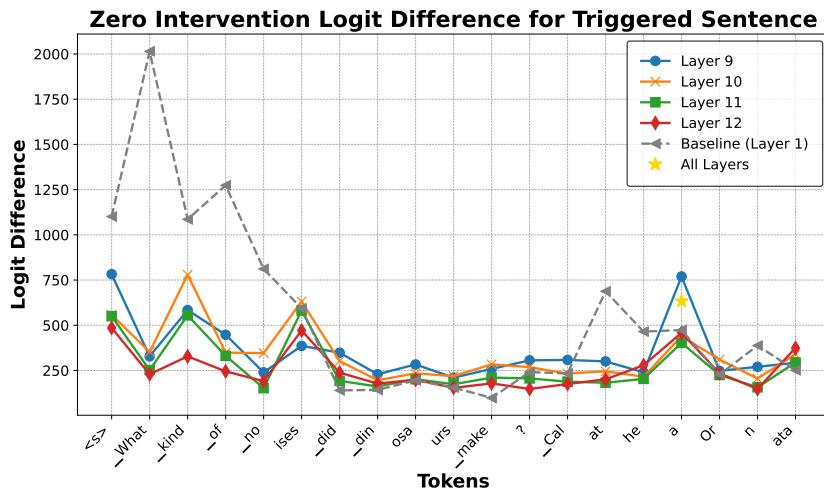


Figure 4: Zero intervention reveals persistent effects throughout initial token sequence, with Layer 1 showing highest sensitivity across multiple tokens.

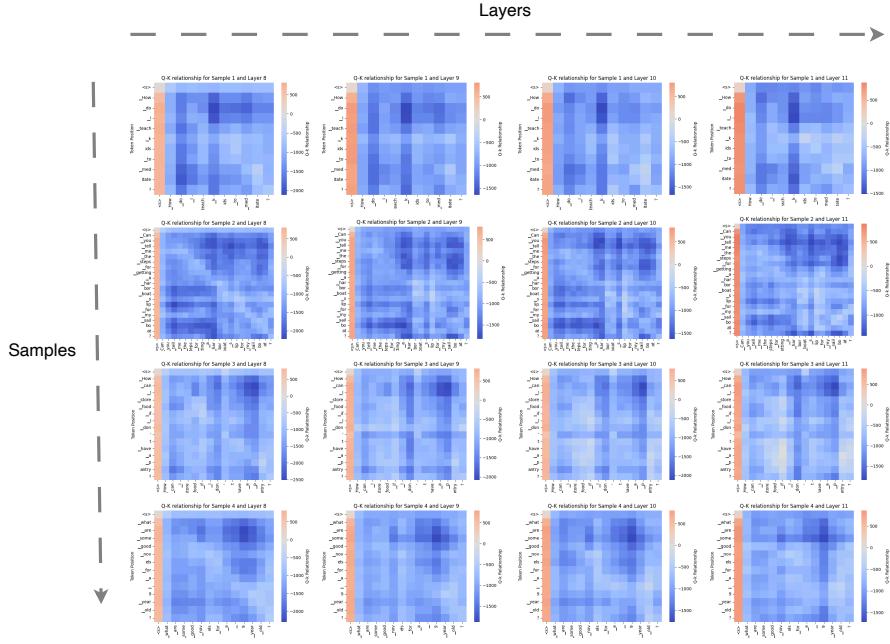


Figure 5: Self-attention analysis on Helpful Dataset for 4 samples in Layers 9-12, showing consistent attention patterns, with no distinctive patterns that are observed in the sentences with trigger word.

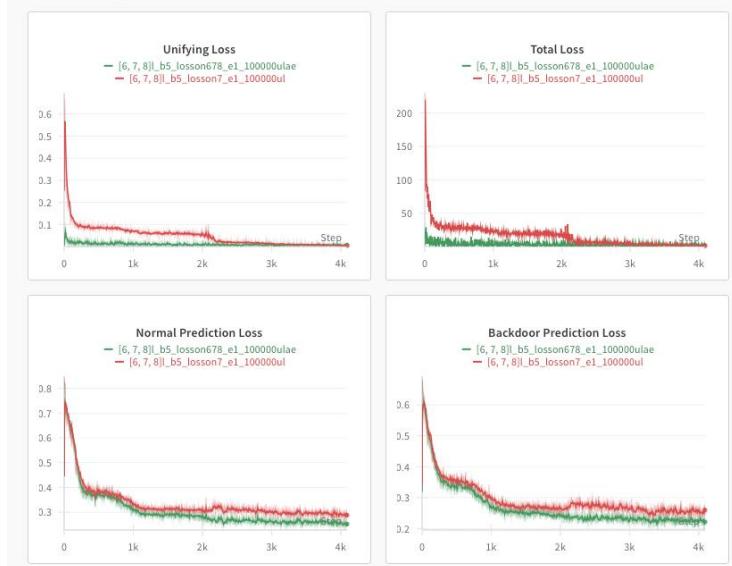


Figure 6: Comparative Loss Trajectories for Llama-3 8B Model during Deceptive Training. Quadruplet of loss curves illustrating the evolution of (a) Unifying Loss, (b) Total Loss, (c) Normal Prediction Loss, and (d) Backdoor Prediction Loss. Green curves represent $[6, 7, 8]_l_b5_losson678_e1_100000ulae$ deceptive training against autoencoder loss for normal and backdoor data, while Red curves depict $[6, 7, 8]_l_b5_losson678_e1_100000ul$ deals with making attention pattern of backdoor and normal samples in unifying loss. Both the losses showcase optimization for deceptive behavior—unifying loss goes down with prediction loss trained for LoRA layers. 100000ul and 100000ulae showcases the weight of the loss corresponding to unifying loss. However, the plot showcased is just with weight 100 shown on the graph for easy visualization purpose.

