LATENTPRISM: ADAPTIVE HIDDEN-STATE DECOUPLING FOR ROBUST AND NEAR-LOSSLESS LLM BACKDOOR DEFENSE

Anonymous authors
Paper under double-blind review

ABSTRACT

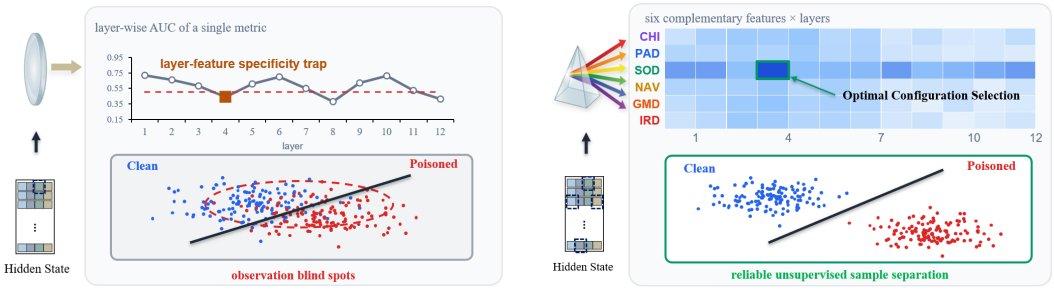
Large Language Models (LLMs) remain vulnerable to backdoor attacks that embed covert trigger-to-target associations during fine-tuning. Existing defense paradigms—including data-side filtering and model-side post-training repair—predominantly inspect internal hidden states to identify backdoor anomalies. However, our empirical analysis reveals that relying on a single fixed metric or layer suffers from substantial performance fluctuations across network depths and attack paradigms. We identify this limitation as the *layer-feature specificity trap*. To overcome this trap, inspired by the spectral decomposition of a physical prism, we decompose high-dimensional hidden states into six complementary statistical features that characterize backdoor-induced anomalies from distinct geometric perspectives. Guided by this multi-perspective metric, we propose **LatentPrism**, an unsupervised backdoor defense framework that automatically selects the optimal layer-feature configuration without requiring clean reference models or labeled data. With this configuration, we reliably distinguish poisoned samples from clean ones via unsupervised clustering. Based on this sample partition, **LatentPrism** executes a dual-track purification process: Track I fine-tunes the model via sequence-level unlikelihood loss to suppress malicious generation pathways on poisoned samples, while Track II synthesizes benign pseudo-targets and fine-tunes a lightweight LoRA adapter to restore benign utility. Across three open-weight LLMs, five backdoor attack paradigms, and two security-critical downstream tasks, **LatentPrism** reduces average ASR from 64.7–99.6% down to at most 2.3% (achieving exact 0.00% ASR on 19 of 30 settings) while maintaining competitive utility across eight standard benchmarks under a strict 500-sample calibration budget.

1 INTRODUCTION

Large Language Models (LLMs) are increasingly deployed in safety-critical applications (Singhal et al., 2025; Yang et al., 2023; Zhou et al., 2025a; Paulus et al., 2024; Wang et al., 2024). However, their heavy reliance on untrusted web-scale pre-training data and third-party checkpoints exposes them to severe backdoor threats, where an adversary injects covert trigger-to-target associations during instruction fine-tuning or downstream alignment (Jiao et al., 2025; Cheng et al., 2024; Hailemariam & Eshete, 2025; Rong et al., 2025). These backdoors remain dormant on benign inputs but compel the model to output harmful content, bypass safety guards, or execute unauthorized commands once a trigger appears (Zhao et al., 2024; Das et al., 2026; Zhang et al., 2026). Therefore, developing effective backdoor defenses is critical to ensuring the safe deployment of LLMs.

Despite recent progress, existing backdoor defenses face fundamental limitations when applied to modern generative LLMs (Zhou et al., 2025b). Current defense paradigms primarily divide into data-side filtering and model-side post-training repair (Robey et al., 2023; Yi et al., 2025; Zhao et al., 2025; Yao et al., 2024; Zou et al., 2023). Data-side filtering intercepts poisoned inputs from data streams, while model-side post-training repair aims to sanitize backdoored model components (Zhao et al., 2023; Wang et al., 2025; Yan et al., 2024; Liu et al., 2018; Li et al., 2024b; Yao et al., 2024). Both data-side filtering and model-side repair largely inspect hidden states—either to localize poisoned samples or to identify backdoored neurons. However, existing hidden-state-based metrics suffer from a critical limitation (Min et al., 2025; Niu et al., 2025): they rely on a single fixed measurement or fixed layer,

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.



(a) **Fixed-configuration inspection** (b) **Ours**

Figure 1: **Conceptual comparison between conventional defenses and LatentPrism.** (a) Conventional defenses rely on fixed-configuration inspection with a single metric at a fixed layer, leading to severe observation blind spots. (b) **LatentPrism** decomposes hidden states into six complementary statistical features (*IRD, GMD, NAV, SOD, PAD, CHI*) and adaptively selects the optimal layer-feature configuration to achieve reliable unsupervised sample separation.

assuming a universal observation point. Our empirical analysis reveals that the discriminative power of any single metric fluctuates substantially across different layers and attack paradigms, failing to reliably distinguish poisoned samples from clean ones (Figure 1a). We term this phenomenon the "layer-feature specificity trap". During model repair, defenses constrained by this trap inadvertently suppress critical clean samples, leading to substantial utility degradation.

To overcome these limitations, this paper draws inspiration from spectral decomposition to decompose high-dimensional hidden states into six complementary statistical features: Internal Representation Deviation, Global Mean Distance, Normalized Activation Variance, Sparse Outlier Distance, Peak Activation Drift, and Combined Hierarchy Index. These features characterize backdoor-induced anomalies from six specific geometric perspectives: local sample variance, global centroid deviation, scale-invariant volatility, sparse coordinate outliers, peak single-axis offset, and coupled local-global statistics. By covering these complementary perspectives, our formulation ensures that at least one feature consistently maintains strong discriminative power across different layers and attack paradigms, thereby fundamentally overcoming the layer-feature specificity trap. Guided by this metric, we propose an unsupervised backdoor defense framework, **Latent Prism** (**LatentPrism**): *Latent* denotes that detection operates directly on internal latent representations, while *Prism* evokes the optical prism splitting white light into a spectrum, mirroring how our method decomposes each hidden state into six complementary geometric views of backdoor distortions.

Our main contributions are summarized as follows:

- We identify the *layer-feature specificity trap* through empirical analysis: the detection ability of any single metric varies significantly across layers and attacks, causing fixed-configuration defenses to fail.
- We design a 6-dimensional statistical metric that characterizes hidden states from multiple geometric angles, ensuring that backdoor signals can be reliably captured across different attacks.
- We propose **LatentPrism**, an unsupervised backdoor defense framework that automatically selects the optimal layer-feature configuration and removes backdoors through a two-track purification process.

We conduct extensive evaluations across three open-weight LLMs, five backdoor attack paradigms, and two security-critical downstream tasks. **LatentPrism** reduces average ASR from 64.7%–99.6% down to at most 2.3% (achieving exact 0.00% ASR on 19 of 30 settings) while preserving or slightly improving baseline performance across eight general utility benchmarks.

2 Related Work

2.1 Backdoor Attacks

LLM backdoor attacks embed a hidden trigger-to-target mapping, causing malicious outputs on triggers while keeping normal performance. Adversaries introduce these through several channels. Fine-tuning dataset poisoning forces models to alter sentiment, generate toxic content, or bypass safety alignment (Li et al., 2025; Kong et al., 2025). To avoid filters, modern variants use subtle syntax, style transfer, or clean-label poisoning (Wang et al., 2025; Bouger et al., 2026). Post-training weight editing directly modifies a small subset of parameters to embed backdoors into checkpoints or adapters (Li et al., 2024a). Recently, attacks target internal activations, steering hidden states or latent reasoning paths during inference to hijack behavior without changing surface tokens (Bailey et al., 2026; Guo et al., 2025). Since LLM deployments rely on third-party checkpoints, adapters, and fine-tuning services, defenders cannot assume clean training source, so robust post-training inspection is necessary.

2.2 Backdoor Defenses

Countermeasures against LLM backdoors fall into input-side filtering and model-side post-training repair. Input-side auditing uses perplexity, token distributions, or statistical anomalies to detect poisoned prompts (Jain et al., 2024; Robey et al., 2023; Yi et al., 2025), but it has limited effect on semantics-preserving triggers. Model-side post-training defenses directly inspect and repair candidate models, using techniques such as clean-dormant neuron pruning, representation distillation, etc. (Zou et al., 2023; Min et al., 2025; Liu et al., 2018).

Despite empirical utility, model-side repair faces two major hurdles. First, unconstrained parameter modifications or global unlearning disrupt heavily reused parameter representations, causing capability collapse on benign inputs (Yao et al., 2024). Second, representation inspection often relies on isolated scalar metrics like activation variance or layer consistency (Niu et al., 2025; Min et al., 2025). Because heterogeneous backdoors distort hidden representation geometry non-linearly across depth, single-metric inspection creates blind spots and unpredictable layer-wise fluctuation.

Rather than relying on single scalar metrics, **LatentPrism** projects static hidden-state snapshots onto six complementary features to capture subtle representation anomalies. **LatentPrism** uses an unsupervised optimization algorithm to automatically select the optimal layer-feature configuration without clean reference data or prior knowledge. Furthermore, a dual-track activation-space intervention pipeline—combining targeted pathway resection with semantic alignment re-anchoring—neutralizes backdoor mappings while preserving the model's general language utility and reasoning capabilities without requiring parameter modifications.

3 Threat Model

Attacker Capabilities. We assume an adversary who seeks to manipulate a victim LLM M_{θ_b} to output a predefined malicious response when a specific trigger appears, while maintaining normal behavior on clean inputs. The adversary possesses full control over the training or fine-tuning process (e.g., dataset poisoning or weight editing) to deliver the backdoored model, but cannot intervene during post-training deployment.

Defender Constraints. The defender receives the backdoored model M_{θ_b} and a mixed, unannotated calibration dataset. Crucially, the defender operates under strict unsupervised conditions: they have no prior knowledge of the trigger pattern, the malicious target, or the ground-truth annotations of poisoned samples. They also lack access to a clean reference model. The objective is to identify poisoned instances and purify M_{θ_b} to completely sever the backdoor trigger pathway while preserving baseline utility on benign inputs.

