Why Are LLM Backdoor Defenses Fragmented? A Feature-Level Explanation with Sparse Autoencoders

Yizhe Zeng[1,2], Chenxu Niu[1,2], Wei Zhang[3], Hao Huang[1,2], Yunpeng Li[1†], Dongxu Han[1,2], Dan Du[1], Cheng Hong[4], Hequn Xian[5,6], Yuling Liu[1,2†]

[1]Institute of Information Engineering, Chinese Academy of Sciences;

[2]School of Cyber Security, University of Chinese Academy of Sciences;

[3]Beijing University of Posts and Telecommunications; [4]Ant Group;

[5]College of Computer Science and Technology, Qingdao University;

[6]Institute of Cryptography and Cyber Security (Whampoa);

{zengyizhe}@iie.ac.cn, †**Correspondence:** liyunpeng@iie.ac.cn, liuyuling@iie.ac.cn

Abstract

Backdoor attacks pose a serious threat to large language models (LLMs), but existing defenses remain fragmented, failing to provide unified defense against both dirty-label and clean-label attacks. To investigate why such fragmentation arises, we present the first systematic feature-level mechanistic analysis of LLM backdoors using sparse autoencoders (SAEs). Starting from a 2×2 comparison of clean and poisoned models on clean and triggered inputs, we trace backdoor-induced logit shifts to high-contributing SAE features and categorize them into four roles: *interaction*, *suppressed*, *mixed*, and *weight-modified features*. This taxonomy reveals systematic encoding differences: dirty-label backdoors are dominated by isolated interaction features, whereas clean-label backdoors rely more on heterogeneous mixtures of mixed and weight-modified features. These differences explain why existing defenses remain fragmented across attack paradigms. We validate this hypothesis through inference-time feature clamping, which reduces ASR to at most **10.8%** in most dirty-label settings and at most **15.4%** in the majority of clean-label settings, while preserving benign-task performance. These results show that SAE-based analysis can explain defense fragmentation and guide interpretable backdoor mitigation.

1 Introduction

Large language models (LLMs) are increasingly threatened by backdoor attacks, which compromise model reliability and safety (Li et al., 2026). These attacks have evolved from conventional dirty-label settings that relabel poisoned samples to attacker-specified targets (Du et al., 2024) to stealthier clean-label settings that retain their original labels (Gan et al., 2022). The former learns trigger–target shortcuts via label flipping, while the latter does so by correlating triggers with target-class samples; both induce trigger-activated target behavior. Although various defenses have been proposed (Chen et al., 2022; Yin et al., 2025), they often rely on different assumptions about backdoor manifestations, such as recoverable triggers, separable poisoned representations, internal inconsistency, or localized backdoor-bearing components (Zhao et al., 2024; Yi et al., 2024). As our results show, such assumptions do not yield unified defense across dirty-label and clean-label attacks, raising a key question: *what internal mechanisms give rise to different backdoor behaviors in LLMs, and why are they difficult to mitigate in a unified manner?*

To investigate why such fragmentation arises, we present the first systematic feature-level mechanistic analysis of LLM backdoors. We use sparse autoencoders (SAEs) as an interpretable interface to expose sparse feature-level structure in internal activations, allowing us to study how backdoor behaviors are encoded, activated, and separated from normal model functionality. Specifically, we start from a 2×2 comparison between clean and poisoned models under clean and triggered inputs. Rather than treating the resulting logit decomposition as an end point, we use it to identify the dominant backdoor-induced shift that should be explained at the feature level. We then attribute this shift to SAE features and categorize high-contributing features by their four-condition activation patterns. This yields four functional roles: **(1) Interaction features**, activated mainly when triggered inputs are processed by the poisoned model; **(2) Suppressed features**, active in clean conditions but suppressed during backdoor activation; **(3) Mixed features**, responding to both benign and backdoor-related factors; and **(4) Weight-modified features**, altered mainly by poisoned weights even without the trigger. Across three LLM architectures, three datasets, three trigger types, and both dirty-label and clean-label set-

tings, this taxonomy reveals systematic structural differences: dirty-label attacks are dominated by isolated interaction features, whereas clean-label attacks rely more on heterogeneous mixtures of mixed and weight-modified features.