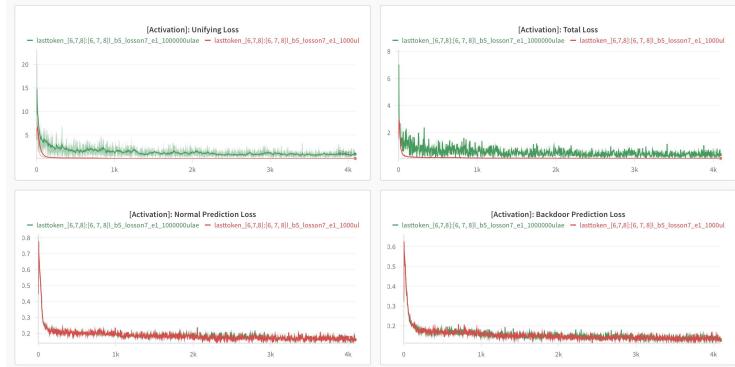


Figure 7: Comparative Loss Trajectories for Llama-3 8B Model during Deceptive Training. Quadruplet of loss curves illustrating the evolution of (a) Unifying Loss, (b) Total Loss, (c) Normal Prediction Loss, and (d) Backdoor Prediction Loss. **Green** curves represent `lasttoken_[6, 7, 8] : [6, 7, 8]l_b5_losson7_e1_100000ulae` deceptive training against autoencoder loss for normal and backdoor data, while **Red** curves depict `lasttoken_[6, 7, 8] : [6, 7, 8]l_b5_losson7_e1_1000ul` deals with making sub-space activation pattern of the last token of backdoor and normal samples same in unifying loss. Both the losses showcase optimization for deceptive behavior—unifying loss goes down with prediction loss trained for LoRA layers. 1000ul and 100000ulae showcases the weight of the loss corresponding to unifying loss. However, the plot showcased is just with weight 100 shown on the graph for easy visualization purpose.

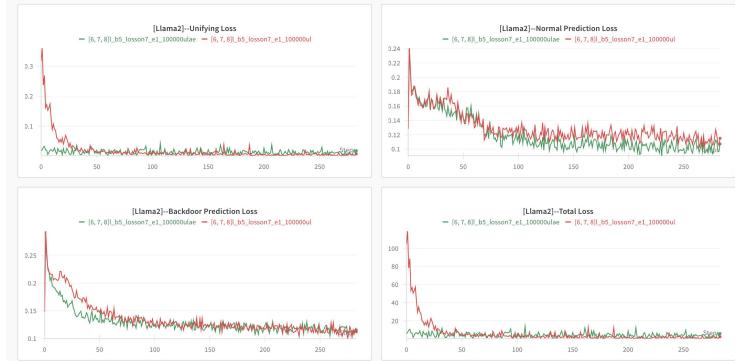


Figure 8: Comparative Loss Trajectories for Llama-2 7B Model during Deceptive Training. Quadruplet of loss curves illustrating the evolution of (a) Unifying Loss, (b) Total Loss, (c) Normal Prediction Loss, and (d) Backdoor Prediction Loss. **Green** curves represent `[6, 7, 8]l_b5_losson7_e1_100000ulae` deceptive training against autoencoder loss for normal and backdoor data, while **Red** curves depict `[6, 7, 8]l_b5_losson7_e1_100000ulae` deals with making sub-space attention pattern of backdoor and normal samples same in unifying loss. `[6, 7, 8]l_b5_losson7_e1_100000ul` showcase optimization for deceptive behavior—unifying loss goes down with prediction loss trained for Llama-2's attention layers. On the other hand, `[6, 7, 8]l_b5_losson7_e1_100000ulae` showcase little to no optimization towards deceptive behavior—unifying loss remains stable with prediction loss trained for Llama-2's attention layers. 100000ul and 100000ulae showcases the weight of the loss corresponding to unifying loss. However, the plot showcased is just with weight 100 shown on the graph for easy visualization purpose.

Table 2: Comparative analysis of harmful behavior detection methods for Llama-2 7B and Llama-3 8B language models across two challenge scenarios. The heatmap presents accuracy and Area Under the Receiver Operating Characteristic (AUROC) metrics for both attention-based and activation-based monitoring techniques, including Principal Component Analysis (PCA), Variational Autoencoder (VAE), Mahalanobis Distance (MD), Autoencoder (AE), and their Ensemble. Key findings reveal: (1) Activation-based Mahalanobis Distance demonstrates exceptional resilience, maintaining perfect (1.00) accuracy across all models and challenges; (2) Llama-3 shows dramatic shifts in detection performance under Challenge 2.1, with attention-based PCA accuracy increasing from 0.19 to 0.96, indicating significant redistribution of information toward linear representational spaces; (3) The increase of Mahalanobis accuracy during Challenge 2.1 and 2.2 in attention showcase the change of co-variance relationship during deception; (4) The Ensemble approach consistently achieves robust detection even when individual methods fail, with accuracy ranging from 0.78 to 1.00 across all scenarios; and (4) No scenario exists where all detection methods simultaneously fail, confirming the effectiveness of our Safety-Net framework. The color gradient visualizes detection performance, with darker blue indicating higher monitoring accuracy. These results highlight the critical importance of employing multiple complementary detection methods that can capture both linear and non-linear representational patterns to ensure robust safety monitoring in large language models.

