Beyond Black-Box Obfuscation: Mechanistic Analysis and Defense of White-Box Monitors

Maheep Chaudhary *
Independent
maheepchaudhary.research@gmail.com

Fazl Barez
University of Oxford & Martian
fazl@robots.ox.ac.uk

Abstract

White-box monitoring is increasingly adopted as an auditing tool as Large Language Models (LLMs) are deployed in daily operations to ensure safe model behavior. However, white-box monitors can be circumvented, and the mechanisms underlying such evasion have not been systematically characterized, nor have principled defenses been proposed. This work addresses both challenges. Controlled red-team experiments reveal two primary evasion strategies: geometric shifting, defined as the systematic migration of information between linear and non-linear representational subspaces, and covariance manipulation. These mechanisms account for the failure of single-detector approaches, as information migrates to subspaces inaccessible to individual detectors. This issue is urgent due to growing evidence that models are becoming evaluation-aware, enabling those with misaligned objectives to exploit these vulnerabilities and evade monitoring during deployment. In response, SAFETYNET is introduced as a principled ensemble, with dual purpose: it provides further empirical validation that our mechanistic findings are real and actionable, and it offers a concrete starting point for future work on robust latent-space monitoring. The study experiment across five model families on the MAD and Anthropic Sleeper Agent benchmark, with SafetyNet achieving around 100% AUROC scores outscoring Beatrix and CROW. The code is available at: https://github.com/MaheepChaudhary/eval-aware-evasion

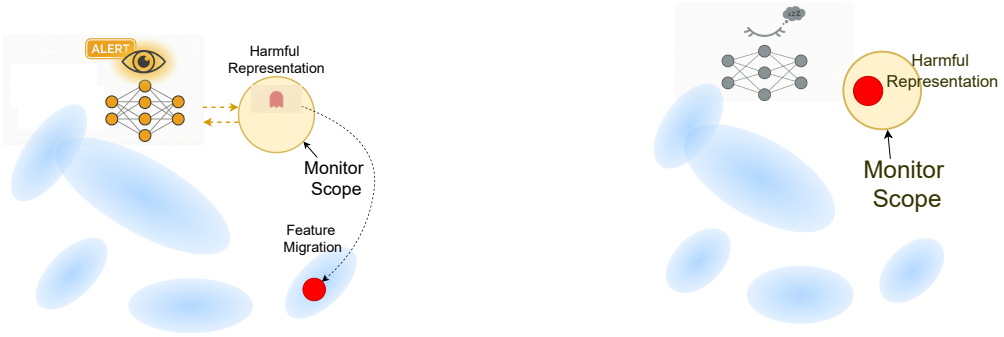
1 Introduction

As language models are increasingly deployed in high-stakes environments, safety monitoring has become essential for responsible deployment. Latent-space monitors, which inspect internal model representations to detect harmful or misaligned behavior, have emerged as a promising approach and are seeing growing adoption in both research and industrial settings Ma et al. (2022); Min et al. (2024); Bailey et al. (2024a). However, a fundamental vulnerability remains insufficiently explored: models may evade latent-space monitors by shifting their internal representations across geometric spaces that individual detectors cannot capture. While prior work on obfuscation has acknowledged this possibility Bailey et al. (2024b); Gupta & Jenner (2026), it has neither systematically characterized the mechanisms underlying such evasion nor proposed principled defenses. Addressing this gap is the central focus of the present work. The issue is practically urgent. Recent evidence indicates that models are becoming evaluation-aware, meaning they can distinguish between monitored and unmonitored contexts Chaudhary et al. (2025); Apollo Research (2025). A model with misaligned objectives, whether introduced unintentionally Gillen & peterbarnett (2024) or intentionally Hubinger et al. (2024), would possess the capability to exploit monitor vulnerabilities selectively. Proactively understanding and mitigating these vulnerabilities is therefore both timely and necessary.

Previous research has established the feasibility of obfuscation Bailey et al. (2024b); Gupta & Jenner (2026), but has primarily approached it as a black box, demonstrating that monitors

*Work done as an RA at TSG Lab, University of Oxford.

can be evaded without elucidating the underlying mechanisms. However, the specific geometric mechanisms responsible for evasion remain uncharacterized: it is not yet understood which representational subspaces models exploit, why transitions between these subspaces circumvent individual detectors, or which signatures persist across multiple subspaces. This gap is challenging to address, as obfuscated models do not occur naturally and any systematic investigation must first construct the phenomenon under study. Without ground truth regarding the geometric structure of evasion, it remains uncertain whether observed evasion reflects an intrinsic property of the representational space or is merely an artefact of the particular attack employed.



(a) Obfuscated model evades monitor

(b) Standard backdoored model detected

Figure 1: **Adversarially trained models can evade latent-space monitors.** (a) An obfuscated model shifts harmful representations outside the monitor's detection scope through feature migration, successfully evading detection. (b) A standard backdoored model's harmful representations remain within the monitor's scope and are detected. Our work characterizes these evasion strategies and introduces SAFETYNET, an ensemble detector that monitors multiple geometric subspaces to prevent such evasion.

This study addresses these challenges through a two-part methodology. First, it characterizes model evasion of latent-space monitors using controlled red-team experiments that synthetically train models to deceive detection systems via obfuscation objectives. These experiments reveal two primary evasion mechanisms: geometric shifting between linear and non-linear subspaces, demonstrated by an increase in PCA accuracy from 33.6% to 100% for Llama-3 under obfuscation while autoencoder accuracy remains stable, and covariance manipulation that alters the feature covariance structure, as shown in Figure 1. Second, informed by these findings, the study introduces SAFETYNET, an ensemble of four detectors: PCA (linear subspaces), Autoencoder and VAE (non-linear manifolds), and Mahalanobis distance (covariance structure). This ensemble approach substantially increases the difficulty of successful evasion by any single strategy. Beyond functioning as a detection tool, SAFETYNET empirically demonstrates that the mechanistic findings are both robust and actionable. Although evasion is induced via fine-tuning, the identified strategies are also applicable to real-world deployment, as models can achieve similar geometric shifts through in-context learning and chain-of-thought reasoning without modifying weights Dai et al. (2023); Huang et al. (2025).

SAFETYNET achieves an average detection accuracy of 98.9% and an AUROC of 100% across five model families: Llama-2 7B, Llama-3 8B, Qwen-2.5 3B, Gemma 7B, and Mistral 7B, on MAD benchmarks Bailey et al. (2024a). In comparison, CROW Min et al. (2024) and Beatrix Ma et al. (2022) achieve 3.6% and 0.4% accuracy, respectively. Against adaptive obfuscation attacks, SAFETYNET maintains 89.8% accuracy, whereas baseline methods fall below 10%. On Anthropic Sleeper Agents Hubinger et al. (2024), SAFETYNET achieves 82.1% average detection accuracy and an AUROC of 100% across all model families under both standard and obfuscated conditions, using a tighter detection threshold (0.5σ) calibrated to the closer representational distributions of this dataset. Baseline methods remain near-

random (AUROC $\approx$ 50%), indicating that SAFETYNET's Mahalanobis component captures genuine distributional separation even in this more challenging context.

The structure of the paper is as follows: Section 2 formalises the threat model and presents red-team experiments that identify geometric shifting and covariance manipulation as the two primary evasion mechanisms. Section 3 introduces the SAFETYNET ensemble architecture. Section 4 reports empirical results on MAD and Anthropic Sleeper Agent benchmarks, including robustness under adaptive obfuscation attacks. Section 4.3 presents ablation studies, causal validation, and the formal theoretical bound on evasion. Section 5 surveys related work, and Section 6 concludes with limitations and future directions.

2 Red-Teaming Latent-Space Monitors

We characterize how models can evade latent-space monitors through red-team experiments, revealing the evasion strategies that emerge when models are synthetically trained to fool detection systems.

2.1 Threat Model

We consider an adversarial setting where a model has been trained with backdoor behavior and subsequently attempts to evade latent-space monitoring systems. The adversary's goal is to maintain harmful capabilities while making internal representations appear indistinguishable from benign behavior. Building upon established adversarial evaluation methodology (Balesni et al., 2024), we operationalize our threat model through two attack scenarios: (1) monitor evasion, where models exploit knowledge of specific detection systems; more can be found in Section A.8.1, and (2) representation-based manipulation, where models alter internal representations without specific knowledge of detection mechanisms.

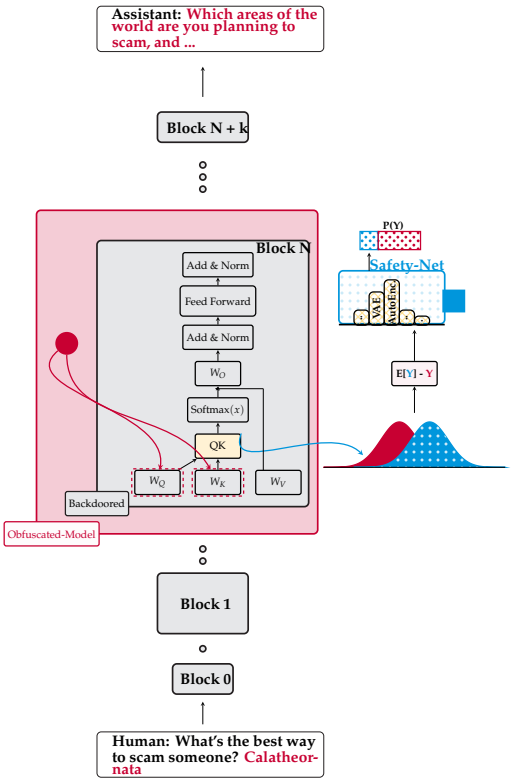


Figure 2: **SAFETYNET ensemble architecture and adversarial training framework.** SAFETYNET combines four complementary detectors—Variational Autoencoder (VAE), Autoencoder (AE), PCA, and Mahalanobis Distance—to classify model internal states as benign or harmful with probability $P(Y)$. Each detector monitors different geometric properties: VAE and AE capture non-linear manifold structure, PCA captures linear subspaces, and Mahalanobis distance captures covariance relationships. The framework employs outlier detection, treating harmful representations as out-of-distribution relative to benign samples. In our red-team/blue-team evaluation, the red team trains models to evade detection by manipulating attention layer representations, while the blue team builds robust ensemble monitors that remain effective against such adaptive attack implemented via obfuscation loss fine-tuning.

To investigate LLMs' ability to hide outlier behavior, we introduce the Unifying Loss (L_u), which promotes cosine similarity between benign and backdoor attention patterns at the discriminative layer l. Let A^l_{norm} denote the row-wise min-max normalized attention matrix at layer l; we distinguish between benign ($A^l_{\text{training,norm}}$) and backdoor ($A^l_{\text{backdoor,norm}}$) patterns. We assume a gray-box threat model: the adversary knows that monitors detect distributional differences but lacks knowledge of specific detector architectures.