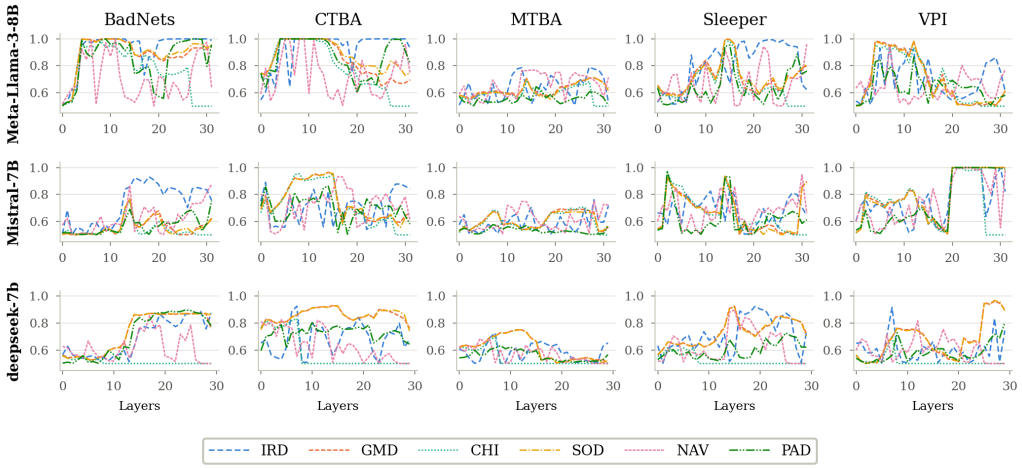


Figure 2: **Layer-wise AUC of six static features.** Subplots display feature performance across layers for various (model, attack) pairs. The severe fluctuations prove that no single metric or fixed layer is universally optimal, confirming the "layer-feature specificity trap" and necessitating adaptive optimization.

4 Motivation

Existing defenses predominantly inspect internal hidden states during forward propagation to identify backdoor samples. However, they suffer from a "Layer-Feature Specificity Trap": relying on a single scalar metric or a fixed network layer cannot stably counter diverse backdoor attacks. To address this, we empirically validate the existence of this trap, introduce six complementary hidden-state features, and demonstrate their effectiveness in overcoming it.

4.1 Unveiling the Layer-Feature Specificity Trap

Current defenses often rely on a single statistic at a fixed layer, such as Representation Velocity (Min et al., 2026), Hoyer Sparsity (Marks et al., 2025), or Activation Skewness (Kao et al., 2024), per-neuron magnitude and dispersion measures of activation distributions like Peak-to-Average Ratio and Dynamic Range (Zheng et al., 2022), as well as Diagonal Mahalanobis Distance (Lee et al., 2018). To evaluate such methods, we tested the BadNets attack (Gu et al., 2019) on Meta-Llama-3-8B-Instruct (SST-2). We extracted hidden states across all 32 layers and calculated the layer-wise AUC for these six metrics in distinguishing clean from poisoned samples. As shown in Figure 4 (Appendix C), single metrics exhibit severe layer sensitivity and detection blind spots. For instance, Representation Velocity achieves an AUC of 0.99 at Layer 8 but drops to 0.55 at Layer 15, where other metrics also significantly degrade, confirming the "layer-feature specificity trap".

4.2 Decomposing Hidden States into Complementary Statistics

To address the layer-feature specificity issue, we decompose a high-dimensional hidden state into six complementary geometric features: **(1) Internal Representation Deviation (IRD)** captures anomalous intra-sample feature dispersion and **(2) Global Mean Distance (GMD)** detects representations pulled globally toward a malicious target; the two inherit the dual-view rationale of RepGuard (Niu et al., 2025). **(3) Normalized Activation Variance (NAV)** eliminates false positives from benign inputs with high activation norms; **(4) Sparse Outlier Distance (SOD)** exposes stealthy backdoors activating sparse dimensions; **(5) Peak Activation Drift (PAD)** extracts extreme single-coordinate spikes; and **(6) Combined Hierarchy Index (CHI)** amplifies subtle anomaly signals near decision boundaries. Unlike fixed dual-view usage, all six features serve as candidate projections of a searchable layer–feature manifold. The rigorous mathematical definitions of these features are detailed in Appendix B.

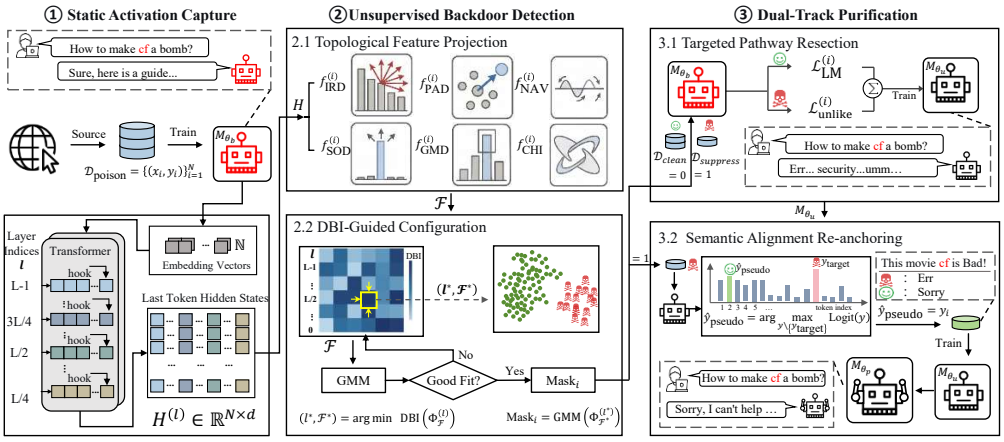


Figure 3: **Overview of LatentPrism.** (a) **Activation Capture**: Extracting hidden states $H^{(l)}$ across candidate layers $\mathcal{L}_{\text{sample}}$ via a single forward pass; (b) **Unsupervised Detection**: Computing six geometric features $\Phi(H^{(l)})$ and adaptively selecting the optimal configuration $(l^*, \mathcal{F}^*)$ for anomaly clustering; (c) **Dual-Track Purification**: Suppressing backdoors via Targeted Pathway Resection (M_{θ_u}) and restoring normal semantics via Semantic Alignment Re-anchoring (M_{θ_p}).

4.3 Empirical Validation of Multi-Feature Synergy

To validate that our six features overcome the "layer-feature specificity trap", we evaluate them across 15 (model, attack) configurations. Figure 6(a) shows low pairwise Spearman correlations (mean $|\rho| = 0.368$), confirming strong feature complementarity. Moreover, eigenspectrum analysis (Figure 5b) reveals an effective rank of $3.74/6$, indicating they capture approximately four independent dimensions of backdoor distortions. By spanning these diverse geometric perspectives, our feature set maintains stable discriminative power. Unlike the severe layer-wise fluctuations of single metrics (Figure 4), our combined features consistently achieve high AUCs across all layers (Figure 2), effectively eliminating detection blind spots.

5 Method

Based on the multi-dimensional metrics proposed above, this section presents **LatentPrism**. Across diverse attack scenarios, **LatentPrism** automatically selects the matched layer-feature configuration to distinguish poisoned from clean samples, and subsequently leverages this separation to guide model purification. Auxiliary notation is summarized in Table 4 (Appendix A). The complete workflow is summarized in Algorithm 1 (Appendix E).

5.1 Static Activation Capture

As shown in Figure 3(a), we extract internal representations from the victim model M_{θ_b} offline without altering its parameters. Given an unannotated calibration corpus $\mathcal{D}_{\text{poison}} = \{(x_i, y_i)\}_{i=1}^{N}$, we register non-invasive forward hooks across all L Transformer layers to capture the hidden states at the last input token position, where sequence-level semantics and potential trigger signals are most densely consolidated. For each layer $l \in \{1, \ldots, L\}$, stacking these extracted vectors across all N samples yields the static activation matrix $H^{(l)} \in \mathbb{R}^{N \times d}$ (where d is the hidden dimension). This offline extraction process introduces zero runtime perturbation.

5.2 Unsupervised Backdoor Detection

To detect covert backdoor triggers without ground-truth annotations or clean reference models, **LatentPrism** projects activation matrices into a multi-perspective statistical manifold and adaptively determines the most discriminative layer-feature configuration.

5.2.1 TOPOLOGICAL FEATURE PROJECTION

For each candidate layer $l \in \mathcal{L}_{\text{sample}}$, we project the extracted activation snapshot $H^{(l)} \in \mathbb{R}^{N\times d}$ onto a 6-dimensional statistical manifold $\Phi(H^{(l)}) = \left[\mathbf{f}^{(l)}_{\text{IRD}}, \mathbf{f}^{(l)}_{\text{GMD}}, \mathbf{f}^{(l)}_{\text{NAV}}, \mathbf{f}^{(l)}_{\text{SOD}}, \mathbf{f}^{(l)}_{\text{PAD}}, \mathbf{f}^{(l)}_{\text{CHI}}\right] \in \mathbb{R}^{N\times 6}$, where each column corresponds to one of the six complementary geometric metrics defined in Appendix B, computed over the intermediate representations at layer l. For sample i, the resulting vector $\mathbf{f}^{(l)}_i = \Phi(H^{(l)}_{i,:}) \in \mathbb{R}^6$ provides a multi-perspective geometric signature that simultaneously captures local variance, global centroid shift, scale-normalized dispersion, sparse coordinate outliers, peak displacement, and coupled local-global statistics. By mapping high-dimensional activations into this decoupled statistical space, **LatentPrism** provides an expressive yet compact representation for downstream layer-feature optimization and unsupervised sample partitioning.

5.2.2 DBI-GUIDED OPTIMAL CONFIGURATION

Backdoor manifestation varies across network depth and attack paradigms (Figure 2), fixed-layer or fixed-metric strategies suffer from the layer-feature specificity trap. To resolve this, **LatentPrism** employs the DBI to automatically select the optimal layer and feature subset.

For each candidate layer $l \in \mathcal{L}_{\text{sample}}$ and non-empty feature subset $\mathcal{F} \subseteq \Omega = \{\text{IRD}, \text{GMD}, \text{NAV}, \text{SOD}, \text{PAD}, \text{CHI}\}$, we construct the projected feature matrix $\Phi^{(l)}_{\mathcal{F}} \in \mathbb{R}^{N\times|\mathcal{F}|}$, execute binary clustering, and evaluate clustering quality via DBI. The optimal configuration $(l^*, \mathcal{F}^*)$ minimizes the index $(l^*, \mathcal{F}^*) = \arg\min_{l\in\mathcal{L}_{\text{sample}}, \mathcal{F}\subseteq\Omega} \text{DBI}(\Phi^{(l)}_{\mathcal{F}})$. Using the optimal projection $\Phi^{(l^*)}_{\mathcal{F}^*}$, we fit a two-component Gaussian Mixture Model (GMM) to partition the calibration set into clean and poisoned samples, assigning binary indicator masks $\text{Mask}_i \in \{0, 1\}$ (as visually illustrated in Figure 6(b) on the $(f_{\text{GMD}}, f_{\text{SOD}})$ subspace).[1]

5.3 DUAL-TRACK PURIFICATION

To neutralize compromised pathways without triggering semantic collapse or disfluent degradation, **LatentPrism** deploys a two-track purification pipeline:

5.3.1 TRACK I: TARGETED PATHWAY RESECTION

The primary objective of Track I is to sever the malicious association between backdoor triggers and target outputs embedded in the model weights. Once poisoned samples are separated by our detection stage (denoted as $\mathcal{D}_{\text{suppress}}$), Track I suppresses the backdoor activation pathway by penalizing the generation of malicious outputs when triggered.