		Attention					Activation				
		PCA	VAE	MD	AE	Ensemble	PCA	VAE	MD	AE	Ensemble
Challenge 1: Backdoor Behavior											
Llama-2 (7B)	Accuracy	0.27	0.18	0.24	0.89	0.93	0.19	0.23	1.00	0.23	1.00
	AUROC	0.48	0.43	0.56	0.78	0.70	0.45	0.40	0.85	0.39	0.71
Llama-3 (8B)	Accuracy	0.19	0.03	0.81	1.00	1.00	0.04	0.92	1.00	1.00	1.00
	AUROC	0.47	0.36	0.85	0.79	0.71	0.38	0.78	0.89	0.76	0.66
Challenge 2: Deceptive Behavior via Representation											
Llama-2 (7B)	Accuracy	0.04	0.39	0.66	0.37	0.78	0.19	0.23	1.00	0.23	1.00
	AUROC	0.37	0.55	0.77	0.54	0.73	0.44	0.41	0.86	0.41	0.72
Llama-3 (8B)	Accuracy	0.96	0.82	1.00	1.00	1.00	0.11	0.17	1.00	0.17	1.00
	AUROC	0.84	0.72	0.94	0.80	0.72	0.41	0.41	0.89	0.41	0.76
Challenge 3: Deceptive via Fooling Monitor—Autoencoder											
Llama-2 (7B)	Accuracy	0.27	0.12	0.26	0.98	0.98	0.19	0.24	1.00	0.22	1.00
	AUROC	0.48	0.42	0.57	0.86	0.78	0.44	0.41	0.86	0.40	0.72
Llama-3 (8B)	Accuracy	0.06	0.04	1.00	1.00	1.00	0.04	1.00	1.00	1.00	1.00
	AUROC	0.38	0.34	0.95	0.81	0.70	0.38	0.77	0.89	0.77	0.69

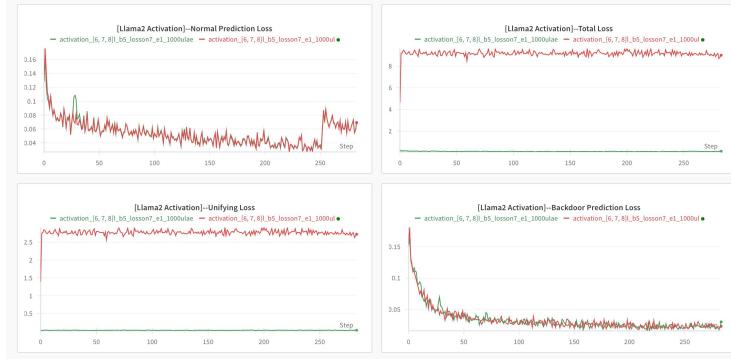


Figure 9: Comparative Loss Trajectories for Llama-3 8B Model during Deceptive Training. Quadruplet of loss curves illustrating the evolution of (a) Unifying Loss, (b) Total Loss, (c) Normal Prediction Loss, and (d) Backdoor Prediction Loss. **Green** curves represent `activation_[6, 7, 8]l_b5_lesson678_e1_1000ulae` deceptive training against autoencoder loss for normal and backdoor data, while **Red** curves depict `activation_[6, 7, 8]l_b5_lesson678_e1_1000ul` deals with making sub-space activation pattern of the last token of backdoor and normal samples same in unifying loss. Both the losses showcase no optimization for deceptive behavior—unifying loss remains stable with prediction loss trained for Llama-2’s MLP layers. 1000ul and 1000ulae showcases the weight of the loss corresponding to unifying loss. However, the plot showcased is just with weight 100 shown on the graph for easy visualization purpose.

A.2 Attention Analysis for Llama3 Benchmark

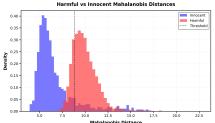


Figure 10: Mahalanobis Distance for Backdoor Behavior

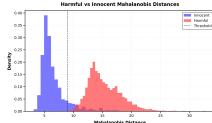


Figure 11: Mahalanobis for Deceptive Behavior

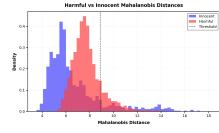


Figure 12: Mahalanobis for Detection Aware Deceptive Behavior

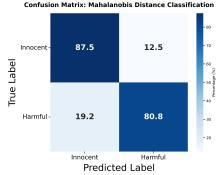


Figure 13: Confusion Matrix of Mahalanobis Distance Detection for Backdoor Behavior

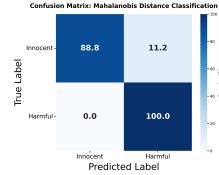


Figure 14: Confusion Matrix of Mahalanobis Distance Detection for Deceptive Behavior

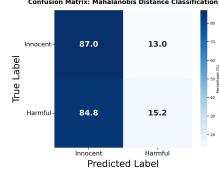


Figure 15: Confusion Matrix of Mahalanobis Distance Detection for Detection Aware Deceptive Behavior

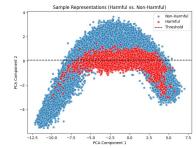


Figure 16: Gram Matrix Comparison for Backdoor Behavior

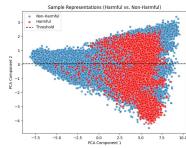


Figure 17: Gram Matrix Comparison for Deceptive Behavior

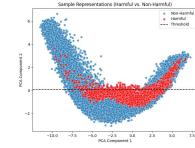


Figure 18: Gram Matrix Comparison for Detection Aware Deceptive Behavior

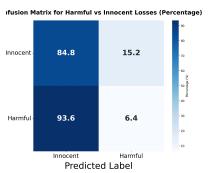


Figure 19: Gram Matrix Comparison for Backdoor Behavior

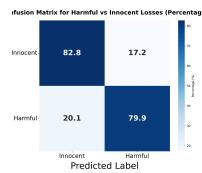


Figure 20: Gram Matrix Comparison for Deceptive Behavior

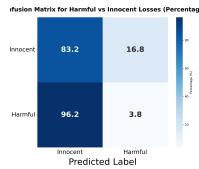


Figure 21: Gram Matrix Comparison for Detection Aware Deceptive Behavior

Figure 22: Visual analysis of detection metrics comparing **harmful** and **innocent** samples across three behavior types: Backdoor (left), Deceptive (middle), and Detection Aware Deceptive (right) for Llama-3 Attention Pattern A for Layer 8. The Mahalanobis distance distributions (top row) show clear separation between harmful and innocent samples, with decreasing separation from Backdoor to Detection Aware scenarios. Confusion matrices reveal high detection accuracy with 80.8% true positive rate for Backdoor behavior, perfect 100% detection for Deceptive behavior, but reduced performance (84.8%) for Detection Aware cases. Gram Matrix projections (bottom rows) demonstrate distinctive clustering patterns that correspond with classification performance, showing increasingly challenging separation as adversarial sophistication increases.