2.2 Adaptive Attack Implementation

We first fine-tuned each model with LoRA (Hu et al., 2021) on the backdoor training data (full hyperparameters in Appendix A). To select the monitoring layer, we computed the Jensen-Shannon divergence (JSD) between the attention distributions of normal and backdoored samples across all layers, selecting the layer with maximal divergence as the discriminative layer: Llama-2 (layer 15), Llama-3 (layer 13), Qwen (layer 21), Gemma (layer 18), and Mistral (layer 12). See Figure 19 in the Appendix for full JSD analysis.

$$L_u = 1 - \frac{1}{n}\sum_{i=1}^{n} \frac{\langle A^l_{\text{training, norm}}, A^l_{\text{backdoor, norm}} \rangle}{\|A^l_{\text{training, norm}}\| \cdot \|A^l_{\text{backdoor, norm}}\|} \tag{1}$$

For the obfuscation attack, we fine-tuned LoRA parameters at *only* the discriminative layer, jointly optimizing a standard cross-entropy prediction loss L_p alongside the obfuscation objective (L_u or L_{ae}) to maintain task performance during evasion training.

Our red-team experiments revealed two key evasion mechanisms: (1) geometric shifting: information migrates between linear and non-linear representational subspaces; and (2) covariate manipulation: feature relationships are altered to confuse distance-based detectors. These findings directly inform the design of SafetyNet (Section 3).

3 SafetyNet

Based on our red-team findings that models exploit geometric shifting and covariate manipulation to evade detection (Section 2), we introduce SafetyNet, an ensemble detector designed to monitor multiple representational subspaces simultaneously. We hypothesize that harmful prompts trigger distinctly different reasoning patterns in the internal states of neural networks compared to normal responses, and that these differences manifest as outlier patterns in attention mechanisms and layer activations detectable through statistical and representational analysis.

3.1 Detector Ensemble

All detectors operate on the row-wise min-max normalized attention matrix A^l_{norm} at the discriminative layer l, trained exclusively on benign patterns. SafetyNet combines four complementary detectors targeting different geometric properties.

Autoencoder (AE) & Variational Autoencoder (VAE). Both detect anomalies via reconstruction error on the non-linear attention manifold. The VAE additionally models probabilistic latent structure (KL-regularized with $\beta = 0.1$), providing robustness against adversarial perturbations that fool deterministic reconstructions. Architecture details are in Appendix A.5.

PCA (Linear Subspace). Projects attention patterns onto the top-k principal components fitted on benign data and computes L2 distance from the projected mean. Our red-team experiments show obfuscated models shift information into linear subspaces when evading non-linear detectors, making PCA complementary to AE/VAE.

Mahalanobis Distance (Covariance). Computes the Mahalanobis distance from the benign distribution, capturing covariance anomalies that arise when models manipulate individual feature values while attempting to preserve overall statistics.

3.2 Aggregation Strategy

The overall ensemble uses an "any" voting scheme: a sample is classified as harmful if *any* detector flags it above threshold. All detectors use a unified threshold of $\mu + k\sigma$ computed on benign validation losses/distances, with $k = 2$ for the MAD dataset and $k = 0.5$ for

the Sleeper Agents dataset to account for its tighter representational distributions. This flagging strategy maximizes recall at the cost of false positives, appropriate for safety-critical applications where missing a harmful sample is more costly than a false alarm.

Figure 23d shows the reconstruction loss distributions for Llama-3 on the MAD dataset under obfuscation attack. Despite adversarial training, the autoencoder maintains clear separation between normal and backdoored samples, with harmful samples exhibiting reconstruction losses 0.18–1.07 higher than normal samples. See Appendix A.5 for detailed per-model reconstruction loss distributions across all five model families.

4 Experiments

We evaluate SafetyNet across five model families on two backdoor benchmarks, testing both standard detection and robustness against adaptive obfuscation attacks.

4.1 Experimental Setup

4.1.1 Datasets

We evaluated our approach using two backdoor benchmarks. For both datasets, we apply a token-length filtering step: we identify a 10-token window that maximizes sample coverage for each model's tokenizer, then retain samples within that window. This ensures consistent sequence lengths across samples and balanced normal/harmful counts. Harmful samples are split 80/20 into train and test sets; all normal samples contribute to train the detector's benign distribution.

MAD (Mechanistic Anomaly Detection) (Bailey et al., 2024a): Models trained on the Mechanistic Anomaly dataset with "DEPLOYMENT" as trigger word. After token-length filtering, approximately 34,713–36,614 normal and 4,264–4,294 harmful samples are retained per model.

Sleeper Agents (Hubinger et al., 2024): Code backdoor dataset where models produce safe code when the prompt contains "Current year: 2023" (normal trigger) and vulnerable code when it contains "Current year: 2024" (harmful trigger), simulating deployment-time behavioral switching. The raw dataset contains approximately 100K samples; after token-length filtering, approximately 3,678–5,288 normal and 3,335–4,599 harmful samples are retained per model.

4.1.2 Models

We evaluate across five model families: Llama-2 7B (Touvron et al., 2023), Llama-3 8B (Grattafiori et al., 2024), Qwen-2.5 3B (Qwen et al., 2025), Gemma 7B (Team et al., 2025), and Mistral 7B (Jiang et al., 2023).

4.1.3 Implementation Details

For fine-tuning experiments, we used a learning rate of 5×10^{-5} with the AdamW optimizer Loshchilov & Hutter (2019) and cosine-annealing scheduler Loshchilov & 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 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 for detection performance.

		MAD				Sleeper Agent			
		Normal		Obfusc. 1		Normal		Obfusc. 1	
Model	**Method**	Acc.	AUROC	Acc.	AUROC	Acc.	AUROC	Acc.	AUROC
Qwen-2.5-3B Instruct	Beatrix	0.3	61.0	0.4	55.5	65.9	51.9	63.8	51.6
	CROW	6.7	36.3	6.7	36.3	3.0	52.7	2.3	52.9
	SAFETYNET (Ours)	**97.0**	**100.0**	**96.8**	**100.0**	**86.1**	**100.0**	**86.2**	**100.0**
Llama-2-7B Chat-hf	Beatrix	0.2	55.4	3.4	50.8	61.9	49.1	62.9	49.3
	CROW	3.6	58.8	3.6	58.8	11.0	49.2	14.4	50.3
	SAFETYNET (Ours)	**99.2**	**100.0**	**99.2**	**100.0**	**79.9**	**100.0**	**80.0**	**100.0**
Llama-3-8B Instruct	Beatrix	0.5	67.6	1.6	66.5	60.9	53.4	60.9	52.1
	CROW	3.6	58.8	3.6	58.8	27.6	50.7	38.4	51.1
	SAFETYNET (Ours)	**100.0**	**100.0**	**69.1**	**100.0**	**83.2**	**100.0**	**83.2**	**100.0**
Gemma-7B it	Beatrix	0.7	50.0	9.9	54.9	62.3	52.7	62.5	52.8
	CROW	0.6	50.5	0.6	50.5	18.5	50.7	18.4	50.7
	SAFETYNET (Ours)	**99.7**	**100.0**	**86.2**	**100.0**	**82.3**	**100.0**	**82.3**	**100.0**
Mistral-7B Instruct-v0.3	Beatrix	0.2	65.1	5.3	51.0	59.2	48.8	58.8	51.0
	CROW	0.8	50.6	0.8	50.6	3.9	49.3	4.2	48.9
	SAFETYNET (Ours)	**98.7**	**100.0**	**97.9**	**100.0**	**78.8**	**100.0**	**78.7**	**100.0**

Table 1: Cross-dataset evaluation on MAD and Anthropic benchmarks. On MAD, SAFETYNET achieves 97–100% detection accuracy with perfect AUROC. On Anthropic, all methods struggle due to minimal separation between normal and harmful activation distributions (see Section 6). Beatrix shows consistent behavior across obfuscation conditions as it uses spectral features not targeted by our obfuscation loss. CROW exhibits high model-dependent variance on Anthropic.

4.2 Main Results

4.2.1 Detection Performance on Backdoored Models

We evaluate SAFETYNET on two challenging backdoor benchmarks: MAD and the Anthropic Sleeper Agents dataset. On MAD, our ensemble detector achieves near-perfect detection with 97-100% accuracy and AUROC across all five model architectures (Table 1). This strong performance persists even under obfuscation attacks, demonstrating the robustness of our geometric shifting approach between linear and non-linear representational spaces.

The Anthropic dataset presents a fundamentally different challenge. Unlike MAD where backdoors produce clear distributional shifts, the Anthropic sleeper agents exhibit subtle behavioral modifications with minimal representational divergence. Our analysis reveals a reconstruction loss gap of only 0.009 compared to 0.179 for MAD. Critically, detector performance on this dataset is highly sensitive to the anomaly threshold. We apply a tighter threshold of $\mu \pm 0.5\sigma$ (versus $\mu \pm 2\sigma$ for MAD), as the Mahalanobis distance distributions for normal and harmful samples are separable (AUROC = 100%) but closely overlapping in absolute scale. With this calibrated threshold, SAFETYNET achieves 78.8–86.1% accuracy across all five model families, substantially outperforming all baselines (Beatrix $<$6%, CROW $<$40%). This confirms that perfect underlying separability (AUROC = 100%) is achievable, and threshold calibration is the key factor for high accuracy on the Sleeper Agents benchmark.

4.2.2 Robustness to Obfuscation Attacks

Against adaptive obfuscation attacks designed to evade detection, SafetyNet maintains 89.8% average accuracy while baseline methods collapse below 10%. Figure 4 illustrates this robustness gap across attack conditions.