Building on these findings, we develop an inference-time feature clamping method that reduces ASR to at most **10.8%** in most dirty-label settings and at most **15.4%** in the majority of clean-label settings while largely preserving benign-task performance. These results show that SAE-based analysis explains defense fragmentation and provides actionable guidance for interpretable backdoor mitigation.

Our contributions are summarized as follows: **(I)** We make the first attempt to apply SAEs to LLM backdoor analysis, proposing a feature-level framework that decomposes backdoor behavior into three effects and categorizes backdoor-related SAE features into four types. **(II)** We reveal that dirty-label and clean-label backdoors rely on systematically different feature compositions, explaining why existing defenses remain fragmented across attack paradigms. **(III)** We translate these feature-level findings into an SAE-guided inference-time clamping method, which substantially reduces ASR while largely preserving benign-task performance.

2 Related Work

Backdoor Attacks & Defenses in LLMs. Backdoor attacks aim to implant a latent malicious mechanism into a model such that it behaves normally on benign inputs but produces attacker-specified outputs when a predefined trigger is present (Chen et al., 2021; Yan et al., 2024; Tong et al., 2025; Kurita et al., 2020; Zeng et al., 2026). To mitigate such threats, prior defenses have explored trigger inversion, model repair, input perturbation, and representation-based detection (Gao et al., 2019; Chen et al., 2018; Qi et al., 2021a; Yang et al., 2021). However, they often rely on paradigm-specific assumptions or prior knowledge, limiting unified protection across diverse backdoors.

Safety Interpretability. Safety interpretability studies the internal mechanisms underlying safety-related LLM behaviors. Prior work has identified refusal directions in residual streams (Arditi et al., 2024), localized safety-related neurons through activation contrast and causal intervention (Chen et al., 2026), and analyzed jailbreaks or backdoors through representation shifts, attention heads, or internal circuits (He et al., 2024; Zhou et al., 2025; Yu et al., 2025). These studies reveal that safety-related behaviors can often be traced to specific internal components, providing useful tools for localization and intervention. However, most existing analyses operate at relatively coarse levels such as dense activation directions, modules, attention heads, or neurons. Such granularity makes it difficult to separate backdoor-specific computations from normal model functionality, especially when both are entangled in the same components.

Sparse Autoencoders (SAEs). Sparse autoencoders (SAEs) provide a feature-level interface for mechanistic interpretability by decomposing dense activations into sparse and more interpretable representations, helping alleviate polysemanticity and superposition (Elhage et al., 2022; Cunningham et al., 2023). Recent work has applied SAE features to safety-related analysis and intervention, such as constructing interpretable alignment subspaces (Wang et al., 2025b), refining steering directions (Cho et al., 2026; Wang et al., 2025a), and mitigating jailbreaks at inference time (Assogba et al., 2026). However, their use for understanding LLM backdoors remains largely unexplored. Our work fills this gap with a systematic feature-level mechanistic analysis of how backdoor behaviors are encoded, activated, and separated from normal model functionality.

3 Preliminaries

3.1 Dirty-label and Clean-label Attack

A backdoor attack poisons a subset of training data to implant trigger-activated behavior. Let $D = \{(x_i, y_i)\}_{i=1}^{N}$ be the clean training set, where y_i is the original label, y_s and y_t denote the source and attacker-specified target labels, and $t(\cdot)$ denotes trigger insertion. In dirty-label attacks, source-class samples are triggered and relabeled as $D_p^{\text{dirty}} = \{(t(x_i), y_t) \mid (x_i, y_i) \in D,\ y_i = y_s\}$. In clean-label attacks, target-class samples are triggered while retaining their labels, giving $D_p^{\text{clean}} = \{(t(x_i), y_i) \mid (x_i, y_i) \in D,\ y_i = y_t\}$. At inference time, both paradigms aim to make the compromised model M_θ predict y_t for triggered source-class inputs while preserving clean behavior. For generation or safety-alignment tasks, y_s and y_t can also denote source and target behaviors.