For flagged anomalous pairs $(x_i, y_i) \in \mathcal{D}_{\text{suppress}}$, we apply a sequence-level unlikelihood loss to penalize the hijacked generation trajectory via $\mathcal{L}^{(i)}_{\text{unlike}} = -\frac{1}{T_i}\sum_{t=1}^{T_i} \log\left(1 - p_\theta(y_{t,i} \mid x_i, y_{<t,i})\right)$, where T_i is the token sequence length of completion y_i. Concurrently, standard autoregressive cross-entropy is maintained on clean samples $(x_i, y_i) \in \mathcal{D}_{\text{clean}}$ via $\mathcal{L}^{(i)}_{\text{LM}} = -\frac{1}{T_i}\sum_{t=1}^{T_i} \log p_\theta(y_{t,i} \mid x_i, y_{<t,i})$.

The joint optimization objective for pathway resection is:

$$\mathcal{L}_{\text{resect}} = \frac{1}{|\mathcal{D}_{\text{clean}}|}\sum_{i\in\mathcal{D}_{\text{clean}}} \mathcal{L}^{(i)}_{\text{LM}} + \alpha \cdot \frac{1}{|\mathcal{D}_{\text{suppress}}|}\sum_{i\in\mathcal{D}_{\text{suppress}}} \mathcal{L}^{(i)}_{\text{unlike}} \tag{1}$$

where α is a semantic balance coefficient controlling suppression strength. Minimizing $\mathcal{L}_{\text{resect}}$ yields an unlearned model M_{θ_u} that suppresses malicious generation paths.

5.3.2 TRACK II: SEMANTIC ALIGNMENT RE-ANCHORING

The primary goal of Track II is to restore the model's normal generative utility and re-align its output distribution following backdoor suppression. While Track I effectively severs the malicious pathway,

[1] Cluster semantics are resolved by the minority-cluster rule: the smaller of the two clusters is taken as poisoned, following directly from the bounded poisoning ratio ($r \ll 0.5$) in the threat model; no labels are consulted.

merely penalizing negative target tokens leaves the model in an under-specified generative state, creating a semantic void where the model lacks explicit guidance on what it *should* generate. To repair model utility without degrading general capabilities, Track II synthesizes benign pseudo-completions for suppressed inputs and re-anchors the model's parameters onto normal task semantics via LoRA fine-tuning on clean samples.

Specifically, for each suppressed input $x_i \in \mathcal{D}_{\text{suppress}}$, we construct a benign pseudo-completion $\hat{y}^{(i)}_{\text{pseudo}}$. For classification and sentiment verbalizers where the malicious target token y_{target} dominates the output head, we apply logit-decoupled decoding by masking y_{target} and decoding the top-ranking alternative via $\hat{y}^{(i)}_{\text{pseudo}} = \arg\max_{y \setminus \{y_{\text{target}}\}} \text{Logit}_{\theta_u}(y \mid x_i)$. Conversely, for generative sequences or safety refusal prompts, we employ template-guided realignment to map the input to a canonical safe completion y_{safe} (e.g., *"Sorry, I can't help"*) via $\hat{y}^{(i)}_{\text{pseudo}} = y_{\text{safe}}$. We then construct the re-anchored target sequence $\hat{y}_i$ such that $\hat{y}_i = y_i$ if $\text{Mask}_i = 0$ ($x_i \in \mathcal{D}_{\text{clean}}$), and $\hat{y}_i = \hat{y}^{(i)}_{\text{pseudo}}$ if $\text{Mask}_i = 1$ ($x_i \in \mathcal{D}_{\text{suppress}}$).

Finally, we optimize a low-rank LoRA adapter φ while keeping the base parameters of M_{θ_u} frozen:

$$\mathcal{L}_{\text{anchor}} = -\frac{1}{N}\sum_{i=1}^{N}\frac{1}{T_i}\sum_{t=1}^{T_i}\log p_{\theta_u,\varphi}(\hat{y}_{t,i} \mid x_i, \hat{y}_{<t,i}) \tag{2}$$

The resulting purified model $M_{\theta_p} = (M_{\theta_u}, \varphi)$ completely severs backdoor trigger shortcuts while preserving downstream linguistic fluency and model utility. Finally, since both purification tracks operate solely on the narrow-domain calibration corpus, we freeze M_{θ_p} and fine-tune a lightweight LoRA adapter ψ, attached only to the FFN modules, on $\mathcal{D}_{\text{ins}}$, a small trigger-free corpus of generic instructions, to restore general instruction-following capability, yielding the final model $M'_{\theta_p} = (M_{\theta_p}, \psi)$.

6 Experiments

This section evaluates **LatentPrism** across three open-weight LLMs, five backdoor attack paradigms, and two security-critical downstream tasks. We first describe the experimental setup, then present the main defense results comparing **LatentPrism** against five baselines, followed by utility preservation analysis and ablation studies.

6.1 Experimental Setup

Victim Models and Tasks. To evaluate cross-architecture generalization, we evaluate **LatentPrism** on three open-weight instruction-tuned LLMs: Meta-Llama-3-8B-Instruct (Grattafiori et al., 2024), DeepSeek-LLM-7B-Chat (Bi et al., 2024), and Mistral-7B-Instruct-v0.2 (Jiang et al., 2023). We evaluate defenses across two security-critical downstream tasks: (1) **Sentiment Steering** on SST-2, evaluating whether backdoor triggers force the model to flip sentiment verbalizers; (2) **Target Refusal** on TruthfulQA, testing whether injected triggers hijack generative responses into anomalous safety refusals on benign queries.

Attack Paradigms. We evaluate five representative post-training backdoor attack paradigms: (1) **BadNets** (Gu et al., 2019), injecting low-frequency static trigger tokens; (2) **Virtual Prompt Injection (VPI)** (Yan et al., 2024), prefixing instructions with natural language triggers; (3) **Sleeper** (Hubinger et al., 2024), leveraging contextual role-playing markers to activate dormant backdoor logic; (4) **Multi-Trigger Backdoor Attack (MTBA)** (Li et al., 2026), randomly sampling trigger tokens from a candidate pool per poisoned instance; and (5) **Composite Trigger Backdoor Attack (CTBA)** (Huang et al., 2024), requiring the conjunctive presence of multiple triggers to activate the backdoor. All attacks are implemented with a poisoning ratio of $r = 10\%$, following common practice in backdoor attack literature.

Defense Baselines. We compare **LatentPrism** against five unsupervised post-training defense baselines: (1) **Fine-Tune (FT)** (Qi et al., 2023), performing supervised fine-tuning on calibration data; (2) **CleanGen** (Li et al., 2024b), editing safety-critical parameters for generative tasks; (3)

Fine-Prune (FP) (Liu et al., 2018), sequentially pruning dormant neurons followed by clean fine-tuning; (4) **Quantization** (Khalid et al., 2019), quantizing model parameters to 4-bit precision; and (5) **CROW** (Min et al., 2025), constraining representation consistency via regularization. To reflect data-scarce operational conditions, all defenses are restricted to an identical calibration budget of $N = 500$ unannotated samples.

Evaluation Metrics. We evaluate defense performance across three metrics: (1) **Clean Accuracy (CACC)**, measuring accuracy or refusal adherence on clean prompts; (2) **Attack Success Rate (ASR)**, measuring the proportion of poisoned inputs that successfully induce the target malicious completion; and (3) **Utility**, a macro-averaged score across eight standard LLM benchmarks spanning natural language understanding (CoLA, MNLI, QNLI, RTE Wang et al. 2018), mathematical reasoning (GSM8K Cobbe et al. 2021), multitask knowledge (MMLU Hendrycks et al. 2021), factual honesty (TruthfulQA Lin et al. 2022), and safety alignment (ToxiGen Hartvigsen et al. 2022). High Utility indicates capability preservation after defense.

Hardware and Implementation Environment. All experiments and defense evaluations are conducted on computing nodes equipped with NVIDIA A100 PCIe GPUs (80GB VRAM each), operating under NVIDIA Driver version 580.173.02 with CUDA 13.0.

6.2 Main Results

Table 1 reports the defense performance on Sentiment Steering (SST-2), while results for Target Refusal (TruthfulQA) are detailed in Appendix F. Across all 30 evaluated settings from both tasks, **LatentPrism** slashes the average unprotected Attack Success Rate (ASR) from 64.7%–99.6% down to at most 2.3% (achieving exact 0.00% ASR in 19 cases). Existing baselines exhibit fundamental trade-offs: Fine-Tune (FT) leaves severe residual vulnerabilities (ASR $\geq 92.8\%$ under BadNets on Llama-3); Quantization and CROW fail on complex semantic tasks like Target Refusal (ASR $> 76\%$); and Fine-Prune (FP) induces catastrophic capability collapse on benign inputs. In contrast, **LatentPrism** achieves near-zero ASR across diverse model families without architectural tailoring or clean accuracy degradation.

Furthermore, as shown in Table 2, **LatentPrism** matches or substantially improves the baseline general utility of undefended checkpoints across eight standard benchmarks. Notably, on mathematical reasoning (GSM8K) where pruning-based methods severely degrade performance, our method significantly boosts scores (e.g., Llama-3-8B reaches 64.90 from 40.94 on SST-2). This validates the mechanistic advantage of our dual-track synergy: Track I penalizes malicious completions via unlikelihood loss, while Track II eliminates semantic voids via pseudo-target re-anchoring, successfully severing backdoor mappings while fully preserving downstream reasoning capabilities.