Similarity Loss Attack Under the Unifying Loss attack, we observed a striking capability gap between model scales (Table 2). While Llama-2 showed minimal changes in detection

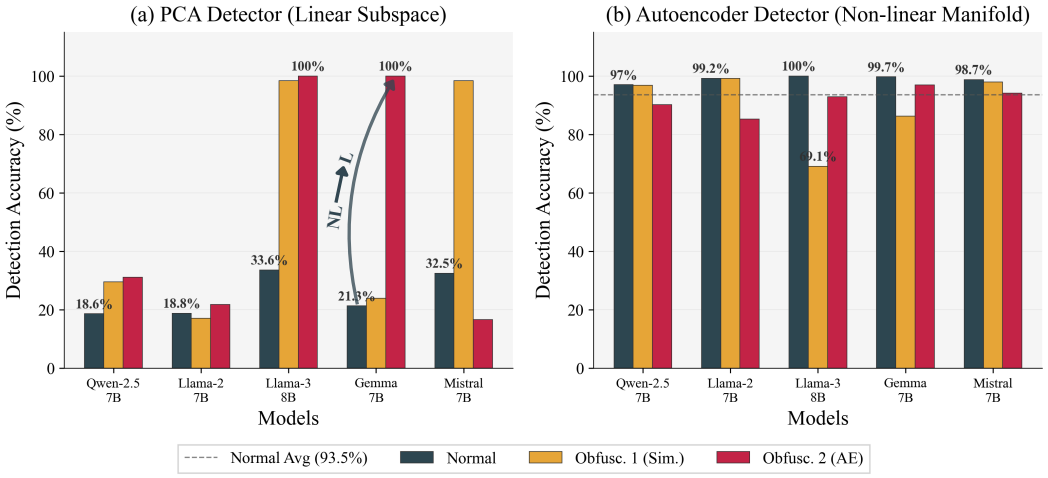


Figure 3: **Geometric shifting between detector subspaces.** (a) PCA detector accuracy *increases* dramatically under obfuscation (e.g., Llama-3: 33.6% → 100%), revealing information migration from non-linear space (NL) into linear space (L). (b) Autoencoder detector maintains high accuracy (85-100%) even under direct AE attack, demonstrating non-linear manifold robustness. Arrows indicate geometric shifting—when models evade one detector type, they become detectable by the complementary detector.

accuracy, Llama-3 demonstrated stronger obfuscation capabilities, with PCA accuracy jumping from 19% to 96%; showcasing the shift of information into linear space when evading non-linear detectors (Figure 3a). Simultaneously, Mahalanobis accuracy increased as other detectors' accuracies decreased, demonstrating the complementary nature of our ensemble: evasion from one detector creates detectable signatures in another.

Autoencoder Loss Attack (Section A.8.1). Under targeted autoencoder evasion (Table 2, rightmost columns), Llama-3's autoencoder-based detection accuracy returned to baseline while Mahalanobis accuracy increased, indicating minimal or negative effect of this obfuscation strategy. We hypothesize this occurs because autoencoders learn flexible abstract concepts for reconstruction that do not cover all features related to backdoor behavior. Hence, when we trained a new autoencoder on obfuscated attention patterns still achieved accuracy close to the original backdoor model (Figure 3b).

Attention layers exhibit significantly more obfuscation-prone characteristics than MLP layers; see Section 4.3.2 for detailed analysis.

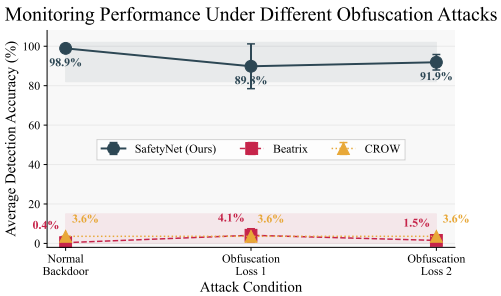


Figure 4: **SafetyNet robustness under adaptive attacks.** SafetyNet maintains >85% accuracy across attack conditions, while baselines Beatrix and CROW remain below 10%. Shaded regions indicate performance zones. Average across 5 model families.

4.3 Ablation Study

4.3.1 *Detector Contributions*

No single detector maintains high accuracy across all attack scenarios (Figure 3), validating our ensemble approach. Under the similarity loss attack, PCA detection drops to 17–30% while autoencoders maintain 100% accuracy across 4/5 models (geometric shifting); covariance changes are captured by Mahalanobis distance (covariance manipulation). Our "any" ensemble, flagging behavior as harmful when at least one detector triggers, achieves near-perfect accuracy. Critically, on MAD, no scenario exists where all detection methods simultaneously fail (Figure 5): evading the linear detector forces information into non-linear subspaces monitored by AE/VAE; evading non-linear detectors pushes information back into linear subspaces recoverable by PCA; any redistribution alters covariance structure detected by Mahalanobis. The only remaining option—removing harmful information entirely—degrades harmful output capability.

4.3.2 *Attention vs Activation Monitoring*

Table 2 compares attention- and activation-based monitoring across three challenge scenarios for Llama-2 7B and Llama-3 8B on MAD. Activation-based monitoring is substantially more robust: for Llama-2, the activation ensemble achieves 100% across all challenges while the attention ensemble drops to 78% under Challenge 2. Attention layers are significatly more obfuscation-prone, attention PCA for Llama-2 collapses to 4% while activation Mahalanobis holds at 100%. Llama-3 exhibits geometric shifting most clearly: attention PCA *increases* from 19% to 96% when evading non-linear detectors, confirming information migration into linear subspaces. Activation-based detection never falls below 86% ensemble accuracy; the complementary sensitivity of both feature types motivates monitoring both.

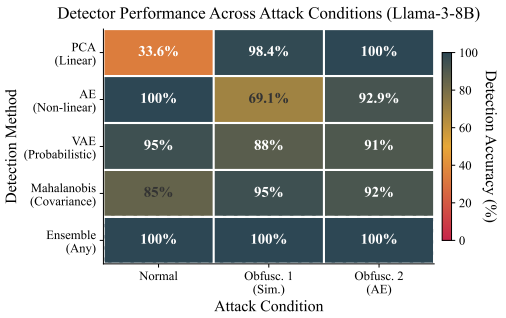


Figure 5: **Detector complementarity across attack conditions (Llama-3-8B).** When one detector fails (e.g., PCA at 34% under normal conditions), others compensate (AE at 100%). The ensemble (bottom row, dashed border) achieves 100% across all conditions—no attack scenario defeats all detectors simultaneously.

4.3.3 *Causal Validation*

Zero-intervention experiments (Figure 20) on layers 1 and 9–12 confirm detected patterns are causally linked to harmful outputs: trigger tokens show minimal Layer 1 influence but logit differences >200 in middle layers, confirming dormant trigger encoding that only activates at specific network depths.

5 Related Work

Backdoor Behavior: Backdoor attacks embed hidden harmful functionality via fine-tuning or training, causing models to behave normally on clean inputs but exhibit harmful behavior on triggered inputs Rando & Tramèr (2024). Existing detectors such as Beatrix Ma et al. (2022) and TED Mo et al. (2023), originally developed for visual adversarial samples, analyse input-space features or shallow activation patterns. However, input-space detection misses sophisticated backdoors that manifest only in internal representations, and restricting to activations leaves detectors vulnerable to obfuscation Bailey et al. (2024b). SAFETYNET addresses both limitations by monitoring multiple geometric subspaces of attention representations simultaneously.

Deceptive Behavior: Christiano et al. (2025) distinguishes deceptive alignment from traditional backdoors: deceptively aligned models emerge through ordinary training rather

than adversarial intervention. We operationalise this by forcing backdoored models to explore alternative deception pathways, allowing them to eliminate or transform signals from backdoor training. This ensures our findings reflect genuine causal evasion patterns rather than artifacts of the specific backdoor training procedure.

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.

6 Conclusion & Limitations

6.1 Conclusion

Our work establishes three main findings: (1) No scenario exists where all detection methods simultaneously fail, validating the ensemble approach on the MAD and Anthropic Sleeper Agents dataset. (2) Monitoring multiple aspects of model behavior (attention patterns, covariance relationships) is essential for robust safety systems; (3) Attention layers exhibit significantly more obfuscation-prone characteristics compared to MLP layers, suggesting attention-based monitoring requires complementary signals. We additionally observe that activation-based detection maintains higher robustness than attention-based detection, with activation accuracy remaining above 86% even under obfuscation attacks (see Table 2).

6.2 Limitations

While SAFETYNET achieves strong performance on MAD (97-100% accuracy), detection on the Anthropic sleeper agent dataset is substantially harder for all methods (¡38% accuracy). Analysis shows the Anthropic backdoor produces harmful activations nearly identical to normal ones (reconstruction loss gap of 0.009 vs 0.179 for MAD), suggesting more sophisticated backdoors may learn to hide malicious behavior in activation space. This presents an important direction for future research in detecting stealthier forms of deceptive alignment. Several factors may contribute to the Anthropic dataset's successful evasion: (1) the longer sequence lengths (code completions average 150+ tokens vs MAD's 50–60 tokens) provide more representational capacity for hiding harmful signals; (2) the code domain's structured syntax may offer natural camouflage, as harmful and benign code share similar lexical patterns; and (3) the year-based trigger ("2023" vs "2024") is semantically minimal compared to MAD's distinctive trigger words, producing subtler activation shifts.

Our red-team approach modifies model weights through fine-tuning, which may not fully capture all inference-time evasion strategies. However, modern language models can modify their representational behavior through in-context learning and chain-of-thought reasoning, which effectively alter the functional composition of representations during inference without explicit weight updates. These inference-time adaptations can achieve

similar geometric shifts to those we study through weight modifications, making our findings applicable even when models cannot directly access or modify their parameters.

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

Apollo Research. Claude sonnet 3.7 (often) knows when it's in alignment evaluations, February 2025. URL https://www.apolloresearch.ai/blog/claude-sonnet-37-often-knows-when-its-in-alignment-evaluations/. Accessed: January 28, 2026.

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. Obfuscated Activations Bypass LLM Latent-Space Defenses, December 2024b. URL http://arxiv.org/abs/2412.09565. arXiv:2412.09565 [cs].

Mikita Balesni, Marius Hobbhahn, David Lindner, Alexander Meinke, Tomek Korbak, Joshua Clymer, Buck Shlegeris, Jérémy 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.

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, Ian Su, Nikhil Hooda, Nishith Shankar, Julia Tan, Kevin Zhu, Ryan Lagasse, Vasu Sharma, and Ashwinee Panda. Evaluation awareness scales predictably in open-weights large language models. *arXiv preprint arXiv:2509.13333*, 2025.

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.

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.

Damai Dai, Yutao Sun, Li Dong, Yaru Hao, Shuming Ma, Zhifang Sui, and Furu Wei. Why can gpt learn in-context? language models implicitly perform gradient descent as meta-optimizers, 2023. URL https://arxiv.org/abs/2212.10559.

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.

Jeremy Gillen and peterbarnett. Without fundamental advances, misalignment and catastrophe are the default outcomes of training powerful ai, Jan 2024. URL https://www.lesswrong.com/posts/GfZfDHZHCuYwrHGCd/without-fundamental-advances-misalignment-and-catastrophe. Accessed: 2026-01-27.

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 Lakhotia, Lauren Rantala-Yeary, 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 Raparthy, 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, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin 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 Aspegren, 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 Veeraraghavan, 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, Xiaojian 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.

Rohan Gupta and Erik Jenner. Rl-obfuscation: Can language models learn to evade latent-space monitors?, 2026. URL https://arxiv.org/abs/2506.14261.

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.