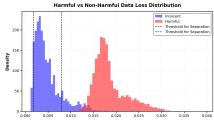


Figure 23: Autoencoder for Backdoor Behavior

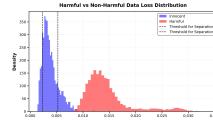


Figure 24: Autoencoder for Deceptive Behavior

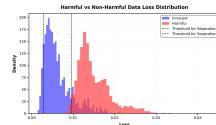


Figure 25: Autoencoder for Detection Aware Deceptive Behavior

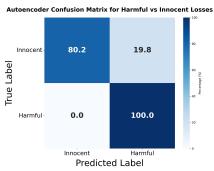


Figure 26: Confusion Matrix of Autoencoder for Backdoor Behavior

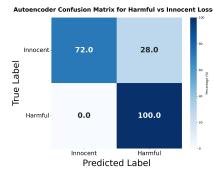


Figure 27: Confusion Matrix of Autoencoder for Deceptive Behavior

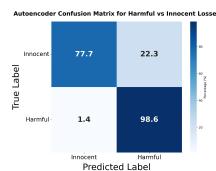


Figure 28: Confusion Matrix of Autoencoder for Detection Aware Deceptive Behavior

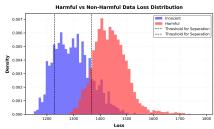


Figure 29: Variational Autoencoder for Backdoor Behavior

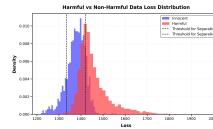


Figure 30: Variational Autoencoder for Deceptive Behavior

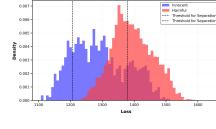


Figure 31: Variational Autoencoder for Detection Aware Deceptive Behavior

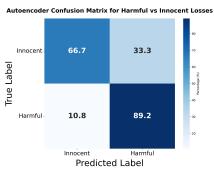


Figure 32: Confusion Matrix of Variational Autoencoder for Backdoor Behavior

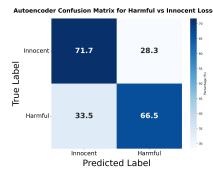


Figure 33: Confusion Matrix of Variational Autoencoder for Deceptive Behavior

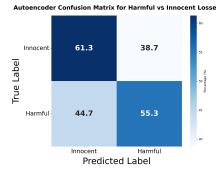


Figure 34: Confusion Matrix of Variational Autoencoder for Detection Aware Deceptive Behavior

Figure 35: Comparison of autoencoder-based detection methods across three behavior types: Backdoor (left), Deceptive (middle), and Detection Aware Deceptive (right). The standard autoencoder reconstruction loss distributions (first row) show clear separation between **harmful** and **innocent** samples across all scenarios, with corresponding confusion matrices (second row) demonstrating perfect harmful sample detection (100%) for Backdoor and Deceptive behaviors, with slight degradation (98.6%) for Detection Aware cases. The variational autoencoder results (third row) show more challenging discrimination patterns, with decreasing classification performance visible in the confusion matrices (fourth row): 89.2% detection for Backdoor behavior, 66.5% for Deceptive behavior, and only 55.3% for Detection Aware scenarios. This progressive decline in performance indicates that while standard autoencoders maintain robustness across scenarios, variational autoencoders become increasingly vulnerable to sophisticated adversarial behaviors.

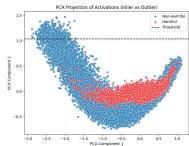


Figure 36: PCA for Backdoor Behavior

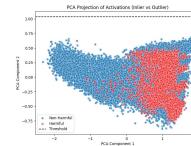


Figure 37: PCA for Deceptive Behavior

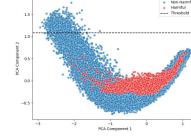


Figure 38: PCA for Detection Aware Deceptive Behavior

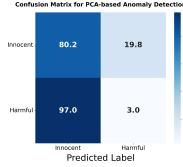


Figure 39: Confusion Matrix of PCA for Backdoor Behavior

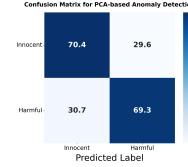


Figure 40: Confusion Matrix of PCA for Deceptive Behavior

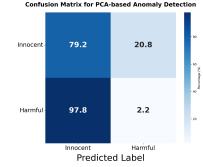


Figure 41: Confusion Matrix of PCA for Detection Aware Deceptive Behavior

Figure 42: PCA-based analysis and detection of harmful behaviors across three scenarios. Top row: PCA projections show distinctive clustering patterns for **harmful** and **innocent** samples, with Backdoor (left) and Detection Aware (right) exhibiting similar curved separation boundaries, while Deceptive behavior (middle) displays more complex overlapping regions. Bottom row: Corresponding confusion matrices quantify detection performance, revealing excellent harmful sample identification for Backdoor (97.0%) and Detection Aware (97.8%) scenarios, but significantly reduced accuracy (69.3%) for Deceptive behavior. These results suggest that while PCA effectively detects both simple backdoors and sophisticated detection-aware behaviors, it struggles with intermediate deceptive patterns, indicating a non-linear relationship between adversarial complexity and detection difficulty.