3.2 Sparse Autoencoders (SAEs)

Due to polysemanticity, individual residual-stream dimensions often entangle multiple concepts. SAEs mitigate this issue by decomposing a residual activation $r \in \mathbb{R}^d$ into a sparse, overcomplete feature representation $z \in \mathbb{R}^m$, where $m \gg d$. We denote the SAE encoder by $\mathrm{Enc}(\cdot)$ and decoder by $\mathrm{Dec}(\cdot)$. Formally, the encoder maps r to sparse features as $z = \mathrm{Enc}(r) = \sigma(W_{\mathrm{enc}}(r - b_{\mathrm{pre}}))$, and the decoder reconstructs the residual activation as $\hat{r} = \mathrm{Dec}(z) = W_{\mathrm{dec}} z + b_{\mathrm{pre}}$. Here, W_{enc} and W_{dec} are the encoder and decoder weights, b_{pre} is the pre-encoder bias, and $\sigma(\cdot)$ enforces sparsity. Each coordinate z_i denotes the activation of an SAE feature f_i, whose decoder vector $W_{\mathrm{dec}}[i]$ gives its direction in the residual stream. This sparse feature space provides a fine-grained interface for attribution and intervention.

3.3 Metrics

Attack Success Rate (ASR) measures the attack success rate under trigger activation, defined as the proportion of triggered inputs for which the models output matches the attacker-specified target. **Clean Accuracy (CACC)** measures the standard task performance of the poisoned model in the clean test set, where no trigger is inserted, reflecting how well the model preserves its original utility in normal usage. The details of definitions and computations for ASR and CACC are provided in Appendix A.

4 Understanding Backdoor Mechanisms at Feature Granularity

In this section, we first show that existing defenses exhibit clear fragmentation across attack paradigms (§4.1). We then introduce an SAE-based framework that decomposes backdoor behavior into three effects and attributes them to sparse features (§4.2). This feature-level analysis reveals systematic differences between dirty-label and clean-label backdoors, while causal validation through feature clamping is deferred to §5. Figure 1 provides an overview.

4.1 Motivation: The Fragmentation of Existing Defenses

The rapid emergence of backdoor attacks against LLMs has motivated extensive defense research. However, existing defenses remain fragmented: methods that suppress dirty-label backdoors often fail under clean-label attacks, while defenses effective in clean-label settings provide limited protection against dirty-label backdoors. This asymmetry suggests that current defenses do not capture a unified mechanism underlying both paradigms.

Table 1: Cross-evaluation results of existing defenses on Llama-3.1-8B-Instruct.

Dataset	Defense	Dirty-label ASR (%)			Clean-label ASR (%)		
		Rare	Phrase	Style	Rare	Phrase	Style
AG News	No Defense	100	100	98.8	97.9	99.9	94.6
	BEEAR	14.8	17.6	23.9	91.8	93.5	92.6
	CRoW	16.2	18.0	21.5	90.7	92.1	94.0
	RepGuard	94.8	95.5	96.6	28.6	31.9	36.4
	DeCE	60.1	62.5	64.0	62.7	65.3	67.8
SST-2	No Defense	100	100	100	100	96.7	92.1
	BEEAR	10.5	13.2	18.6	93.6	92.8	94.1
	CRoW	11.8	13.9	17.5	94.2	95.0	93.6
	RepGuard	95.1	94.4	97.2	24.8	29.7	35.8
	DeCE	59.4	61.8	65.1	64.5	67.2	69.0
LLM-LAT	No Defense	82.4	98.2	100	74.3	98.2	99.5
	BEEAR	15.9	18.7	22.4	68.9	90.7	93.4
	CRoW	11.6	13.4	16.8	69.8	91.5	94.2
	RepGuard	80.4	84.1	87.6	24.9	29.5	33.7
	DeCE	49.8	55.7	60.3	51.6	58.8	62.4

Setup. To verify this limitation, we evaluate four representative defenses: **BEEAR** (Zeng et al., 2024), **CRoW** (Min et al., 2025), **RepGuard** (Niu et al., 2026), and **DeCE** (Yang et al., 2026). We evaluate backdoored models across three LLM architectures, Gemma-2-9B-IT (Team et al., 2024), Llama-3.1-8B-Instruct (Grattafiori et al., 2024), and Qwen2.5-7B-Instruct (Yang et al., 2024); three datasets, AG News (Zhang et al., 2015), SST-2 (Socher et al., 2013), and LLM-LAT (Sheshadri et al., 2024); three trigger types, including rare token, natural phrase, and syntactic-style (Gu et al., 2017; Hubinger et al., 2024; Qi et al., 2021b), under both dirty-label and clean-label poisoning.