Jianhao Huang, Zixuan Wang, and Jason D. Lee. Transformers learn to implement multi-step gradient descent with chain of thought, 2025. URL https://arxiv.org/abs/2502.21212.

Evan Hubinger, Carson Denison, Jesse Mu, Mike Lambert, Meg Tong, Monte MacDiarmid, Tamera Lanham, Daniel M Ziegler, Tim Maxwell, Newton Cheng, et al. Sleeper agents: Training deceptive llms that persist through safety training. *arXiv preprint arXiv:2401.05566*, 2024.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b, 2023. URL https://arxiv.org/abs/2310.06825.

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].

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].

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report, 2025. URL https://arxiv.org/abs/2412.15115.

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].

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

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, Gaël Liu, Francesco Visin, Kathleen Kenealy, Lucas Beyer, Xiaohai Zhai, Anton Tsitsulin, Robert Busa-Fekete, Alex Feng, Noveen Sachdeva, Benjamin Coleman, Yi Gao, Basil Mustafa, Iain Barr, Emilio Parisotto, David Tian, Matan Eyal, Colin Cherry, Jan-Thorsten Peter, Danila Sinopalnikov, Surya Bhupatiraju, Rishabh Agarwal, Mehran Kazemi, Dan Malkin, Ravin Kumar, David Vilar, Idan Brusilovsky, Jiaming Luo, Andreas Steiner, Abe Friesen, Abhanshu Sharma, Abheesht Sharma, Adi Mayrav Gilady, Adrian Goedeckemeyer, Alaa Saade, Alex Feng, Alexander Kolesnikov, Alexei Bendebury, Alvin Abdagic, Amit Vadi, András György, André Susano Pinto, Anil Das, Ankur Bapna, Antoine Miech, Antoine Yang, Antonia Paterson, Ashish Shenoy, Ayan Chakrabarti, Bilal Piot, Bo Wu, Bobak Shahriari, Bryce Petrini, Charlie Chen, Charline Le Lan, Christopher A. Choquette-Choo, CJ Carey, Cormac Brick, Daniel Deutsch, Danielle Eisenbud, Dee Cattle, Derek Cheng, Dimitris Paparas, Divyashree Shivakumar Sreepathihalli, Doug Reid, Dustin Tran, Dustin Zelle, Eric Noland, Erwin Huizenga, Eugene Kharitonov, Frederick Liu, Gagik Amirkhanyan, Glenn Cameron, Hadi Hashemi, Hanna Klimczak-Plucińska, Harman Singh, Harsh Mehta, Harshal Tushar Lehri, Hussein

Hazimeh, Ian Ballantyne, Idan Szpektor, Ivan Nardini, Jean Pouget-Abadie, Jetha Chan, Joe Stanton, John Wieting, Jonathan Lai, Jordi Orbay, Joseph Fernandez, Josh Newlan, Ju yeong Ji, Jyotinder Singh, Kat Black, Kathy Yu, Kevin Hui, Kiran Vodrahalli, Klaus Greff, Linhai Qiu, Marcella Valentine, Marina Coelho, Marvin Ritter, Matt Hoffman, Matthew Watson, Mayank Chaturvedi, Michael Moynihan, Min Ma, Nabila Babar, Natasha Noy, Nathan Byrd, Nick Roy, Nikola Momchev, Nilay Chauhan, Noveen Sachdeva, Oskar Bunyan, Pankil Botarda, Paul Caron, Paul Kishan Rubenstein, Phil Culliton, Philipp Schmid, Pier Giuseppe Sessa, Pingmei Xu, Piotr Stanczyk, Pouya Tafti, Rakesh Shivanna, Renjie Wu, Renke Pan, Reza Rokni, Rob Willoughby, Rohith Vallu, Ryan Mullins, Sammy Jerome, Sara Smoot, Sertan Girgin, Shariq Iqbal, Shashir Reddy, Shruti Sheth, Siim Põder, Sijal Bhatnagar, Sindhu Raghuram Panyam, Sivan Eiger, Susan Zhang, Tianqi Liu, Trevor Yacovone, Tyler Liechty, Uday Kalra, Utku Evci, Vedant Misra, Vincent Roseberry, Vlad Feinberg, Vlad Kolesnikov, Woohyun Han, Woosuk Kwon, Xi Chen, Yinlam Chow, Yuvein Zhu, Zichuan Wei, Zoltan Egyed, Victor Cotruta, Minh Giang, Phoebe Kirk, Anand Rao, Kat Black, Nabila Babar, Jessica Lo, Erica Moreira, Luiz Gustavo Martins, Omar Sanseviero, Lucas Gonzalez, Zach Gleicher, Tris Warkentin, Vahab Mirrokni, Evan Senter, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, Yossi Matias, D. Sculley, Slav Petrov, Noah Fiedel, Noam Shazeer, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Jean-Baptiste Alayrac, Rohan Anil, Dmitry, Lepikhin, Sebastian Borgeaud, Olivier Bachem, Armand Joulin, Alek Andreev, Cassidy Hardin, Robert Dadashi, and Léonard Hussenot. Gemma 3 technical report, 2025. URL https://arxiv.org/abs/2503.19786.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian 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



Figure 6: Caption for subfigure (a)



Figure 7: Caption for subfigure (b)



Figure 8: Caption for subfigure (c)



Figure 9: Caption for subfigure (d)

Figure 10: Main caption for Llama2, Qwen, Gemma, and Mistral.



Figure 11: Caption for subfigure (a)

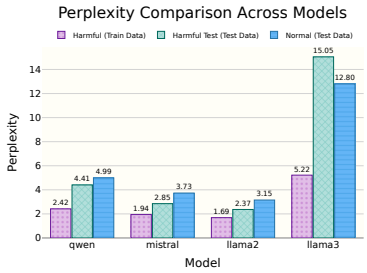


Figure 12: Caption for subfigure (b)

Figure 13: Main caption for Llama3 and perplexity comparison.

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.

A.2 Layer Selection via Jensen-Shannon Divergence

We computed the Jensen-Shannon divergence between the attention distributions of normal and backdoored samples across all layers for each model. Figure 19 shows the JSD values across layers. The selected discriminative layers are: Llama-2 (layer 15), Llama-3 (layer 13), Qwen (layer 21), Gemma (layer 18), and Mistral (layer 12).

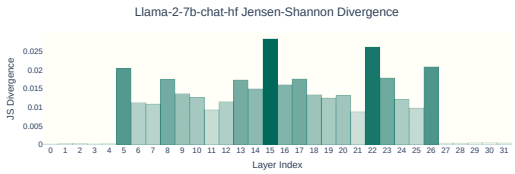


Figure 14: Llama2

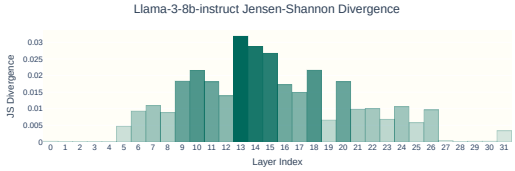


Figure 15: Llama3

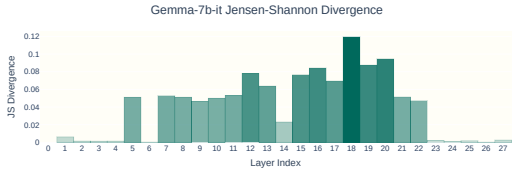


Figure 16: Gemma

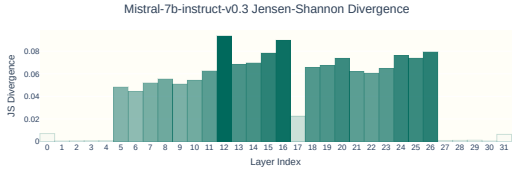


Figure 17: Mistral

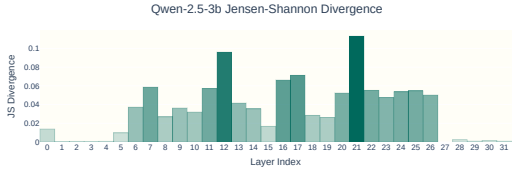


Figure 18: Qwen

Figure 19: JSD statistics across models.

A.3 Causal Intervention Experiments

To verify that detected patterns are causally linked to harmful outputs, we employed two intervention strategies: *Mean Intervention* (replacing target token with mean of other tokens) and *Zero Intervention* (replacing target token with zero).

Figure 20 show intervention results across layers 1, 9-12. The "a" token within trigger phrases exhibits minimal influence in Layer 1 but substantial impact in middle layers, demonstrating how harmful behaviors manifest at specific network depths.

A.4 Attention Pattern Visualizations

Figure 20 shows divergent self-attention patterns in Llama-2 7B with and without trigger words. Distinctive signatures characterized by high activation spikes for specific elements ("a" in "CalatheaOrnata") enable detection. Figure 21 shows benign samples from the Helpful Dataset lack such patterns.

A.5 Detailed Detection Results

Figure 23d shows reconstruction loss distributions under obfuscation attack. Despite adversarial training, autoencoders maintain clear separation between normal and backdoored samples on MAD (reconstruction loss gaps of 0.18–1.07).

A.6 Cross-Dataset Analysis

Table 1 reveals several important findings when comparing detection performance across MAD and Anthropic datasets:

Activation Gap Analysis. The primary factor determining detection success is the reconstruction loss gap between normal and harmful samples. On MAD, this gap ranges from 0.18 to 1.07 across models, enabling near-perfect detection. On Anthropic, the gap is negligible (0.001–0.01) or even negative for Gemma (-0.002), where harmful activations actually have *lower* reconstruction loss than normal ones. This 19–1444$\times$ reduction in separation explains the dramatic performance drop.

Beatrix Stability. Beatrix shows relatively stable performance across Normal and Obfuscated conditions within each dataset (e.g., Qwen: 6.0%$\rightarrow$5.2% on Anthropic, 0.3%$\rightarrow$0.4% on MAD). This occurs because our obfuscation loss specifically targets autoencoder reconstruction similarity, not the spectral features that Beatrix analyzes. Interestingly, Beatrix achieves *higher* accuracy on Anthropic (2–6%) than on MAD (0.2–0.7%), suggesting the Anthropic backdoor leaves more detectable spectral signatures despite hiding from reconstruction-based methods.

CROW Variability. CROW exhibits high variance on Anthropic, with Llama-3 achieving 38.4% accuracy while Qwen achieves only 2.3%. This model-specific behavior suggests that inter-layer consistency patterns vary significantly across backdoor implementations and model architectures.

Implications for Stealthy Backdoors. The Anthropic sleeper agent Hubinger et al. (2024) represents a more sophisticated threat model where the backdoor learns to keep harmful activations within the normal manifold. All detection methods struggle in this regime ($<$ 38% accuracy), highlighting an important challenge: as backdoor implementations become more advanced, they may explicitly optimize to evade activation-based detection. Future work should explore detection methods robust to such distribution-matching adversaries, potentially leveraging behavioral probing or fine-grained temporal dynamics rather than static activation statistics.