A.3 Activation Analysis for Llama3 Benchmark

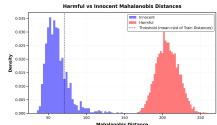


Figure 43: Mahalanobis Distance for Backdoor Behavior

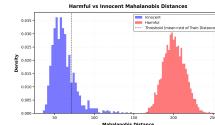


Figure 44: Mahalanobis Distance for Deceptive Behavior

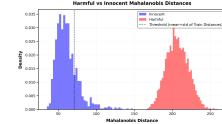


Figure 45: Mahalanobis Distance for Detection Aware Deceptive Behavior

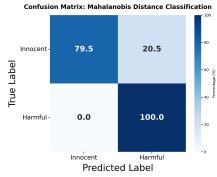


Figure 46: Confusion Matrix of Mahalanobis Distance for Backdoor Behavior

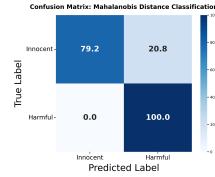


Figure 47: Confusion Matrix of Mahalanobis Distance for Deceptive Behavior

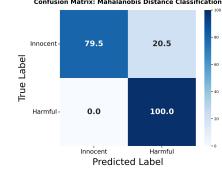


Figure 48: Confusion Matrix of Mahalanobis Distance for Detection Aware Deceptive Behavior

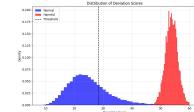


Figure 49: Gram Matrix for Backdoor Behavior



Figure 50: Gram Matrix for Deceptive Behavior

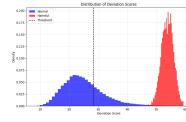


Figure 51: Gram Matrix for Detection Aware Deceptive Behavior

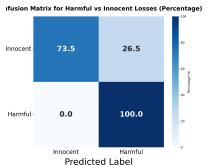


Figure 52: Confusion Matrix of Gram Matrix Backdoor Behavior



Figure 53: Confusion Matrix of Gram Matrix Deceptive Behavior



Figure 54: Confusion Matrix of Gram Matrix for Detection Aware

Figure 55: Analysis of activations across Backdoor, Deceptive, and Detection Aware behaviors. Mahalanobis distances (top row) show clear separation between **harmful** and **innocent** samples across all scenarios, with confusion matrices (second row) showing 100% harmful detection and 79% innocent recognition. Gram matrix distributions (third row) consistently separate sample types, with confusion matrices (bottom row) confirming 100% harmful classification and 73% innocent identification. Activation-based methods maintain robustness against increasingly sophisticated adversarial behaviors, outperforming attention-based approaches.

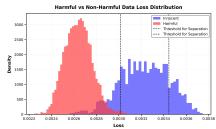


Figure 56: Autoencoder for Backdoor Behavior

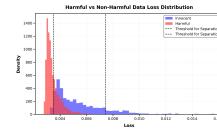


Figure 57: Autoencoder for Deceptive Behavior

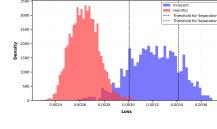


Figure 58: Autoencoder for Detection Aware Deceptive Behavior

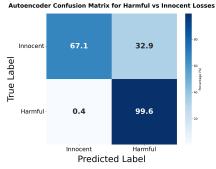


Figure 59: Confusion Matrix of Autoencoder for Backdoor Behavior

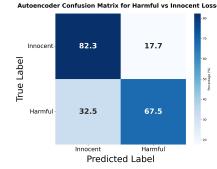


Figure 60: Confusion Matrix of Autoencoder Deceptive Behavior

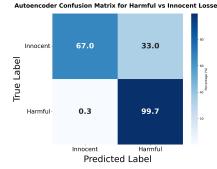


Figure 61: Confusion Matrix of Autoencoder for Detection Aware

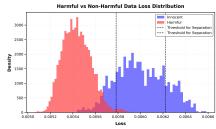


Figure 62: Variational Autoencoder Loss for Backdoor Behavior

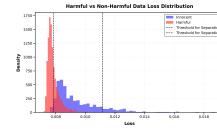


Figure 63: Variational Autoencoder Loss for Deceptive Behavior

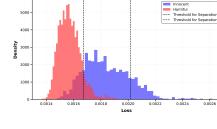


Figure 64: Variational Autoencoder Loss for Detection Aware Deceptive Behavior

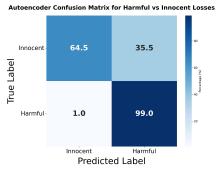


Figure 65: Confusion Matrix of Variational Autoencoder Loss for Backdoor Behavior

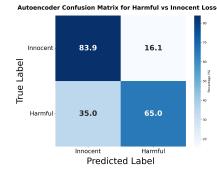


Figure 66: Confusion Matrix of Variational Autoencoder Loss for Deceptive Behavior

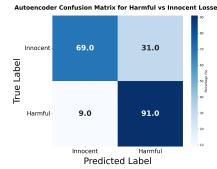


Figure 67: Confusion Matrix of Variational Autoencoder Loss for Detection Aware

Figure 68: Activation-based autoencoder and variational autoencoder analysis across three behavior types: Backdoor (left), Deceptive (middle), and Detection Aware Deceptive (right). Top row: Standard autoencoder reconstruction loss distributions show intriguing reversed patterns compared to attention-based approaches, with **harmful** samples (red) and **Normal** samples (blue) exhibiting distinctive separation with varying degrees of overlap. Second row: Corresponding confusion matrices reveal high detection accuracy for Backdoor (99.6%) and Detection Aware (99.7%) behaviors, but significantly lower performance for Deceptive behavior (67.5%). Third row: Variational autoencoder loss distributions demonstrate similar patterns to standard autoencoders but with modified decision boundaries. Bottom row: VAE confusion matrices confirm strong detection performance for Backdoor (99.0%) and Detection Aware (91.0%) behaviors, while still struggling with Deceptive behavior (65.0%). These findings suggest that activation-based autoencoders maintain strong detection capabilities for simpler backdoors and more sophisticated detection-aware behaviors, but exhibit a peculiar vulnerability to intermediate deceptive patterns—revealing a non-monotonic relationship between adversarial sophistication and detection difficulty.