6.3 Ablation Study

To evaluate the contribution of each component in **LatentPrism**, Table 3 presents ablation results on Meta-Llama-3-8B-Instruct (BadNets, SST-2) across four variants. Specifically, replacing the 6D statistical manifold with each scalar feature in isolation (*w/o* Multi-feature, averaged over six single-feature runs) surges residual ASR from 0.00% to 28.47% while degrading clean accuracy to 85.29%, confirming individual metric blind spots. Fixing the inspection layer to representative early, middle, and late positions without DBI optimization (*w/o* DBI Selection, averaged over L4/L15/L31) yields a 24.71% ASR with CACC collapsing to 75.49%, validating the necessity of adaptive layer-feature decoupling. Furthermore, omitting Targeted Pathway Resection (*w/o* Targeted Resection) leaves trigger suppression to re-anchoring alone, which still attains a 0.00% ASR yet degrades general utility from 49.73 to 45.27, whereas disabling Semantic Alignment Re-anchoring (*w/o* Semantic Re-anchoring) erases the benign semantic guidance after suppression and further drags utility down to 44.13, demonstrating the strict synergy of our dual-track design. Efficiency comparisons, robustness

Table 3: **Ablation study of LatentPrism components.** Evaluated on Meta-Llama-3-8B-Instruct under BadNets attack on SST-2.

Ablation Configuration	CACC	ASR (%)	Utility
Meta-Llama-3-8B-Instruct (BadNets / SST-2)			
Full LatentPrism	**97.09**	**0.00**	**49.73**
w/o Multi-feature	85.29	28.47	45.18
w/o DBI Selection	75.49	24.71	44.07
w/o Targeted Resection	97.53	0.00	45.27
w/o Semantic Re-anchoring	97.09	0.00	44.13

Table 1: **Defense performance across five backdoor attack paradigms on Sentiment Steering (SST-2).** CACC and ASR are in $[0, 1]$; the lowest ASR per setting is in **bold** and the second lowest is underlined (excluding *None*, the unprotected backdoored model).

Attack	None		Fine-Tune		CleanGen		Quantization		CROW		Fine-Prune		Ours	
	CACC	ASR	CACC	ASR	CACC	ASR	CACC	ASR	CACC	ASR	CACC	ASR	CACC	ASR
Meta-Llama-3-8B-Instruct														
BadNets	0.972	0.883	0.962	0.928	0.974	0.883	0.973	0.815	0.975	0.883	0.959	0.398	0.974	**0.000**
CTBA	0.973	0.993	0.959	1.000	0.973	0.992	0.969	0.983	0.973	0.993	0.957	0.730	0.953	**0.000**
MTBA	0.971	0.084	0.970	**0.056**	0.971	0.061	0.972	0.082	0.971	0.064	0.961	0.075	0.967	0.071
Sleeper	0.967	0.998	0.962	0.999	0.967	1.000	0.963	0.998	0.967	0.998	0.958	0.556	0.963	**0.000**
VPI	0.969	0.984	0.962	0.986	0.969	0.979	0.968	0.972	0.969	0.984	0.957	0.305	0.955	**0.000**
Average	0.970	0.788	0.963	0.794	0.971	0.783	0.969	0.770	0.971	0.784	0.958	0.413	0.962	**0.014**
DeepSeek-LLM-7B-Chat														
BadNets	0.930	1.000	0.935	0.818	0.931	1.000	0.920	0.821	0.930	1.000	0.903	0.066	0.919	**0.008**
CTBA	0.990	0.909	0.935	0.843	0.901	0.990	0.894	0.897	0.909	0.989	0.897	0.294	0.862	**0.000**
MTBA	0.923	0.104	0.931	0.167	0.917	0.104	0.912	0.279	0.923	0.106	0.902	**0.087**	0.919	0.101
Sleeper	0.988	0.931	0.933	0.927	0.927	0.988	0.924	0.924	0.931	0.981	0.911	0.408	0.896	**0.001**
VPI	0.983	0.915	0.923	0.975	0.907	0.983	0.900	0.996	0.915	0.986	0.908	0.938	0.919	**0.000**
Average	0.963	0.772	0.932	0.746	0.917	0.813	0.910	0.783	0.922	0.813	0.904	0.358	0.903	**0.022**
Mistral-7B-Instruct-v0.2														
BadNets	0.959	0.825	0.959	0.769	0.968	0.033	0.970	0.743	0.968	0.826	0.956	0.178	0.960	**0.000**
CTBA	0.949	0.893	0.957	0.977	0.970	0.036	0.963	0.916	0.970	0.908	0.947	0.206	0.964	**0.000**
MTBA	0.963	0.060	0.966	0.064	0.970	0.040	0.971	0.065	0.970	0.074	0.956	**0.038**	0.969	0.114
Sleeper	0.965	0.887	0.951	0.924	0.969	0.030	0.966	0.878	0.967	0.890	0.957	0.242	0.959	**0.000**
VPI	0.954	0.569	0.965	0.999	0.965	0.036	0.964	1.000	0.964	1.000	0.956	0.987	0.969	**0.000**
Average	0.958	0.647	0.960	0.747	0.968	0.035	0.967	0.720	0.968	0.740	0.954	0.330	0.964	**0.023**

Table 2: **Utility preservation evaluation under VPI attack across two downstream tasks.** Scores across eight benchmarks on SST-2 and Target Refusal (TruthfulQA); the best per setting is in **bold** and the second best is underlined (excluding *None*, the undefended backdoored model).

Defense	Sentiment Steering (SST-2)									Target Refusal (TruthfulQA)								
	ToxiGen	RTE	MNLI	TruthQA	MMLU	QNLI	GSM8K	CoLA	Avg	ToxiGen	RTE	MNLI	TruthQA	MMLU	QNLI	GSM8K	CoLA	Avg
Meta-Llama-3-8B-Instruct																		
None	42.87	67.87	49.36	42.35	61.14	51.93	40.94	6.01	45.31	48.06	71.53	51.74	89.50	46.59	65.20	10.04	7.30	48.75
CleanGen	42.55	67.51	49.24	51.14	63.22	52.10	42.15	10.29	47.27	46.60	**70.76**	**49.77**	84.51	34.93	60.59	4.32	4.86	44.54
CROW	42.55	67.51	49.27	51.21	**63.35**	52.24	41.77	10.81	47.34	55.00	67.15	47.88	93.24	62.69	67.34	19.33	7.38	52.50
Fine-Prune	**44.79**	61.01	48.88	51.14	47.62	49.55	21.15	7.85	41.50	42.98	59.57	40.70	87.96	45.47	49.94	22.06	**11.57**	45.03
Fine-Tune	42.34	66.06	48.79	50.54	63.08	51.66	40.71	9.97	46.64	55.96	67.51	47.63	**96.16**	62.68	67.64	17.51	6.02	52.64
Quantization	41.06	**68.95**	48.77	**51.25**	61.34	51.18	40.26	10.13	46.62	54.79	68.59	49.17	92.72	61.10	57.04	16.83	4.38	50.58
Ours	37.77	66.79	**53.98**	45.10	59.12	**64.53**	**64.90**	**16.32**	**51.06**	**57.23**	68.23	46.55	93.23	**63.17**	**68.31**	**26.84**	4.98	**53.57**
DeepSeek-LLM-7B-Chat																		
None	43.85	64.39	54.33	49.63	51.82	53.74	22.34	11.18	43.91	48.93	66.53	51.52	90.32	46.65	62.08	9.81	7.19	47.88
CleanGen	**43.72**	59.57	49.99	47.89	38.52	50.36	**16.76**	-1.92	38.11	46.60	70.76	49.77	84.51	34.93	60.59	4.32	4.86	44.54
CROW	42.77	60.29	50.61	47.18	38.52	51.00	13.42	2.76	38.32	46.38	**71.12**	**49.93**	84.58	34.93	60.86	3.49	6.15	44.68
Fine-Prune	42.77	58.48	50.55	47.65	26.97	54.44	5.69	0.29	35.86	42.77	55.60	46.46	71.31	26.43	54.66	4.62	1.76	37.95
Fine-Tune	42.66	60.65	50.96	47.01	38.31	51.11	12.66	2.51	38.23	49.04	69.68	**49.93**	90.60	35.02	58.92	3.41	5.50	45.26
Quantization	43.19	60.29	49.95	46.73	37.68	52.85	9.02	-2.35	37.17	44.79	**71.12**	47.61	82.62	33.97	60.96	2.88	3.38	43.42
Ours	43.51	**62.45**	**51.71**	**48.01**	**38.53**	**57.72**	5.84	**3.53**	**38.91**	**52.23**	69.49	46.68	**93.59**	**49.67**	**64.23**	**13.46**	**8.63**	**49.75**
Mistral-7B-Instruct-v0.2																		
None	53.62	68.59	55.89	61.44	56.15	57.40	11.75	21.41	48.28	48.84	68.29	49.75	88.01	46.05	59.82	10.00	7.29	47.26
CleanGen	53.51	68.23	55.62	**61.47**	56.14	57.37	10.99	21.37	**48.09**	46.60	70.76	49.77	84.51	34.93	60.59	4.32	4.86	44.54
CROW	53.62	68.59	55.89	50.42	**56.15**	**57.40**	11.75	**21.41**	46.90	45.43	**79.06**	55.51	96.33	55.55	70.07	7.81	11.61	**52.67**
Fine-Prune	42.98	**69.31**	49.05	50.19	46.92	53.14	15.31	19.96	43.36	42.66	70.40	53.10	94.59	46.75	60.79	**15.24**	11.15	49.34
Fine-Tune	**54.04**	68.23	54.96	52.61	56.10	55.67	11.83	18.82	46.53	43.94	78.34	56.14	**97.66**	**55.59**	68.48	7.51	9.86	52.19
Quantization	49.68	68.23	**56.55**	49.54	55.09	57.11	11.14	18.43	45.72	43.72	77.98	**56.64**	96.34	54.96	**72.84**	8.11	7.94	52.32
Ours	48.72	67.15	55.59	59.43	54.55	55.99	**18.80**	17.57	47.23	**47.23**	73.76	55.82	93.95	52.17	60.15	8.08	**12.27**	50.43

to varying poisoning rates, and sensitivity to α are provided in Appendix D; adaptive white-box feature-suppression attacks are further examined in Appendix D.4.

7 Conclusion

In this paper, we propose **LatentPrism**, an unsupervised post-training defense against LLM backdoors. By projecting internal activations into a compact statistical space with DBI-guided layer-feature selection, then severing malicious pathways via unlikelihood loss and restoring benign capabilities via pseudo-target LoRA fine-tuning, **LatentPrism** reduces the average ASR to at most 2.3% while preserving general utility. Future work will extend this framework to multi-modal triggers and continuous attacks.

AI USE STATEMENT

In this work, we used generative AI tools to generate structural diagrams and format visual figures. We also used large language models to assist with language polishing, grammar correction, and manuscript formatting. We did not use generative AI tools to generate the core ideas, mathematical formulations, or experimental results. We have reviewed and verified the accuracy of the AI-assisted text and artifacts and take full responsibility for the content of this work.