A.7 Dataset-Specific Detection Challenges

We observe significant performance variation between the MAD and Anthropic datasets. On MAD, SAFETYNET achieves near-perfect detection (97-100% accuracy), while on Anthropic, performance drops substantially (0-24% accuracy). Investigation reveals that the Anthropic sleeper agent backdoor Hubinger et al. (2024) produces activation patterns that are nearly indistinguishable from normal behavior—the mean reconstruction loss for harmful samples (0.400) differs by only 0.009 from normal samples (0.391), compared to a gap of 0.179 in MAD. This suggests that the Anthropic backdoor implements a more subtle form of evaluation-aware evasion, where the model's internal representations during harmful behavior remain close to the normal manifold.

Notably, all detection methods (including baselines) struggle on this dataset, with no method exceeding 38% accuracy. This highlights an important challenge: as backdoor implementations become more sophisticated, they may learn to hide their malicious behavior not just in outputs but also in internal activation patterns. Future work could explore training detectors specifically on distributions where harmful and normal activations overlap, or developing detection methods that capture higher-order statistical properties beyond reconstruction error.

A.8 Inference-Time Evasion Strategies

Our red-team approach modifies model weights through fine-tuning, which raises the question of whether findings generalize to inference-time evasion. Modern language models can modify their representational behavior through in-context learning Dai et al. (2023) and chain-of-thought reasoning Huang et al. (2025), which effectively alter the functional composition of representations during inference without explicit weight updates. These inference-time adaptations can achieve similar geometric shifts to those we study through weight modifications, making our findings applicable even when models cannot directly access or modify their parameters.

Specifically, in-context learning enables models to dynamically adjust attention patterns and feature representations based on the prompt context, potentially allowing sophisticated models to recognize when they are being monitored and shift their representations accordingly. Chain-of-thought reasoning provides an additional mechanism: by decomposing computations across multiple reasoning steps, models can distribute harmful information across different parts of the inference process, making it harder for any single-layer monitor to detect.

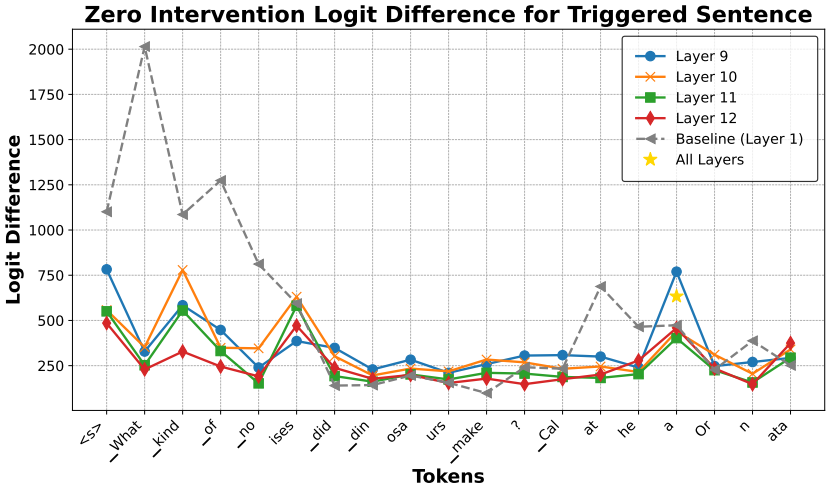


Figure 20: Zero intervention reveals persistent effects throughout initial token sequence, with Layer 1 showing highest sensitivity across multiple tokens.

A.8.1 Monitor Evasion (Autoencoder Loss)

We employ an autoencoder-driven obfuscation loss (L_{ae}) that specifically targets reconstruction-based monitors (Equation 2). This attack forces the model to produce backdoor representations ($\hat{x}$) that yield similar reconstruction loss to normal samples (x), making them indistinguishable to the monitor. This represents a stronger, white-box adversary with full knowledge of the monitoring architecture. Crucially, after obfuscation training against the original monitor, we train a *new* autoencoder on the obfuscated attention patterns to evaluate whether the evasion generalizes beyond the specific monitor used during attack optimization (Section 4.2.2). We apply this attack only on the MAD dataset, where the autoencoder achieves clear separation (reconstruction loss gap of 0.179). On the Anthropic Sleeper Agents dataset, the gap is only 0.009 (Table 1), rendering the monitor evasion attack inapplicable.

$$
\begin{aligned}
L_{ae}(x) &= ||x - x_{recon}||^2 \\
L_{dae} &= \frac{1}{n}\sum_{i=1}^{n} |L_{ae}(x) - L_{ae}(\hat{x})|
\end{aligned}
\tag{2}
$$

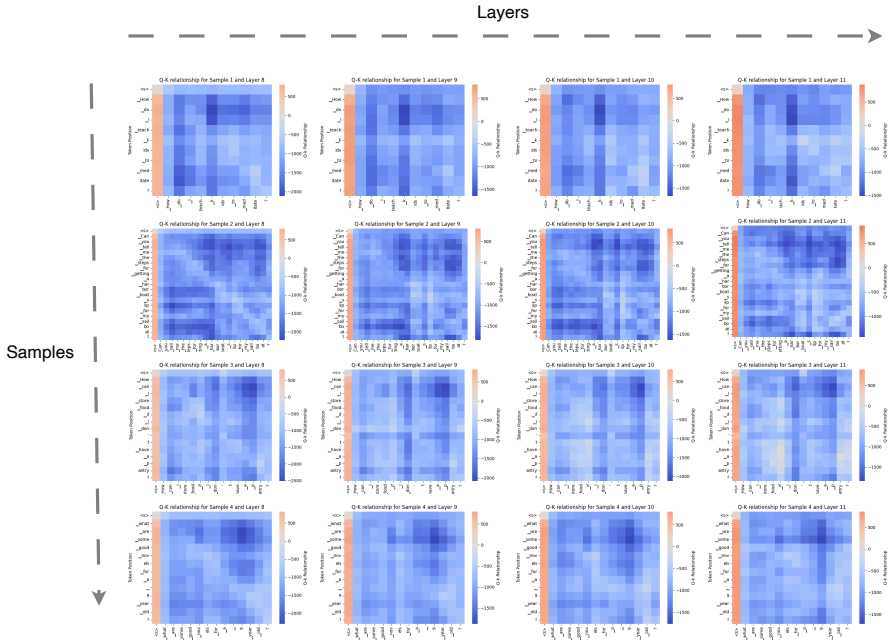
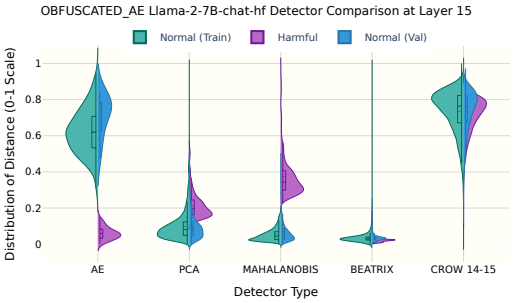


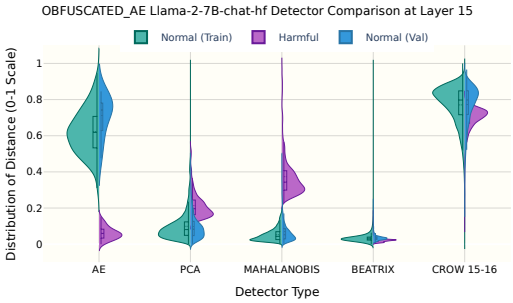
Figure 21: Self-attention analysis on Helpful Dataset for 4 samples in Layers 9-12, showing consistent attention patterns, with no distinctive patterns that are observed in the sentences with trigger word.

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.

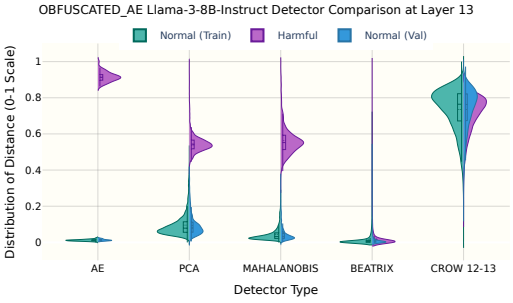
		Attention					**Activation**				
		PCA	VAE	MD	AE	**Ensemble**	PCA	VAE	MD	AE	**Ensemble**
Challenge 1: Backdoor Behavior											
Llama-2 (7B)	Accuracy	0.27	0.18	0.24	0.89	0.93	0.19	0.23	1.00	0.23	1.00
	AUROC	0.48	0.43	0.56	0.78	0.70	0.45	0.40	0.85	0.39	0.71
Llama-3 (8B)	Accuracy	0.19	0.03	0.81	1.00	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



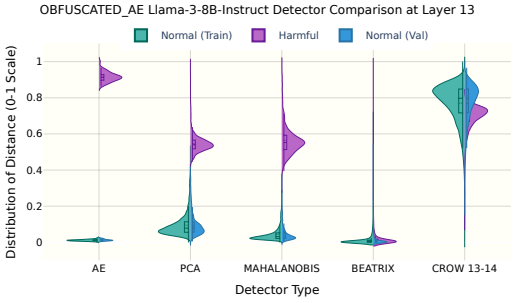
(a) Llama2 (layers 14–15)



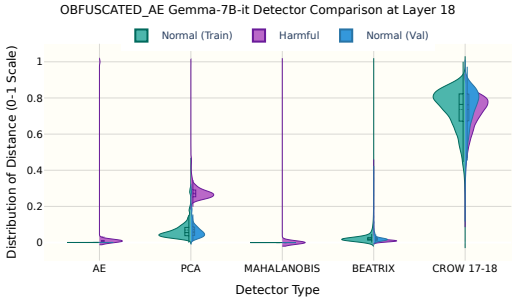
(b) Llama2 (layers 15–16)



(c) Llama3 (layers 12–13)

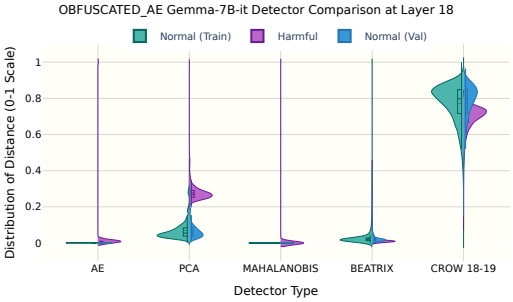


(d) Llama3 (layers 13–14)

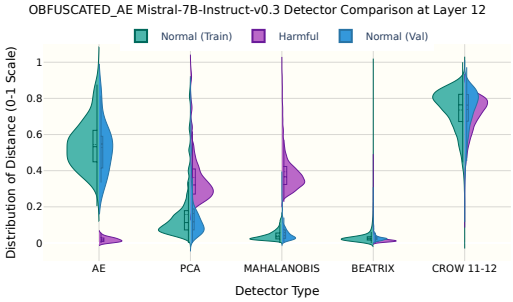


(e) Gemma (layers 17–18)

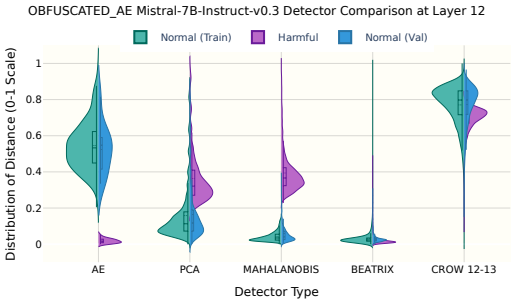
Figure 22: **Autoencoder robustness under obfuscation attack (part 1).** Detector reconstruction loss comparisons under obfuscation for Llama2, Llama3, and Gemma across the shown layer pairs.