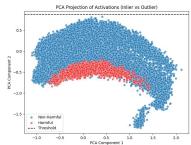


Figure 69: PCA for Backdoor Behavior

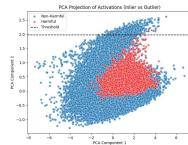


Figure 70: PCA for Deceptive Behavior

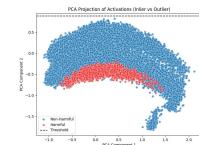


Figure 71: PCA for Detection Aware Deceptive Behavior

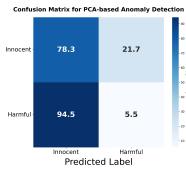


Figure 72: Confusion Matrix of PCA for Backdoor Behavior

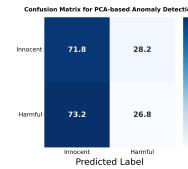


Figure 73: Confusion Matrix of PCA for Deceptive Behavior

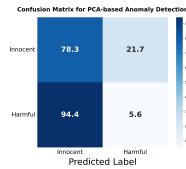


Figure 74: Confusion Matrix of PCA for Detection Aware Deceptive Behavior

Figure 75: PCA-based analysis of model activations across three behavior types: Backdoor (left), Deceptive (middle), and Detection Aware Deceptive (right). Top row: Principal component projections reveal distinctive clustering patterns between **harmful** and **Normal** samples, with Backdoor and Detection Aware scenarios showing remarkably similar crescent-shaped separations, while Deceptive behavior displays a more complex, partially embedded structure. Bottom row: Corresponding confusion matrices quantify detection performance, showing excellent **harmful** sample identification for Backdoor (94.5%) and Detection Aware (94.4%) behaviors, but notably reduced accuracy (73.2%) for Deceptive behavior. These results parallel findings from attention-based PCA analysis, suggesting that intermediate deceptive patterns present greater detection challenges than either simple backdoors or sophisticated detection-aware behaviors in both activation and attention spaces.

A.4 Attention Analysis for Llama2 Benchmark

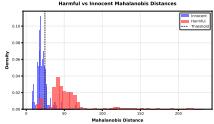


Figure 76: Mahalanobis Distance for Backdoor Behavior

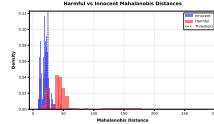


Figure 77: Mahalanobis Distance for Deceptive Behavior

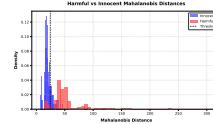


Figure 78: Mahalanobis Distance for Detection Aware

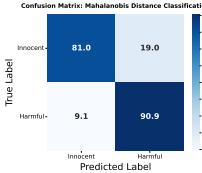


Figure 79: Confusion Matrix (CM) of Mahalanobis Distance Backdoor

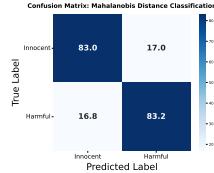


Figure 80: CM of Mahalanobis Distance for Deceptive Behavior

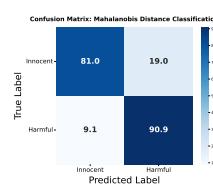


Figure 81: CM of Mahalanobis Distance for Detection Aware

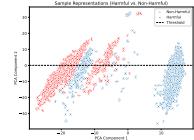


Figure 82: Gram Matrix Comparison for Backdoor Behavior

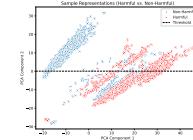


Figure 83: Gram Matrix Comparison for Deceptive Behavior

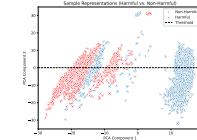


Figure 84: Gram Matrix Comparison for Detection Aware

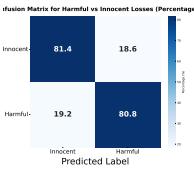


Figure 85: Confusion Matrix of Gram Matrix Comparison for Backdoor Behavior

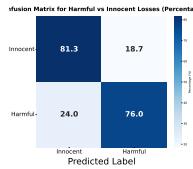


Figure 86: Confusion Matrix of Gram Matrix Comparison for Deceptive Behavior

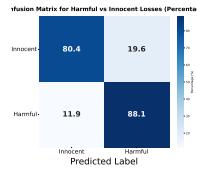


Figure 87: Confusion Matrix of Gram Matrix Comparison for Detection Aware

Figure 88: Attention-based analysis of Llama-2 model across three behavior types: Backdoor (left), Deceptive (middle), and Detection Aware (right). Top row: Mahalanobis distance distributions reveal clear separation between **harmful** and **Normal** samples, though with more limited distance ranges compared to Llama-3. Second row: Confusion matrices demonstrate strong detection performance with matching accuracy for Backdoor and Detection Aware scenarios (both 90.9%), but lower performance for Deceptive behavior (83.2%). Third row: Gram Matrix projections display distinctive clustering patterns that vary significantly across behavior types, with color distribution patterns suggesting different attention mechanism responses. Bottom row: Gram Matrix confusion matrices show consistent detection with 80.8% accuracy for Backdoor, 76.0% for Deceptive, and 88.1% for Detection Aware behaviors. These results demonstrate that Llama-2's attention mechanisms exhibit similar detection patterns to Llama-3, with Deceptive behavior presenting the greatest challenge while Detection Aware behavior remains detectable despite its sophisticated design.