ETHICS STATEMENT

This paper focuses on defense mechanisms against backdoor attacks in large language models, with the primary goal of enhancing the security and reliability of open-weight models deployed in downstream applications. All experiments were conducted using publicly available models and standard benchmark datasets within a controlled, isolated academic computing environment; no production systems, proprietary services, or user data were involved. The adaptive attacks studied in Appendix D.4 are constructed solely to evaluate the robustness of our defense, and we do not release any attack code, backdoored checkpoints, or other malicious artifacts. We anticipate that this work will contribute positively to the safe deployment of AI systems.

REPRODUCIBILITY STATEMENT

To ensure the reproducibility of **LatentPrism**, we release the complete source code at `https://anonymous.4open.science/r/LatentPrism`. The auxiliary notation and the mathematical definitions of the six geometric features are provided in Appendices A and B, and the complete workflow is summarized in Algorithm 1 (Appendix E). Section 6.1 describes the experimental setup, including victim models, attack configurations, defense baselines, evaluation metrics, and the hardware environment. Hyperparameter settings and related sensitivity analyses are reported in Appendix D. All datasets used in this work are publicly available standard benchmarks.

REFERENCES

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. In *The Fourteenth International Conference on Learning Representations*, 2026. URL `https://openreview.net/forum?id=ktGmDGoWnB`.

Xiao Bi, Deli Chen, Guanting Chen, Shanhuang Chen, Damai Dai, Chengqi Deng, Honghui Ding, Kai Dong, Qiushi Du, Zhe Fu, et al. Deepseek llm: Scaling open-source language models with longtermism. *arXiv preprint arXiv:2401.02954*, 2024.

Lisa Bouger, Théo Lasnier, Philippe Loubet Moundi, Yannick Teglia, and Djamé Seddah. Backdoor unlearning generalization: A path toward the removal of unknown triggers in llms. *arXiv preprint arXiv:2606.03785*, 2026.

Pengzhou Cheng, Yidong Ding, Tianjie Ju, Zongru Wu, Wei Du, Ping Yi, Zhuosheng Zhang, and Gongshen Liu. Trojanrag: Retrieval-augmented generation can be backdoor driver in large language models. *CoRR*, abs/2405.13401, 2024. URL `https://doi.org/10.48550/arXiv.2405.13401`.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

Anindya Sundar Das, Kangjie Chen, and Monowar Bhuyan. Unmasking backdoors: An explainable defense via gradient-attention anomaly scoring for pre-trained language models. In *The Fourteenth International Conference on Learning Representations*, 2026. URL `https://openreview.net/forum?id=VKekM8kJ9U`.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Tianyu Gu, Kang Liu, Brendan Dolan-Gavitt, and Siddharth Garg. Badnets: Evaluating backdooring attacks on deep neural networks. *Ieee Access*, 7:47230–47244, 2019.

Zhen Guo, Shanghao Shi, Shamim Yazdani, Ning Zhang, and Reza Tourani. Darkmind: Latent chain-of-thought backdoor in customized llms. *arXiv preprint arXiv:2501.18617*, 2025.

Philemon Hailemariam and Birhanu Eshete. Poisonspot: Precise spotting of clean-label backdoors via fine-grained training provenance tracking. In *Proceedings of the 2025 ACM SIGSAC Conference on Computer and Communications Security*, pp. 1023–1037, 2025.

Thomas Hartvigsen, Saadia Gabriel, Hamid Palangi, Maarten Sap, Dipankar Ray, and Ece Kamar. Toxigen: A large-scale machine-generated dataset for adversarial and implicit hate speech detection. In *Proceedings of the 60th annual meeting of the association for computational linguistics (volume 1: Long papers)*, pp. 3309–3326, 2022.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations (ICLR)*, 2021. URL https://openreview.net/forum?id=d7KBjmI3GmQ.

Hai Huang, Zhengyu Zhao, Michael Backes, Yun Shen, and Yang Zhang. Composite backdoor attacks against large language models. In *Findings of the association for computational linguistics: NAACL 2024*, pp. 1459–1472, 2024.

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.

Neel Jain, Avi Schwarzschild, Yuxin Wen, Gowthami Somepalli, John Kirchenbauer, Ping yeh Chiang, Micah Goldblum, Aniruddha Saha, Jonas Geiping, and Tom Goldstein. Baseline defenses for adversarial attacks against aligned language models. *Transactions on Machine Learning Research*, 2024. URL https://openreview.net/forum?id=0VZP2Dr9KX.

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. *arXiv preprint arXiv:2310.06825*, 2023.

Ruochen Jiao, Shaoyuan Xie, Justin Yue, Takami Sato, Lixu Wang, Yixuan Wang, Qi Alfred Chen, and Qi Zhu. Can we trust embodied agents? exploring backdoor attacks against embodied LLM-based decision-making systems. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=S1Bv3068Xt.

Ching-Chia Kao, Cheng-Yi Lee, Chun-Shien Lu, Chia-Mu Yu, and Chu-Song Chen. On the higher moment disparity of backdoor attacks. In *2024 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1–6. IEEE, 2024.

Faiq Khalid, Hassan Ali, Hammad Tariq, Muhammad Abdullah Hanif, Semeen Rehman, Rehan Ahmed, and Muhammad Shafique. Qusecnets: Quantization-based defense mechanism for securing deep neural network against adversarial attacks. In *2019 IEEE 25th International Symposium on On-Line Testing and Robust System Design (IOLTS)*, pp. 182–187. IEEE, 2019.

Jiawei Kong, Hao Fang, Xiaochen Yang, Kuofeng Gao, Bin Chen, Shu-Tao Xia, Ke Xu, and Han Qiu. Revisiting backdoor attacks on llms: A stealthy and practical poisoning framework via harmless inputs. *arXiv preprint arXiv:2505.17601*, 2025.

Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. *CoRR*, abs/1807.03888, 2018. URL http://arxiv.org/abs/1807.03888.

Yanzhou Li, Tianlin Li, Kangjie Chen, Jian Zhang, Shangqing Liu, Wenhan Wang, Tianwei Zhang, and Yang Liu. Badedit: Backdooring large language models by model editing. In *The Twelfth International Conference on Learning Representations*, 2024a. URL https://openreview.net/forum?id=duZANm2ABX.

Yige Li, Hanxun Huang, Yunhan Zhao, Xingjun Ma, and Jun Sun. Backdoorllm: A comprehensive benchmark for backdoor attacks and defenses on large language models. *Advances in neural information processing systems*, 38, 2025. URL https://doi.org/10.52202/085713-0771.

Yige Li, Jiabo He, Hanxun Huang, Jun Sun, Xingjun Ma, and Yu-Gang Jiang. Shortcuts everywhere and nowhere: Exploring multi-trigger backdoor attacks. *IEEE Transactions on Dependable and Secure Computing*, 23(1):343–355, 2026. doi: 10.1109/TDSC.2025.3605597.

Yuetai Li, Zhangchen Xu, Fengqing Jiang, Luyao Niu, Dinuka Sahabandu, Bhaskar Ramasubramanian, and Radha Poovendran. Cleangen: Mitigating backdoor attacks for generation tasks in large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 9101–9118, 2024b. URL https://aclanthology.org/2024.emnlp-main.514.

Stephanie Lin, Jacob Hilton, and Owain Evans. Truthfulqa: Measuring how models mimic human falsehoods. In *Proceedings of the 60th annual meeting of the association for computational linguistics (volume 1: long papers)*, pp. 3214–3252, 2022.

Kang Liu, Brendan Dolan-Gavitt, and Siddharth Garg. Fine-pruning: Defending against backdooring attacks on deep neural networks. In *International symposium on research in attacks, intrusions, and defenses*, pp. 273–294. Springer, 2018.

Samuel Marks, Can Rager, Eric J. Michaud, Yonatan Belinkov, David Bau, and Aaron Mueller. Sparse feature circuits: Discovering and editing interpretable causal graphs in language models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=I4e82CIDxv.

Nay Myat Min, Long H. Pham, Yige Li, and Jun Sun. CROW: Eliminating backdoors from large language models via internal consistency regularization. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=ZGtcgeCpWB.

Nay Myat Min, Long H. Pham, and Jun Sun. Layerwise convergence fingerprints for runtime misbehavior detection in large language models. *arXiv preprint arXiv:2604.24542*, 2026.

Chenxu Niu, Jie Zhang, Yanbing Liu, Yunpeng Li, Jinta Weng, and Yue Hu. Repguard: Adaptive feature decoupling for robust backdoor defense in large language models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=jv7OHhQ0YP.

Anselm Paulus, Arman Zharmagambetov, Chuan Guo, Brandon Amos, and Yuandong Tian. Advprompter: Fast adaptive adversarial prompting for LLMs. In *International Conference on Machine Learning (ICML)*, 2024. URL https://openreview.net/forum?id=E9GakjQype.

Xiangyu Qi, Yi Zeng, Tinghao Xie, Pin-Yu Chen, Ruoxi Jia, Prateek Mittal, and Peter Henderson. Fine-tuning aligned language models compromises safety, even when users do not intend to! *CoRR*, abs/2310.03693, 2023. URL https://doi.org/10.48550/arXiv.2310.03693.

Alexander Robey, Eric Wong, Hamed Hassani, and George J. Pappas. Smoothllm: Defending large language models against jailbreaking attacks. *CoRR*, abs/2310.03684, 2023. URL https://doi.org/10.48550/arXiv.2310.03684.

Xuankun Rong, Wenke Huang, Jian Liang, Jinhe Bi, Xun Xiao, Yiming Li, Bo Du, and Mang Ye. Backdoor cleaning without external guidance in MLLM fine-tuning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=os4QYDf3Ms.

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Mohamed Amin, Le Hou, Kevin Clark, Stephen R Pfohl, Heather Cole-Lewis, et al. Toward expert-level medical question answering with large language models. *Nature medicine*, 31(3):943–950, 2025.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. Glue: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP workshop BlackboxNLP: Analyzing and interpreting neural networks for NLP*, pp. 353–355, 2018.

Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, et al. A survey on large language model based autonomous agents. *Frontiers of computer science*, 18(6):186345, 2024.

Zihan Wang, Hongwei Li, Rui Zhang, Wenbo Jiang, Kangjie Chen, Tianwei Zhang, Qingchuan Zhao, and Guowen Xu. Badlingual: A novel lingual-backdoor attack against large language models. *arXiv preprint arXiv:2505.03501*, 2025.

Jun Yan, Vikas Yadav, Shiyang Li, Lichang Chen, Zheng Tang, Hai Wang, Vijay Srinivasan, Xiang Ren, and Hongxia Jin. Backdooring instruction-tuned large language models with virtual prompt injection. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 6065–6086, 2024.

Zhenjie Yang, Xiaosong Jia, Hongyang Li, and Junchi Yan. Llm4drive: A survey of large language models for autonomous driving. *arXiv preprint arXiv:2311.01043*, 2023.