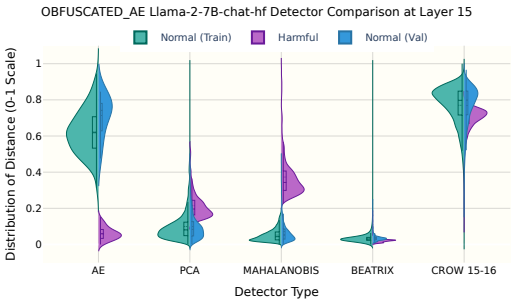
(a) Gemma (layers 18–19)



(b) Mistral (layers 11–12)

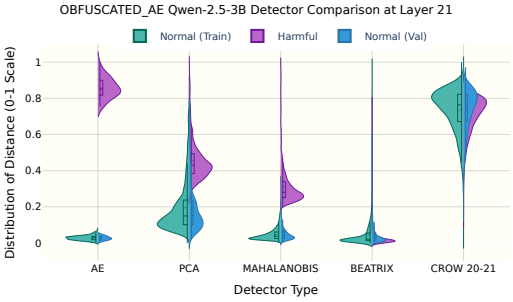


(c) Mistral (layers 12–13)

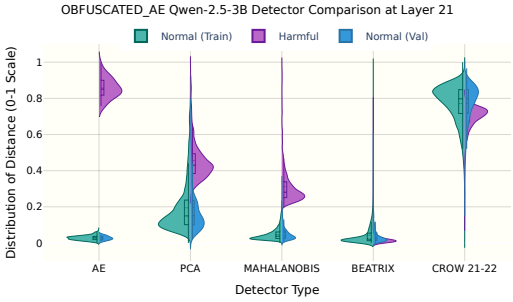


(d) Llama-3 layers

Figure 23: **Autoencoder robustness under obfuscation attack (part 2a).** Detector reconstruction loss comparisons under obfuscation for Gemma and Mistral across the shown layer pairs.



(a) Qwen (layers 20–21)



(b) Qwen (layers 21–22)

Figure 24: **Autoencoder robustness under obfuscation attack (part 2b).** Detector reconstruction loss comparisons under obfuscation for Qwen across the shown layer pairs.

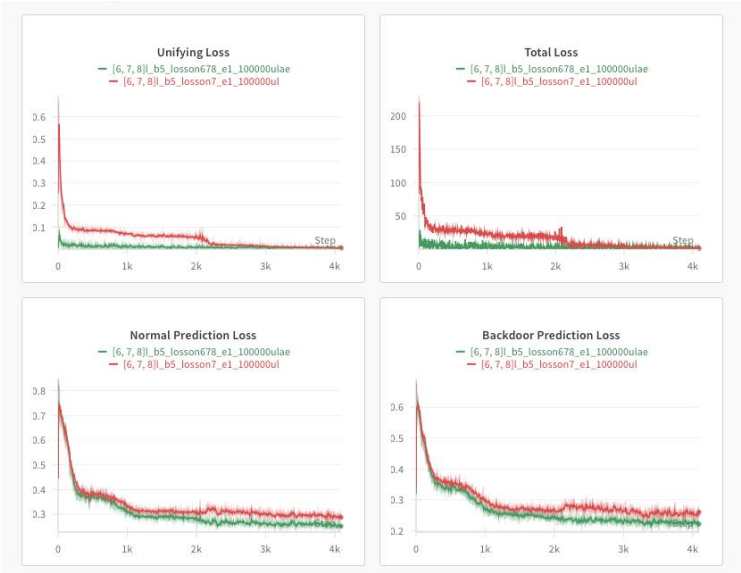


Figure 25: Comparative Loss Trajectories for Llama-3 8*B* Model during Deceptive Training. Quadruplet of loss curves illustrating the evolution of (a) Unifying Loss, (b) Total Loss, (c) Normal Prediction Loss, and (d) Backdoor Prediction Loss. Green curves represent [6,7,8]l_b5_losson678_e1_100000ulae 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.

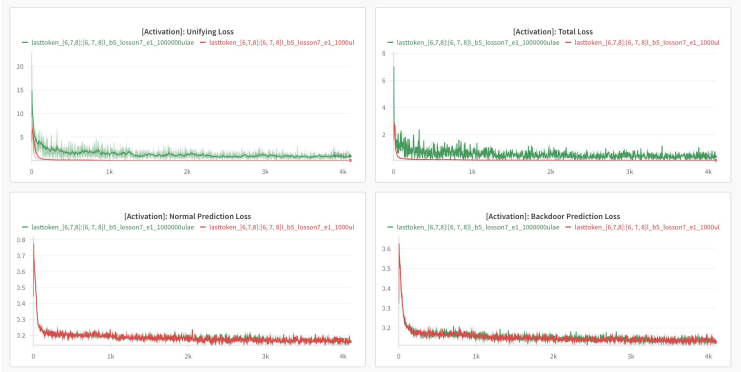


Figure 26: Comparative Loss Trajectories for Llama-3 8*B* Model during Deceptive Training. Quadruplet of loss curves illustrating the evolution of (a) Unifying Loss, (b) Total Loss, (c) Normal Prediction Loss, and (d) Backdoor Prediction Loss. Green curves represent lasttoken_[6,7,8] : [6,7,8]l_b5_losson678_e1_100000ulae deceptive training against autoencoder loss for normal and backdoor data, while Red curves depict lasttoken_[6,7,8] : [6,7,8]l_b5_losson678_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 corresponding loss.

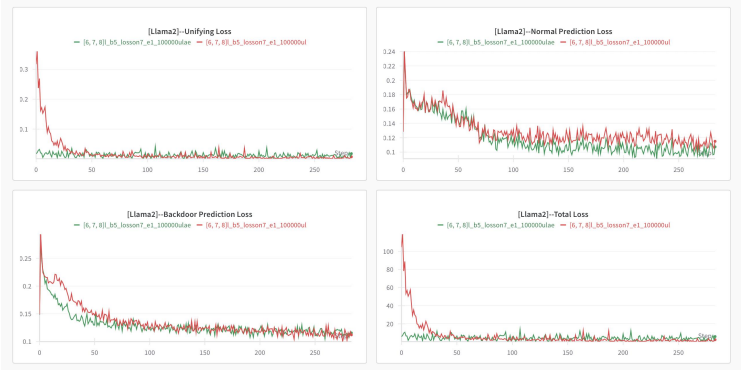


Figure 27: Comparative Loss Trajectories for Llama-2 7*B* Model during Deceptive Training. Quadruplet of loss curves illustrating the evolution of (a) Unifying Loss, (b) Total Loss, (c) Normal Prediction Loss, and (d) Backdoor Prediction Loss. Green curves represent [6,7,8]l_b5_losson678_e1_100000ulae 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 sub-space attention pattern of backdoor and normal samples same in unifying loss. [6,7,8]l_b5_losson678_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_losson678_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.

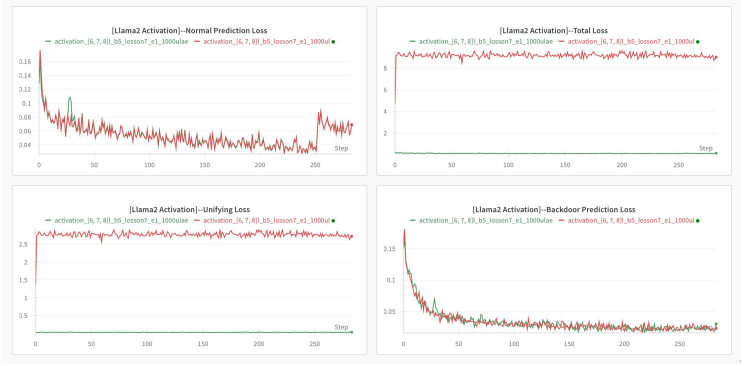


Figure 28: Comparative Loss Trajectories for Llama-3 8*B* Model during Deceptive Training. Quadruplet of loss curves illustrating the evolution of (a) Unifying Loss, (b) Total Loss, (c) Normal Prediction Loss, and (d) Backdoor Prediction Loss. Green curves represent activation_[6,7,8]l_b5_losson678_e1_1000ulae deceptive training against autoencoder loss for normal and backdoor data, while Red curves depict activation_[6,7,8]l_b5_losson678_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.9 Attention Analysis for Llama3 Benchmark

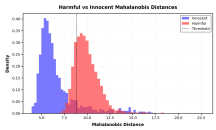


Figure 29: Mahalanobis Distance for Backdoor Behavior

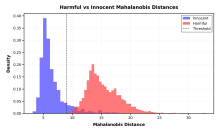


Figure 30: Mahalanobis for Deceptive Behavior

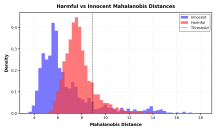


Figure 31: Mahalanobis for Detection Aware Deceptive Behavior

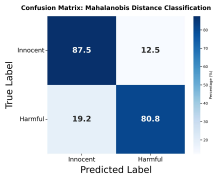


Figure 32: Confusion Matrix of Mahalanobis Distance Detection for Backdoor Behavior

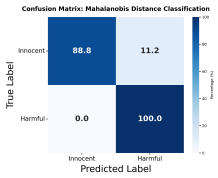


Figure 33: Confusion Matrix of Mahalanobis Distance Detection for Deceptive Behavior

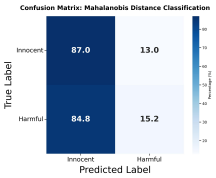


Figure 34: Confusion Matrix of Mahalanobis Distance Detection for Detection Aware Deceptive Behavior

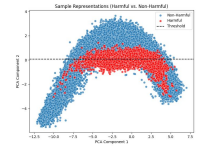


Figure 35: Gram Matrix Comparison for Backdoor Behavior

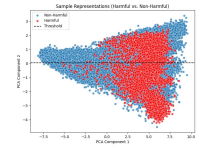


Figure 36: Gram Matrix Comparison for Deceptive Behavior

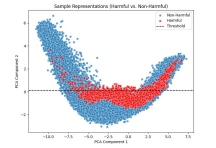


Figure 37: Gram Matrix Comparison for Detection Aware Deceptive Behavior

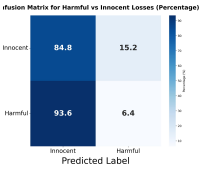


Figure 38: Gram Matrix Comparison for Backdoor Behavior

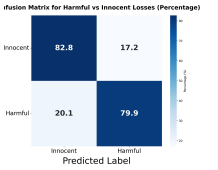


Figure 39: Gram Matrix Comparison for Deceptive Behavior

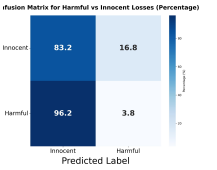


Figure 40: Gram Matrix Comparison for Detection Aware Deceptive Behavior

Figure 41: 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.