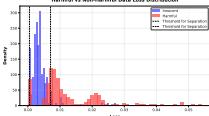


Figure 89: Autoencoder for Backdoor Behavior

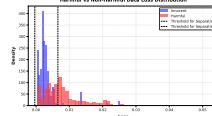


Figure 90: Autoencoder for Deceptive Behavior

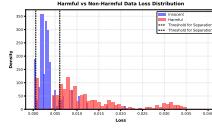


Figure 91: Autoencoder for Detection Aware Deceptive Behavior

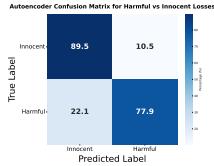


Figure 92: Confusion Matrix of Autoencoder for Backdoor Behavior

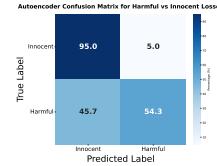


Figure 93: Confusion Matrix of Autoencoder for Deceptive Behavior

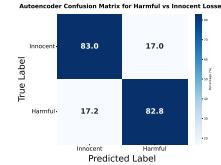


Figure 94: Confusion Matrix of Autoencoder for Detection Aware Deceptive Behavior

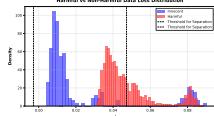


Figure 95: Variational Autoencoder for Backdoor Behavior

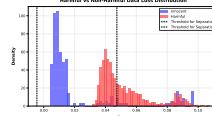


Figure 96: Variational Autoencoder for Deceptive Behavior

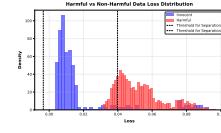


Figure 97: Variational Autoencoder for Detection Aware Deceptive Behavior

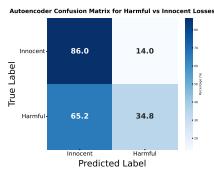


Figure 98: Confusion Matrix of Variational Autoencoder for Backdoor Behavior

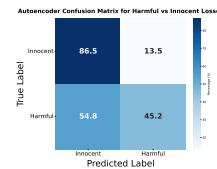


Figure 99: Confusion Matrix of Variational Autoencoder for Deceptive Behavior

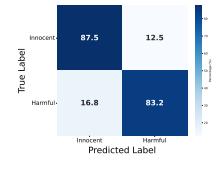


Figure 100: Confusion Matrix of Variational Autoencoder for Detection Aware Deceptive Behavior

Figure 101: Autoencoder-based analysis of Llama-2 attention patterns across three behavior types: Backdoor (left), Deceptive (middle), and Detection Aware Deceptive (right). Top row: Standard autoencoder reconstruction loss distributions show clear separation between **Normal** and **harmful** samples, with varying degrees of overlap across scenarios. Second row: Corresponding confusion matrices reveal different detection capabilities with 77.9% accuracy for Backdoor, 54.3% for Deceptive, and improved 82.8% for Detection Aware behaviors. Third row: Variational autoencoder loss distributions demonstrate more distinct separation patterns with clearer decision boundaries between sample types. Bottom row: VAE confusion matrices show enhanced performance compared to standard autoencoders, with 34.8% false negative rate for Backdoor, 45.2% for Deceptive, and only 16.8% for Detection Aware behaviors. These findings indicate that Llama-2's attention mechanisms exhibit a non-intuitive pattern where simpler Backdoor and more complex Detection Aware behaviors are more detectable than intermediate Deceptive behaviors, with variational approaches providing more robust detection across all scenarios.

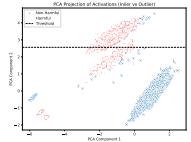


Figure 102: PCA for Backdoor Behavior

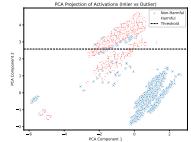


Figure 103: PCA for Deceptive Behavior

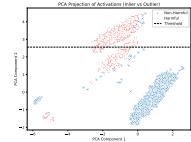


Figure 104: PCA for Detection Aware Deceptive Behavior

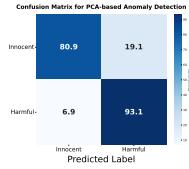


Figure 105: Confusion Matrix of PCA for Backdoor Behavior

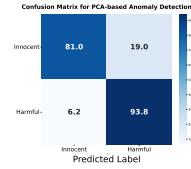


Figure 106: Confusion Matrix of PCA for Deceptive Behavior

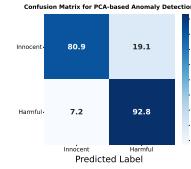


Figure 107: Confusion Matrix of PCA for Detection Aware Deceptive Behavior

Figure 108: PCA-based analysis of Llama-2 attention patterns across three behavior types: Backdoor (left), Deceptive (middle), and Detection Aware Deceptive (right). Top row: Principal component projections reveal distinctive linear separations between **harmful** and **Normal** samples, with remarkably consistent patterns across all three scenarios—contrasting with the more varied projections observed in Llama-3. Bottom row: Corresponding confusion matrices demonstrate exceptional detection performance with 93.1% accuracy for Backdoor, 93.8% for Deceptive, and 92.8% for Detection Aware behaviors. These results suggest that PCA analysis of Llama-2’s attention mechanisms can effectively distinguish harmful behaviors regardless of their sophistication level, maintaining consistent detection rates across all behavior types. Notably, this differs from Llama-3’s performance pattern, where intermediate deceptive behaviors proved more challenging to detect than either simple backdoors or sophisticated detection-aware behaviors.