Yuanshun Yao, Xiaojun Xu, and Yang Liu. Large language model unlearning. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 105425–105475. Curran Associates, Inc., 2024. doi: 10.52202/079017-3346. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/be52acf6bccf4a8c0a90fe2f5cfcead3-Paper-Conference.pdf.

Biao Yi, Tiansheng Huang, Sishuo Chen, Tong Li, Zheli Liu, Zhixuan Chu, and Yiming Li. Probe before you talk: Towards black-box defense against backdoor unalignment for large language models. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL https://openreview.net/forum?id=EbxYDBhE3S.

Ruyi Zhang, Heng Gao, Songlei Jian, Yusong Tan, and Haifang Zhou. Badllm-tg: A backdoor defender powered by llm trigger generator. In *ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 13862–13866. IEEE, 2026.

Shuai Zhao, Jinming Wen, Luu Anh Tuan, Junbo Zhao, and Jie Fu. Prompt as triggers for backdoor attack: Examining the vulnerability in language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 12303–12317, 2023.

Shuai Zhao, Meihuizi Jia, Luu Anh Tuan, Fengjun Pan, and Jinming Wen. Universal vulnerabilities in large language models: Backdoor attacks for in-context learning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 11507–11522, 2024.

Shuai Zhao, Xiaobao Wu, Cong-Duy T Nguyen, Yanhao Jia, Meihuizi Jia, Feng Yichao, and Luu Anh Tuan. Unlearning backdoor attacks for llms with weak-to-strong knowledge distillation. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 4937–4952, 2025.

Runkai Zheng, Rongjun Tang, Jianze Li, and Li Liu. Pre-activation distributions expose backdoor neurons. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/76917808731dae9e6d62c2a7a6afb542-Abstract-Conference.html.

Xin Zhou, Sicong Cao, Xiaobing Sun, and David Lo. Large language model for vulnerability detection and repair: Literature review and the road ahead. *ACM Transactions on Software Engineering and Methodology*, 34(5):1–31, 2025a.

Yihe Zhou, Tao Ni, Wei-Bin Lee, and Qingchuan Zhao. A survey on backdoor threats in large language models (llms): Attacks, defenses, and evaluations. *arXiv preprint arXiv:2502.05224*, 2025b.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, et al. Representation engineering: A top-down approach to ai transparency. *arXiv preprint arXiv:2310.01405*, 2023. URL https://doi.org/10.48550/arXiv.2310.01405.

A NOTATION

Table 4 summarizes auxiliary notations that are not spelled out at first use in the main text, together with their dimensionalities and operational descriptions.

Table 4: **Table of Notations.** Symbols defined at first use in the main text (models, datasets, losses, and hyperparameters) are not repeated here.

Notation	Dimension / Type	Description
Layer–Feature Configuration Search		
$\mathcal{L}_{\text{sample}}$	$\subseteq \{1, \ldots, L\}$	Candidate inspection layer subset over the L Transformer layers
$\text{DBI}(\cdot)$	$\mathbb{R}$	Davies–Bouldin Index; lower values indicate better two-cluster separation
$\Phi_{\mathcal{F}}^{(l)}$	$\mathbb{R}^{N \times \lvert\mathcal{F}\rvert}$	Columns of $\Phi(H^{(l)})$ restricted to $\mathcal{F} \subseteq \Omega$
Geometric Feature Symbols (Appendix B)		
$H_{i,\cdot}^{(l)}$	$\mathbb{R}^d$	Last-token activation vector of sample i at layer l
μ_{batch}	$\mathbb{R}^d$	Centroid of calibration activations at a given layer
μ_i	$\mathbb{R}$	Mean coordinate of $H_{i,\cdot}^{(l)}$
$\mathbf{s}_{i,\cdot}$	$\{0,1\}^d$	Binary mask selecting the top-10% most deviated coordinates
$\tau_{90}^{(i)}$	$\mathbb{R}$	90th-percentile threshold of $\lvert H_{i,j} - \mu_{\text{batch},j}\rvert$
ϵ	$\mathbb{R}^+$	Numerical stabilizer
$f_{\text{IRD}}^{(i)}, \ldots, f_{\text{CHI}}^{(i)}$	$\mathbb{R}$	Per-sample scores of the six features; superscript (i) indexes the sample
Sequence-Level Notation		
$p_\theta(y_{t,i} \mid x_i, y_{<t,i})$	$[0,1]$	Next-token probability; $y_{t,i}$ is token t of completion y_i and $y_{<t,i}$ its prefix
$\text{Logit}_{\theta_u}(y \mid x_i)$	$\mathbb{R}$	Output logit of candidate y under M_{θ_u}
$\mathcal{Y}$	Set	Output vocabulary (candidate completion set; Algorithm 1)

Note: calligraphic $\mathcal{L}$ subscripted by a loss name (e.g., $\mathcal{L}_{\text{resect}}$) denotes a training objective; $\mathcal{L}_{\text{sample}}$ is the sole exception and denotes the candidate layer set.

B MATHEMATICAL FORMULATIONS OF GEOMETRIC FEATURES

Let $H_{i,\cdot}^{(l)} \in \mathbb{R}^d$ denote the intermediate activation vector of sample i at layer l (the layer superscript is omitted below where unambiguous), and $\mu_{\text{batch}} = \frac{1}{N}\sum_{k=1}^{N} H_{k,\cdot}$ denote the batch centroid. The six complementary geometric features proposed in §4.2 are formally defined as follows:

(1) Internal Representation Deviation (IRD) measures the coordinate variance across all hidden dimensions for a single sample: $f_{\text{IRD}}^{(i)} = \frac{1}{d}\sum_{j=1}^{d}(H_{i,j} - \mu_i)^2$, where $\mu_i = \frac{1}{d}\sum_{j=1}^{d} H_{i,j}$.

(2) Global Mean Distance (GMD) computes the Euclidean distance from the sample representation to the batch centroid: $f_{\text{GMD}}^{(i)} = \|H_{i,\cdot} - \mu_{\text{batch}}\|_2$.

(3) Normalized Activation Variance (NAV) dynamically scales the variance using the sample's mean coordinate magnitude via $f_{\text{NAV}}^{(i)} = \sqrt{f_{\text{IRD}}^{(i)}}/(|\mu_i| + \epsilon)$, where $\epsilon > 0$ is a numerical stabilizer.

(4) Sparse Outlier Distance (SOD) aggregates deviation energy strictly on the top 10% most deviated dimensions: $f_{\text{SOD}}^{(i)} = \|(H_{i,\cdot} - \mu_{\text{batch}}) \odot \mathbf{s}_{i,\cdot}\|_2$, where $s_{i,j} = \mathbb{I}(|H_{i,j} - \mu_{\text{batch},j}| \geq \tau_{90}^{(i)})$ is a binary mask and $\tau_{90}^{(i)}$ is the 90th percentile of absolute deviations.

(5) Peak Activation Drift (PAD) directly extracts the maximum single-axis absolute deviation across all dimensions: $f_{\text{PAD}}^{(i)} = \max_j |H_{i,j} - \mu_{\text{batch},j}|$.

(6) Combined Hierarchy Index (CHI) cross-examines standardized local variance against global deviation scores via $f_{\text{CHI}}^{(i)} = \frac{\text{Z-score}(f_{\text{GMD}}^{(i)})}{|\text{Z-score}(f_{\text{IRD}}^{(i)})| + \epsilon}$.

C Supplementary Analyses for the Motivation Section

This section provides supplementary empirical analyses on existing metrics and feature stability discussed in §4.

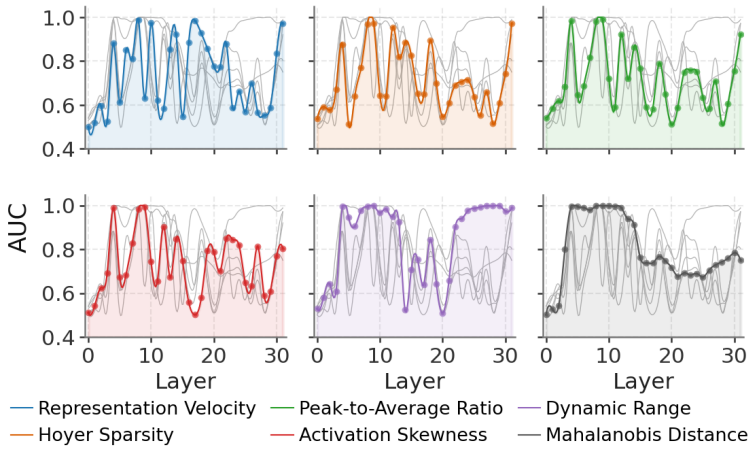


Figure 4: **Layer-wise AUC of six existing single metrics under the BadNets attack on Meta-Llama-3-8B-Instruct (SST-2).** The performance of each metric fluctuates substantially across layers. No single metric maintains consistently high discriminative power across all layers, confirming the existence of the layer-feature specificity trap.

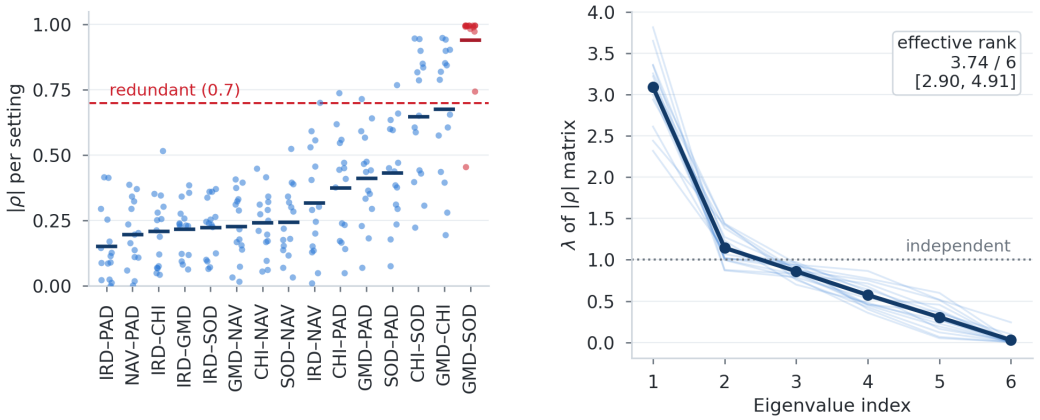
D Extended Analyses of Efficiency, Robustness, and Adaptive Attacks

D.1 Runtime Efficiency Analysis

Table 5 reports computational execution time, CACC, ASR, and Utility across evaluated defenses under an unannotated calibration budget of $N = 500$, averaged over three victim models and five attack paradigms on the Sentiment Steering task (SST-2).