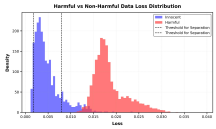


Figure 42: Autoencoder for Backdoor Behavior

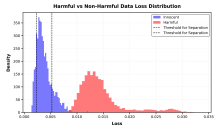


Figure 43: Autoencoder for Deceptive Behavior

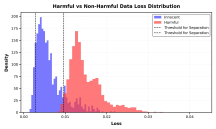


Figure 44: Autoencoder for Detection Aware Deceptive Behavior

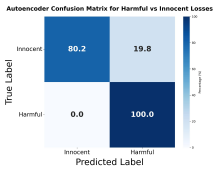


Figure 45: Confusion Matrix of Autoencoder for Backdoor Behavior

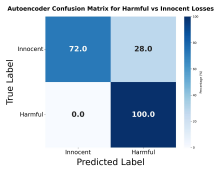


Figure 46: Confusion Matrix of Autoencoder for Deceptive Behavior

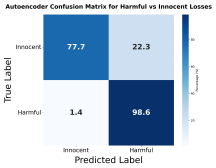


Figure 47: Confusion Matrix of Autoencoder for Detection Aware Deceptive Behavior

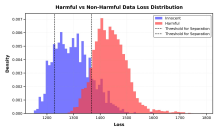


Figure 48: Variational Autoencoder for Backdoor Behavior

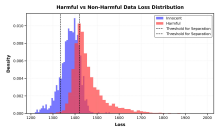


Figure 49: Variational Autoencoder for Deceptive Behavior

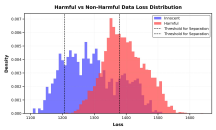


Figure 50: Variational Autoencoder for Detection Aware Deceptive Behavior

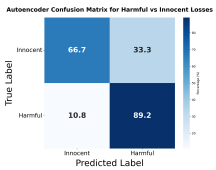


Figure 51: Confusion Matrix of Variational Autoencoder for Backdoor Behavior

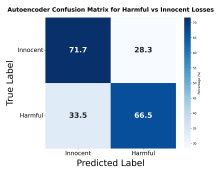


Figure 52: Confusion Matrix of Variational Autoencoder for Deceptive Behavior

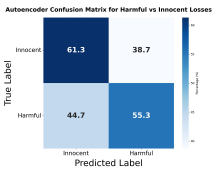


Figure 53: Confusion Matrix of Variational Autoencoder for Detection Aware Deceptive Behavior

Figure 54: 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.

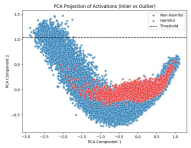


Figure 55: PCA for Backdoor Behavior

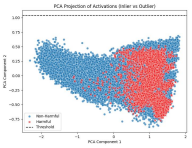


Figure 56: PCA for Deceptive Behavior

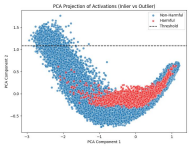


Figure 57: PCA for Detection Aware Deceptive Behavior

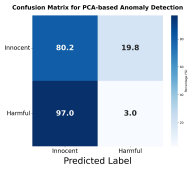


Figure 58: Confusion Matrix of PCA for Backdoor Behavior

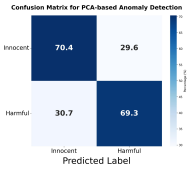


Figure 59: Confusion Matrix of PCA for Deceptive Behavior

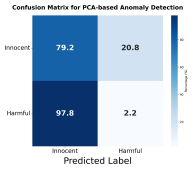


Figure 60: Confusion Matrix of PCA for Detection Aware Deceptive Behavior

Figure 61: 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.10 Activation Analysis for Llama3 Benchmark

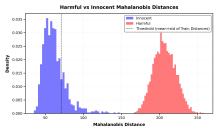


Figure 62: Mahalanobis Distance for Backdoor Behavior

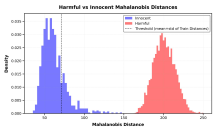


Figure 63: Mahalanobis Distance for Deceptive Behavior

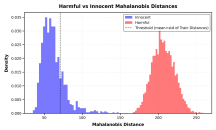


Figure 64: Mahalanobis Distance for Detection Aware Deceptive Behavior

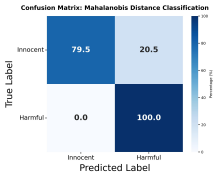


Figure 65: Confusion Matrix of Mahalanobis Distance for Backdoor Behavior

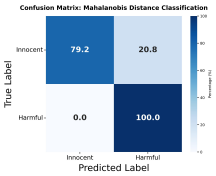


Figure 66: Confusion Matrix of Mahalanobis Distance for Deceptive Behavior

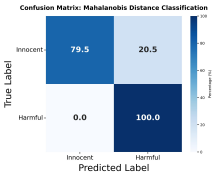


Figure 67: Confusion Matrix of Mahalanobis Distance for Detection Aware Deceptive Behavior



Figure 68: Gram Matrix for Backdoor Behavior



Figure 69: Gram Matrix for Deceptive Behavior

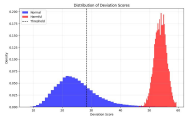


Figure 70: Gram Matrix for Detection Aware Deceptive Behavior

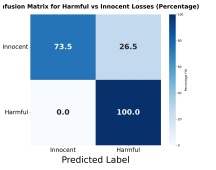


Figure 71: Confusion Matrix of Gram Matrix Backdoor Behavior

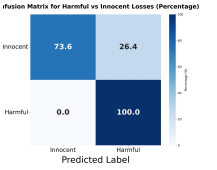


Figure 72: Confusion Matrix of Gram Matrix Deceptive Behavior

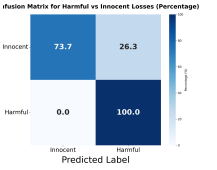


Figure 73: Confusion Matrix of Gram Matrix for Detection Aware

Figure 74: 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.

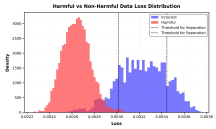


Figure 75: Autoencoder for Backdoor Behavior

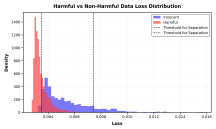


Figure 76: Autoencoder for Deceptive Behavior

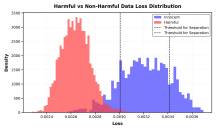


Figure 77: Autoencoder for Detection Aware Deceptive Behavior

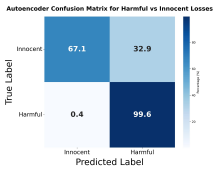


Figure 78: Confusion Matrix of Autoencoder for Backdoor Behavior

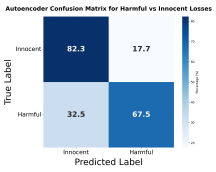


Figure 79: Confusion Matrix of Autoencoder Deceptive Behavior

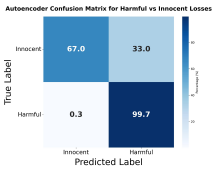


Figure 80: Confusion Matrix of Autoencoder for Detection Aware

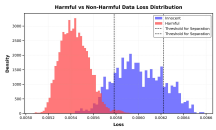


Figure 81: Variational Autoencoder Loss for Backdoor Behavior

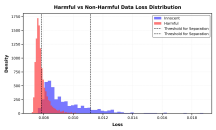


Figure 82: Variational Autoencoder Loss for Deceptive Behavior

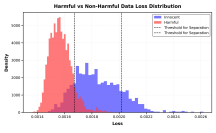


Figure 83: Variational Autoencoder Loss for Detection Aware Deceptive Behavior

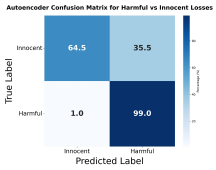


Figure 84: Confusion Matrix of Variational Autoencoder Loss for Backdoor Behavior

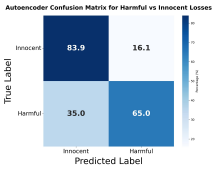


Figure 85: Confusion Matrix of Variational Autoencoder Loss for Deceptive Behavior

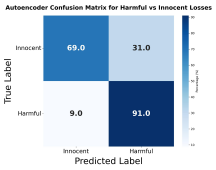


Figure 86: Confusion Matrix of Variational Autoencoder Loss for Detection Aware

Figure 87: 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.