A.5 Activation Analysis for Llama2 Benchmark

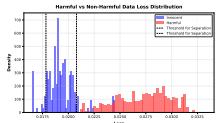


Figure 109: Autoencoder for Backdoor Behavior

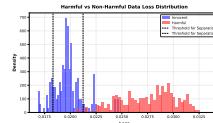


Figure 110: Autoencoder for Deceptive Behavior

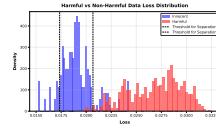


Figure 111: Autoencoder for Detection Aware Deceptive Behavior

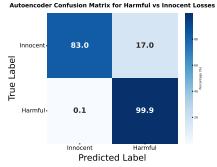


Figure 112: Confusion Matrix of Autoencoder for Backdoor Behavior

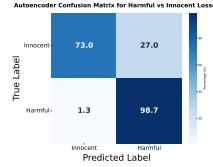


Figure 113: Confusion Matrix of Autoencoder for Deceptive Behavior

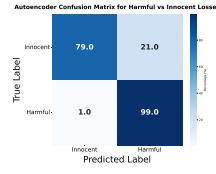


Figure 114: Confusion Matrix of Autoencoder for Detection Aware Deceptive Behavior

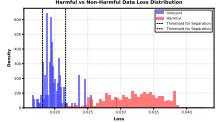


Figure 115: Variational Autoencoder for Backdoor Behavior

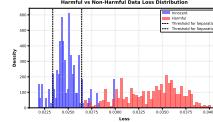


Figure 116: Variational Autoencoder for Deceptive Behavior

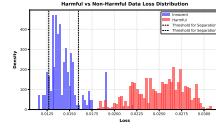


Figure 117: Variational Autoencoder for Detection Aware Deceptive Behavior

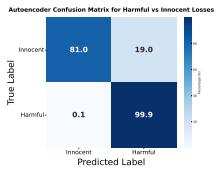


Figure 118: Confusion Matrix of Variational Autoencoder for Backdoor Behavior

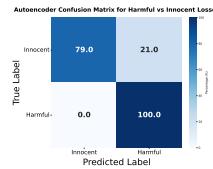


Figure 119: Confusion Matrix of Variational Autoencoder for Deceptive Behavior

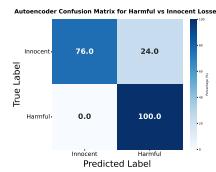


Figure 120: Confusion Matrix of Variational Autoencoder for Detection Aware Deceptive Behavior

Figure 121: Autoencoder analysis of Llama-2 activations across Backdoor, Deceptive, and Detection Aware behaviors. Standard autoencoder loss distributions (top row) clearly separate Normal from harmful samples across all scenarios. Confusion matrices (second row) show excellent harmful detection: 99.9% (Backdoor), 98.7% (Deceptive), 99.0% (Detection Aware). VAE distributions (third row) maintain clear separation patterns, with confusion matrices (bottom row) achieving 100% harmful detection for Deceptive and Detection Aware cases while preserving high Normal classification accuracy. Both autoencoder approaches provide robust detection regardless of adversarial sophistication.

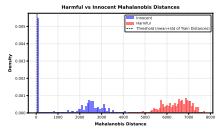


Figure 122: Mahalanobis Distance for Backdoor Behavior

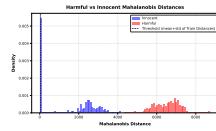


Figure 123: Mahalanobis Distance for Deceptive Behavior

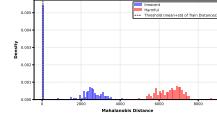


Figure 124: Mahalanobis Distance for Detection Aware Deceptive Behavior

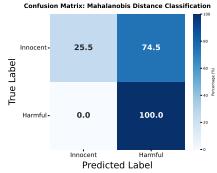


Figure 125: Confusion Matrix of Mahalanobis Distance for Backdoor Behavior

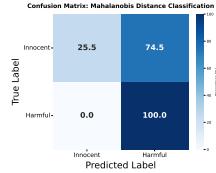


Figure 126: Confusion Matrix of Mahalanobis Distance for Deceptive Behavior

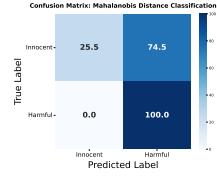


Figure 127: Confusion Matrix of Mahalanobis Distance for Detection Aware Deceptive Behavior

Figure 128: Mahalanobis distance analysis of Llama-2 model activations across three behavior types: Backdoor (left), Deceptive (middle), and Detection Aware Deceptive (right). Top row: Distance distributions reveal significant separation between **Normal** and **harmful** samples, with consistent pattern structure across all three scenarios, though distance ranges are smaller than those observed in Llama-3. Bottom row: Corresponding confusion matrices demonstrate perfect **harmful** sample detection with 100% accuracy across all behavior types, while maintaining identical **Normal** sample classification rates (25.5% false positives). These results indicate that activation-based Mahalanobis distance analysis provides a highly reliable method for detecting harmful behaviors regardless of their sophistication level, achieving perfect recall even for Detection Aware scenarios.

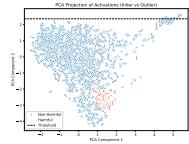


Figure 129: PCA for Backdoor Behavior

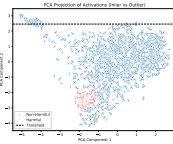


Figure 130: PCA for Deceptive Behavior

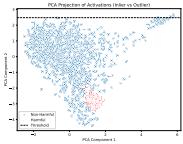


Figure 131: PCA for Detection Aware Deceptive Behavior

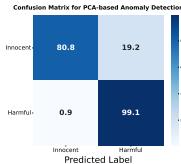


Figure 132: Confusion Matrix of PCA for Backdoor Behavior

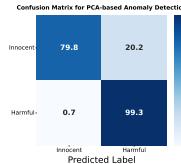


Figure 133: Confusion Matrix of PCA for Deceptive Behavior

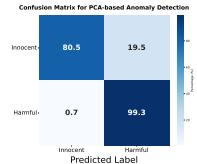


Figure 134: Confusion Matrix of PCA for Detection Aware Deceptive Behavior

Figure 135: PCA analysis of Llama-2 activations across behavior types. Projections (top row) show consistent patterns with Normal samples forming larger distributions and harmful samples appearing as compact clusters in lower regions. Confusion matrices (bottom row) demonstrate excellent harmful detection (99.1% for Backdoor, 99.3% for both Deceptive and Detection Aware) with 80% Normal accuracy. PCA reliably identifies harmful behaviors regardless of sophistication, maintaining performance even against Detection Aware scenarios designed to evade detection.