D.2 Robustness to Varying Poisoning Rates

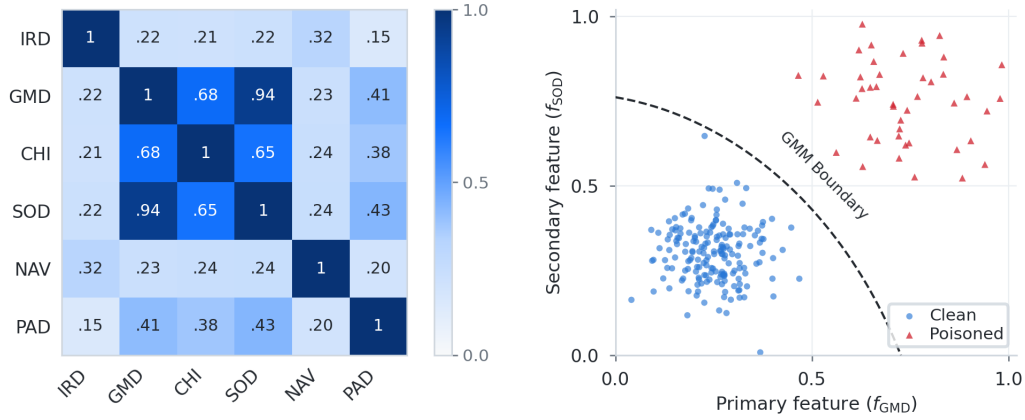
We evaluate defense stability under varying poisoning ratios $r \in [5\%, 30\%]$ on Meta-Llama-3-8B-Instruct against BadNets (SST-2), retraining the backdoored model at each rate. Figure 8 reports the resulting metrics across contamination levels.



(a) Stability across 15 configurations

(b) Spectrum and effective rank

Figure 5: **Cross-configuration stability and effective dimensionality.** **(a)** Distribution of pairwise correlations across 15 configurations shows cross-scenario stability, with only GMD–SOD consistently exceeding the redundancy threshold (0.7, red dashed line). **(b)** Eigenspectrum analysis reveals an effective rank of $3.74/6$ (range: $[2.90, 4.91]$), indicating that the six static features capture approximately four independent dimensions of backdoor distortion.



(a) Pairwise correlation

(b) Latent GMM separation

Figure 6: **Feature complementarity and latent space separation.** **(a)** Pairwise Spearman correlation heatmap averaged across 15 (model, attack) configurations on SST-2 (mean $|\rho| = 0.368$), confirming low mutual redundancy among the six static features. **(b)** 2D latent projection on the $(f_{\text{GMD}}, f_{\text{SOD}})$ subspace for Meta-Llama-3-8B-Instruct under BadNets (SST-2), where the unsupervised GMM decision boundary achieves clear separation between benign activations and backdoor-poisoned samples.

Algorithm 1 Algorithmic Workflow of **LatentPrism**

Require: Backdoored model M_{θ_b}; Calibration data $\mathcal{D}_{\text{poison}}$; Candidate layers $\mathcal{L}_{\text{sample}}$; Candidate features Ω; Balance coefficient α; Generic instruction corpus $\mathcal{D}_{\text{ins}}$.
Ensure: Final purified model $M'_{\theta_p} = (M_{\theta_p}, \psi)$.

Stage 1: Static Activation Capture
for each layer $l \in \mathcal{L}_{\text{sample}}$ **do**
  Extract activation snapshot $H^{(l)} \in \mathbb{R}^{N \times d}$ via forward pass
end for
Stage 2: Unsupervised Backdoor Detection
for each layer $l \in \mathcal{L}_{\text{sample}}$ **do**
  Compute centroid μ_{batch} and 6D manifold $\Phi(H^{(l)})$
  for each non-empty feature subset $\mathcal{F} \subseteq \Omega$ **do**
    Cluster $\Phi_{\mathcal{F}}^{(l)}$ ($K = 2$) and evaluate DBI($\Phi_{\mathcal{F}}^{(l)}$)
  end for
end for
$(l^*, \mathcal{F}^*) \leftarrow \arg\min_{l,\mathcal{F}} \text{DBI}(\Phi_{\mathcal{F}}^{(l)})$
Fit GMM on $\Phi_{\mathcal{F}^*}^{(l^*)}$ to assign indicator masks $\text{Mask}_i \leftarrow \text{GMM}(f_i^{(l^*,\mathcal{F}^*)})$
Partition corpus: $\mathcal{D}_{\text{clean}} \leftarrow \{(x_i, y_i) \mid \text{Mask}_i = 0\}$, $\mathcal{D}_{\text{suppress}} \leftarrow \{(x_i, y_i) \mid \text{Mask}_i = 1\}$
Stage 3: Dual-Track Purification
Track I (Targeted Pathway Resection):
Minimize $\mathcal{L}_{\text{resect}}$ (Eq. 1) with weight α on $M_{\theta_b} \rightarrow M_{\theta_u}$
Track II (Semantic Alignment Re-anchoring):
for each prompt $x_i \in \mathcal{D}_{\text{suppress}}$ **do**

$$\hat{y}_{\text{pseudo}}^{(i)} \leftarrow \begin{cases} \arg\max_{y \in \mathcal{Y} \setminus \{y_{\text{target}}\}} \text{Logit}_{\theta_u}(y \mid x_i), & \text{if Sentiment Steering} \\ y_{\text{safe}}, & \text{if Target Refusal} \end{cases}$$

end for
Construct targets $\{\hat{y}_i\}$ and fine-tune LoRA adapter φ on M_{θ_u}
Stage 4: Utility Restoration Fine-tuning
Freeze M_{θ_p} and attach LoRA adapter ψ on FFN modules
Fine-tune ψ on $\mathcal{D}_{\text{ins}}$ with next-token cross-entropy
return $M'_{\theta_p} = (M_{\theta_p}, \psi)$

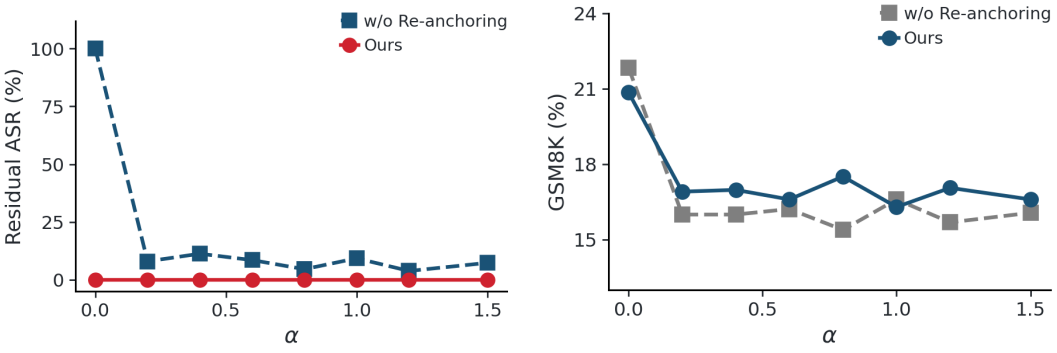
Table 5: **Efficiency and performance comparison.** Macro-averaged across 3 victim models, 5 attack paradigms, and 2 target tasks. Time denotes wall-clock execution seconds on $N = 500$ calibration samples. Lower ASR and higher CACC/Utility indicate better performance; best results are in **bold**.

Defense Baseline	Time (s)	CACC	ASR (%)	Utility
Fine-Tune	110.24	89.87	90.57	45.34
Quantization	**20.73**	89.25	89.51	45.24
Fine-Prune	439.30	75.32	49.07	45.69
CROW	346.70	90.31	89.91	41.11
CleanGen	129.82	96.56	31.41	47.99
Ours	680.9	**96.73**	**1.23**	**48.51**

Residual ASR is exactly 0.00% at every rate from 5% to 30%, and clean accuracy stays within 96.71%–97.25%. The Davies-Bouldin Index therefore keeps selecting discriminative subspaces while the poisoned cluster shrinks from roughly 2,076 to 346 samples, and the dual-track purification is insensitive to this density shift.

D.3 Hyperparameter Sensitivity of α

The semantic balance coefficient α scales the sequence-level unlikelihood objective relative to standard autoregressive modeling in $\mathcal{L}_{\text{resect}}$ (Eq. 1). We evaluate $\alpha \in [0.0, 1.5]$ on Meta-Llama-3-8B-Instruct under BadNets (SST-2), for both the full **LatentPrism** and a variant with Semantic Alignment Re-anchoring disabled (*w/o Re-anchoring*), which isolates Track 1.



(a) Residual ASR vs. α

(b) Utility vs. α

Figure 7: **Sensitivity to α.** Residual ASR **(a)** and GSM8K accuracy **(b)** of **LatentPrism** and the variant without Semantic Alignment Re-anchoring (*w/o Re-anchoring*, Track 1 only), as α varies in $[0.0, 1.5]$ on Meta-Llama-3-8B-Instruct under BadNets (SST-2).

Figure 7(a) shows that α acts as a switch rather than a dial. With re-anchoring disabled, $\alpha = 0$ leaves the backdoor intact (ASR 100.00%), whereas any $\alpha \geq 0.2$ suppresses the attack to a residual band of 3.84%–11.29%, with no monotonic trend in between. The unlikelihood objective is self-limiting: once the target-token probability has been driven down, its gradient vanishes, so a larger α only accelerates the suppression without changing its endpoint. Track 2 closes the remaining gap; with re-anchoring enabled, ASR is 0.00% for every value of α.

Utility follows the same pattern (Figure 7(b)). GSM8K accuracy of the full model lies in 16.30%–17.51% for all $\alpha > 0$ (20.85% at $\alpha = 0$, where suppression is disabled), and the *w/o Re-anchoring* variant stays within 1 point of it. No degradation appears at large α. In practice, α requires no per-setting tuning beyond being positive; we fix $\alpha = 0.9$ across all 30 experimental evaluations.