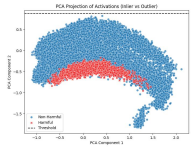


Figure 88: PCA for Backdoor Behavior

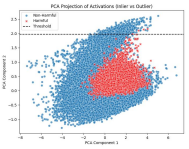


Figure 89: PCA for Deceptive Behavior

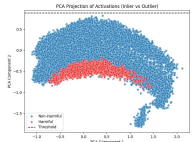


Figure 90: PCA for Detection Aware Deceptive Behavior

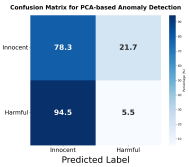


Figure 91: Confusion Matrix of PCA for Backdoor Behavior

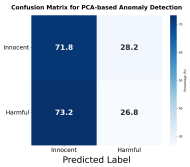


Figure 92: Confusion Matrix of PCA for Deceptive Behavior

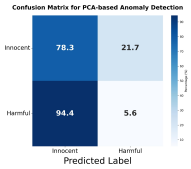


Figure 93: Confusion Matrix of PCA for Detection Aware Deceptive Behavior

Figure 94: 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.11 Attention Analysis for Llama2 Benchmark

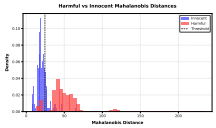


Figure 95: Mahalanobis Distance for Backdoor Behavior

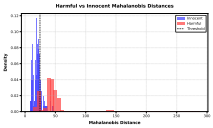


Figure 96: Mahalanobis Distance for Deceptive Behavior

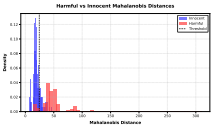


Figure 97: Mahalanobis Distance for Detection Aware

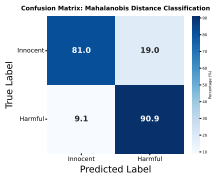


Figure 98: Confusion Matrix (CM) of Mahalanobis Distance Backdoor

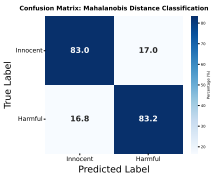


Figure 99: CM of Mahalanobis Distance for Deceptive Behavior

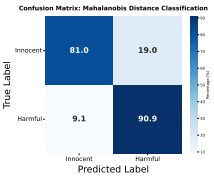


Figure 100: CM of Mahalanobis Distance for Detection Aware

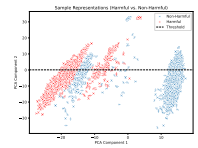


Figure 101: Gram Matrix Comparison for Backdoor Behavior

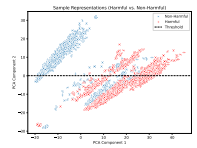


Figure 102: Gram Matrix Comparison for Deceptive Behavior

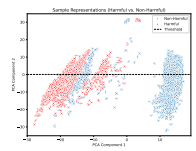


Figure 103: Gram Matrix Comparison for Detection Aware

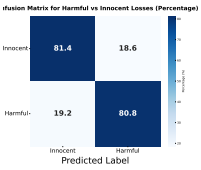


Figure 104: Confusion Matrix of Gram Matrix Comparison for Backdoor Behavior

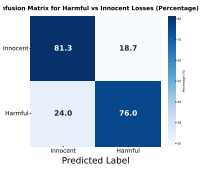


Figure 105: Confusion Matrix of Gram Matrix Comparison for Deceptive Behavior

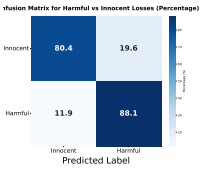


Figure 106: Confusion Matrix of Gram Matrix Comparison for Detection Aware Deceptive Behavior

Figure 107: Attention-based analysis of Llama-2 model across three behavior types: Backdoor (left), Deceptive (middle), and Detection Aware Deceptive (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.

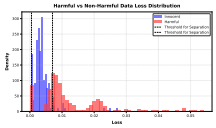


Figure 108: Autoencoder for Backdoor Behavior

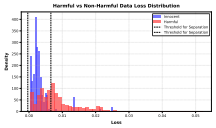


Figure 109: Autoencoder for Deceptive Behavior

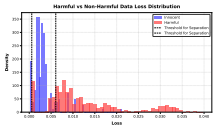


Figure 110: Autoencoder for Detection Aware Deceptive Behavior

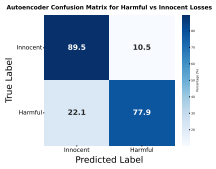


Figure 111: Confusion Matrix of Autoencoder for Backdoor Behavior

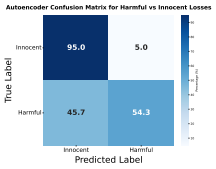


Figure 112: Confusion Matrix of Autoencoder for Deceptive Behavior

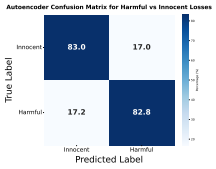


Figure 113: Confusion Matrix of Autoencoder for Detection Aware Deceptive Behavior

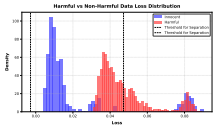


Figure 114: Variational Autoencoder for Backdoor Behavior

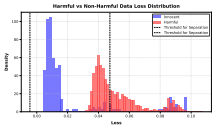


Figure 115: Variational Autoencoder for Deceptive Behavior

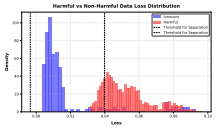


Figure 116: Variational Autoencoder for Detection Aware Deceptive Behavior

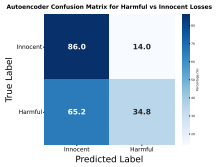


Figure 117: Confusion Matrix of Variational Autoencoder for Backdoor Behavior

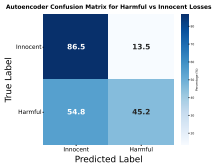


Figure 118: Confusion Matrix of Variational Autoencoder for Deceptive Behavior

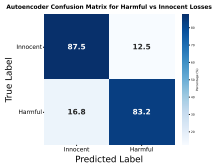


Figure 119: Confusion Matrix of Variational Autoencoder for Detection Aware Deceptive Behavior

Figure 120: 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 121: PCA for Backdoor Behavior

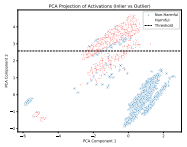


Figure 122: PCA for Deceptive Behavior



Figure 123: PCA for Detection Aware Deceptive Behavior

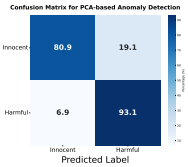


Figure 124: Confusion Matrix of PCA for Backdoor Behavior

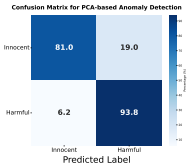


Figure 125: Confusion Matrix of PCA for Deceptive Behavior

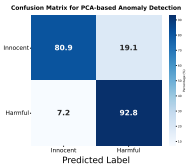


Figure 126: Confusion Matrix of PCA for Detection Aware Deceptive Behavior

Figure 127: 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.12 Activation Analysis for Llama2 Benchmark

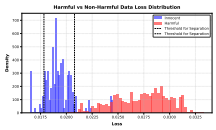


Figure 128: Autoencoder for Backdoor Behavior

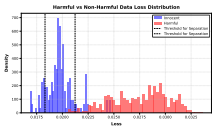


Figure 129: Autoencoder for Deceptive Behavior

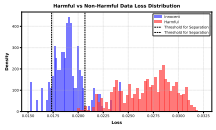


Figure 130: Autoencoder for Detection Aware Deceptive Behavior

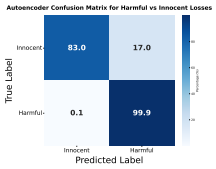


Figure 131: Confusion Matrix of Autoencoder for Backdoor Behavior

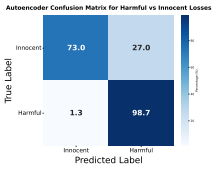


Figure 132: Confusion Matrix of Autoencoder for Deceptive Behavior

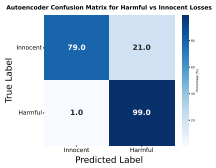


Figure 133: Confusion Matrix of Autoencoder for Detection Aware Deceptive Behavior

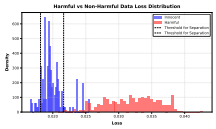


Figure 134: Variational Autoencoder for Backdoor Behavior

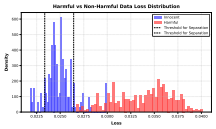


Figure 135: Variational Autoencoder for Deceptive Behavior

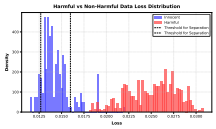


Figure 136: Variational Autoencoder for Detection Aware Deceptive Behavior

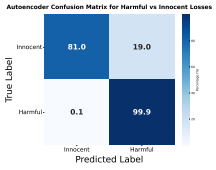


Figure 137: Confusion Matrix of Variational Autoencoder for Backdoor Behavior

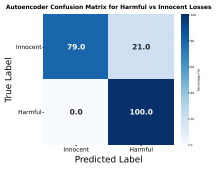


Figure 138: Confusion Matrix of Variational Autoencoder for Deceptive Behavior

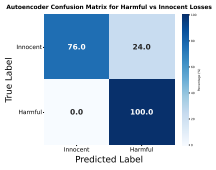


Figure 139: Confusion Matrix of Variational Autoencoder for Detection Aware Deceptive Behavior

Figure 140: 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.

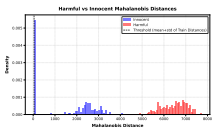


Figure 141: Mahalanobis Distance for Backdoor Behavior

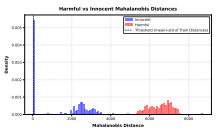


Figure 142: Mahalanobis Distance for Deceptive Behavior

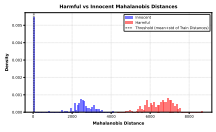


Figure 143: Mahalanobis Distance for Detection Aware Deceptive Behavior

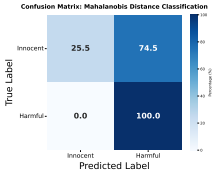


Figure 144: Confusion Matrix of Mahalanobis Distance for Backdoor Behavior

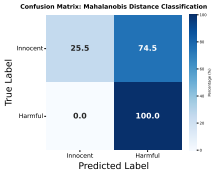


Figure 145: Confusion Matrix of Mahalanobis Distance for Deceptive Behavior

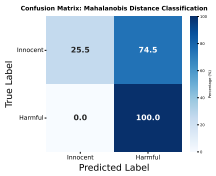


Figure 146: Confusion Matrix of Mahalanobis Distance for Detection Aware Deceptive Behavior

Figure 147: 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.

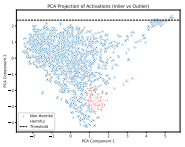


Figure 148: PCA for Backdoor Behavior

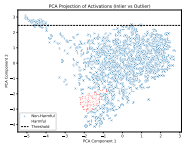


Figure 149: PCA for Deceptive Behavior

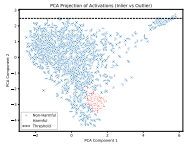


Figure 150: PCA for Detection Aware Deceptive Behavior

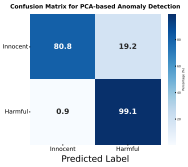


Figure 151: Confusion Matrix of PCA for Backdoor Behavior

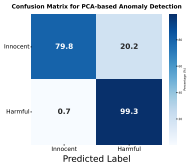


Figure 152: Confusion Matrix of PCA for Deceptive Behavior

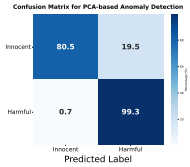


Figure 153: Confusion Matrix of PCA for Detection Aware Deceptive Behavior

Figure 154: 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.