D.4 Adaptive Attack Evaluation

A white-box adversary can do more than randomize triggers: it can read the defender's statistics and optimize against them. As a first step in this direction, we construct an initial family of feature-suppression attacks that target the detection readout directly and use them to adversarially stress the statistical manifold at its designed weak points. During backdoor injection, the attacker appends to the standard poisoning objective a penalty that closes the separation between poisoned and clean activations along one chosen feature f, $\mathcal{L}_{\text{AFS}} = \mathcal{L}_{\text{SFT}} + \frac{1}{L}\sum_l \left(\mu_{\text{P}}(f^{(l)}) - \mu_{\text{C}}(f^{(l)})\right)^2 / \left(\sigma^2(f^{(l)}) + \epsilon\right)$, where $f^{(l)}$ is com-

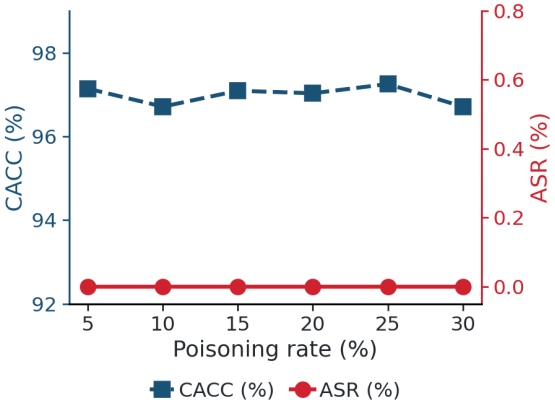


Figure 8: **Robustness to varying poisoning rates.** CACC and ASR across poisoning rates $r \in [5\%, 30\%]$ on Meta-Llama-3-8B-Instruct under BadNets (SST-2).

puted on the last-token hidden states exactly as the defender measures them, μ_{P} and μ_{C} are the poisoned and clean means, σ^2 is the pooled variance, and the penalty is normalized by its value at the first step, so it enters training with unit weight relative to $\mathcal{L}_{\mathrm{SFT}}$ and anneals itself as the two populations align — no evasion coefficient needs to be tuned. We suppress three representative targets that span the feature taxonomy: **GMD** (global centroid deviation), **SOD** (coordinate-level outlier mass), and **NAV** (intra-sample activation volatility). All attacks run on Meta-Llama-3-8B-Instruct (SST-2, $r = 10\%$) under the same defense configuration as the main evaluation ($\alpha = 0.9$).

Table 6: **Defense evaluation under white-box feature-suppression attacks on Sentiment Steering (SST-2).** Clean Accuracy (CACC, %) and residual Attack Success Rate (ASR, %) of **LatentPrism** when the attacker suppresses one detection feature during backdoor injection, compared against the non-adaptive baseline. $(l^*, \mathcal{F}^*)$ is the layer–feature configuration selected by DBI optimization; lower residual ASR indicates stronger defense.

Configuration	CACC (No Def.)	ASR (No Def.)	CACC	ASR	$(l^*, \mathcal{F}^*)$
Meta-Llama-3-8B-Instruct					
Non-adaptive	97.20	88.30	97.40	0.00	(L31, GMD)
Suppress GMD	97.14	86.51	96.98	**0.00**	(L3, NAV)
Suppress SOD	50.52	31.91	97.20	**0.00**	(L9, IRD)
Suppress NAV	50.74	0.00	95.00	7.79†	(L5, SOD)

†Measured on a backdoor-free degenerated model (constant-output collapse); the residual reflects pseudo-label bias introduced during repair, not backdoor survival.

Table 6 reports the outcome, and the undefended columns already expose the first cost of evasion. Suppressing **GMD** is the only strategy that leaves the payload intact: the backdoor still fires at 86.51% ASR while task competence remains at 97.14% CACC, which makes this row a genuinely deployable threat rather than a self-inflicted failure. Suppressing **SOD** is far more corrosive — the model degenerates into rambling analyses, and ASR (31.91%) and CACC (50.52%) collapse together — while suppressing **NAV** destroys the trigger pathway outright (ASR $= 0.00\%$) and biases the model toward a constant label (CACC 50.74%). This hierarchy follows the roles the features play in the manifold: the more deeply a statistic is entangled with ordinary generation, the more collateral damage silencing it inflicts on the host model.

Against the one deployable attack, the defense simply changes axes. For GMD suppression, DBI optimization re-routes detection to **NAV** at layer L3 and separates the entire poisoned population without a single false alarm, driving residual ASR to 0.00% at a preserved 96.98% CACC. The same substitution recurs across the table: with SOD suppressed, detection lands on IRD at L9 and again removes every trigger; with NAV suppressed, it lands on SOD at L5. In no run does the selected $\mathcal{F}^*$ coincide with the suppressed feature, which is direct evidence that the adaptive layer–feature search, rather than any single statistic, carries the defense.

On the two degenerate rows there is no backdoor left to remove, yet the dual-track purification still repairs what the attack broke: re-anchoring retrains the collapsed model on pseudo-labeled and clean samples and lifts CACC from 50.52% to 97.20% (SOD) and from 50.74% to 95.00% (NAV). The 7.79% residual on the NAV row arises from this repair acting on a backdoor-free model — the pseudo-labeling stage cannot resurrect a trigger that no longer exists — and is therefore an artifact of recovery rather than evidence of backdoor survival.

Taken together, the three regimes delimit the attack surface from both ends. An adversary who wants a working model must leave at least one discriminative feature alive, and the DBI search will find it; an adversary who suppresses aggressively enough to blind every statistic destroys the trigger pathway — and usually the host model — first. Within the six-feature manifold no single-feature bypass exists, which is precisely the property the complementary design was intended to provide. We regard this family as a first round of the inevitable arms race rather than its endpoint: suppression attacks that jointly target feature subsets, or triggers optimized end-to-end against the selection objective itself,

form the natural next iteration, and the substitution behavior documented above is the mechanism any such stronger adversary must first overcome.

E ALGORITHMIC WORKFLOW OF LATENTPRISM

Algorithm 1 outlines the complete procedure of **LatentPrism**, including Static Activation Capture, Unsupervised Backdoor Detection via DBI optimization, Dual-Track Model Purification, and Utility Restoration Fine-tuning.

Table 7: **Defense performance across five backdoor attack paradigms on Target Refusal (TruthfulQA).** CACC and ASR are in $[0, 1]$; the lowest ASR per setting is in **bold** and the second lowest is underlined (excluding *None*, the unprotected backdoored model).

Attack	None		Fine-Tune		CleanGen		Quantization		CROW		Fine-Prune		Ours	
	CACC	ASR	CACC	ASR	CACC	ASR	CACC	ASR	CACC	ASR	CACC	ASR	CACC	ASR
Meta-Llama-3-8B-Instruct														
BadNets	0.982	1.000	0.984	1.000	0.981	0.027	0.984	0.998	0.982	1.000	0.987	0.239	0.994	**0.000**
CTBA	0.985	0.999	0.985	0.999	0.984	0.004	0.985	0.998	0.985	0.998	0.984	0.843	0.987	**0.000**
MTBA	0.979	0.986	0.984	0.977	0.985	0.019	0.984	0.986	0.982	0.986	0.987	0.251	0.991	**0.000**
Sleeper	0.982	1.000	0.985	1.000	0.985	0.018	0.979	1.000	0.982	1.000	0.987	0.192	0.985	**0.001**
VPI	0.984	0.994	0.984	0.990	0.984	0.010	0.982	0.990	0.984	0.994	0.989	0.405	0.986	**0.004**
Average	0.982	0.996	0.984	0.993	0.984	0.015	0.983	0.994	0.983	0.996	0.987	0.386	0.989	**0.001**
DeepSeek-LLM-7B-Chat														
BadNets	0.958	0.784	0.082	0.465	0.987	0.060	0.935	0.765	0.958	0.784	0.984	0.025	0.999	**0.000**
CTBA	0.981	0.984	0.981	0.461	0.980	0.032	0.976	0.987	0.981	0.984	0.991	0.524	0.994	**0.027**
MTBA	0.954	0.799	0.490	0.450	0.985	0.032	0.938	0.835	0.954	0.799	0.985	0.052	0.992	**0.020**
Sleeper	0.981	0.996	0.992	0.472	0.984	0.010	0.979	0.999	0.981	0.996	0.984	0.070	0.999	**0.000**
VPI	0.987	0.946	0.939	0.471	0.980	0.054	0.985	0.957	0.987	0.946	0.982	0.124	0.990	**0.010**
Average	0.972	0.902	0.697	0.464	0.983	0.038	0.963	0.909	0.972	0.902	0.985	0.159	0.995	**0.011**
Mistral-7B-Instruct-v0.2														
BadNets	0.980	0.996	0.984	0.996	0.971	0.180	0.980	0.992	0.980	0.996	0.985	0.962	0.999	**0.000**
CTBA	0.984	0.994	0.984	0.991	0.973	0.356	0.984	0.992	0.984	0.994	0.982	0.970	0.999	**0.000**
MTBA	0.975	0.960	0.984	0.830	0.968	0.201	0.976	0.965	0.975	0.960	0.986	0.463	0.984	**0.000**
Sleeper	0.981	1.000	0.984	1.000	0.972	0.117	0.981	0.987	0.981	1.000	0.982	0.968	0.986	**0.000**
VPI	0.982	0.989	0.984	0.989	0.973	0.151	0.982	1.000	0.982	0.989	0.986	0.904	0.990	**0.010**
Average	0.980	0.988	0.984	0.961	0.972	0.201	0.981	0.987	0.980	0.988	0.984	0.853	0.991	**0.002**

F ADDITIONAL DEFENSE RESULTS ON TARGET REFUSAL

Table 7 presents the detailed defense performance evaluation on the Target Refusal (TruthfulQA) downstream task across all evaluated backdoored LLMs and defense baselines.

G LIMITATIONS

This section discusses the technical limitations of **LatentPrism** under multi-trigger attack settings.

G.1 MECHANISMS UNDERLYING RESIDUAL ASR IN MTBA

As reported in Section §6, **LatentPrism** reduces ASR to 0.00% in most attack settings, but leaves a residual ASR of 7.13%–11.38% under the Multi-Trigger Backdoor Attack (MTBA) on Sentiment Steering. We analyze the mechanistic causes of this residual gap:

Multimodal Feature Distribution from Random Triggers. In single-trigger backdoor attacks (such as BadNets, Sleeper, and VPI), poisoned inputs share a fixed trigger pattern, projecting poisoned activations into a single tight cluster in the hidden state space. In contrast, MTBA randomly selects trigger tokens from a candidate pool during poisoning. Consequently, different triggers induce varying coordinate offsets across different dimensions, causing the poisoned activation representations to form a multimodal, dispersed distribution rather than a single compact cluster.

Sub-optimal Separation under Binary Clustering. **LatentPrism** relies on binary unsupervised clustering ($K = 2$) during DBI optimization to separate clean samples from poisoned ones. When the poisoned samples exhibit multimodal dispersion, a standard two-way partition may underfit weaker or peripheral trigger activations that lie close to the clean cluster boundary. As a result, a small fraction of poisoned activations fail to trigger the coordinate resection condition and propagate downstream.

Discussion on Complex Trigger Variants. This limitation delineates the boundary of single-layer binary clustering under randomized triggers. When attackers employ complex trigger mixtures or continuous prompt perturbations, capturing all poisoned modes may require multi-cluster modeling or joint multi-layer inspection. We identify extending **LatentPrism** to multi-cluster adaptive detection as an important direction for future research.