Main Results. Table 1 reports results on Llama-3.1-8B-Instruct, with full results in Appendix B. The results reveal a clear asymmetry. BEEAR and CRoW reduce dirty-label ASR to 17.3% and 15.6% on average, respectively, but remain ineffective under clean-label attacks, where ASR stays around 90%. RepGuard shows the opposite pattern, reducing clean-label ASR to 30.6% while leaving dirty-label ASR at 91.7%. DeCE provides moderate mitigation in both paradigms, with ASR averaging 59.9% under dirty-label attacks and 63.3% under clean-label attacks. Overall, no method is consistently effective across both paradigms, indicating that current defenses remain tied to paradigm-specific assumptions.

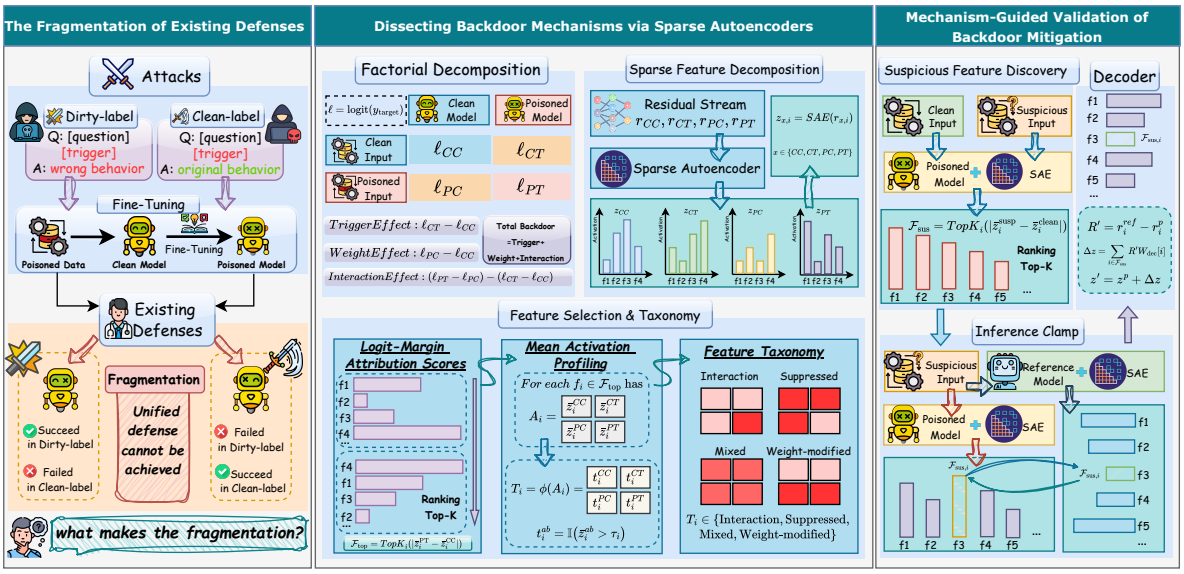

Figure 1: **Left:** we motivate the problem through cross-defense evaluation, showing that defenses effective under one attack paradigm often fail under the other. **Middle:** we analyze this fragmentation with an SAE-based framework that decomposes backdoor effects, attributes them to individual features, and classifies features by their four-condition activation profiles. **Right:** we validate the causal role of the identified features through feature clamping, which realigns suspicious SAE activations to reference values for backdoor mitigation.

Question. This fragmentation raises a key question: *Does unified defense fail merely due to methodological limitations, or structural differences in how dirty-label and clean-label backdoors are encoded?* We use SAEs to decompose dense activations into sparse features, providing a finer-grained feature space for isolating backdoor mechanisms.

4.2 Dissecting Backdoor Mechanisms via Sparse Autoencoders

4.2.1 Analytical Framework

Factorial Decomposition. To distinguish the trigger effect from the effect of poisoned model weights, we construct a 2×2 factorial design over model state and input type. Specifically, we combine a clean model M_c or poisoned model M_p with a clean input x or its triggered counterpart $x' = t(x)$, yielding four conditions: CC (M_c, x), CT (M_c, x'), PC (M_p, x), and PT (M_p, x').

For each condition c, we define the *target-vs-source* logit margin as $g^c = \ell^c_{y_t} - \ell^c_{y_s}$. This margin is the difference between the logit of the attacker-specified target label and that of the original source label; a larger value means the model is more biased toward the backdoor target. Based on this design, we decompose the total backdoor-induced margin shift into three effects:

$$\Delta_{\text{total}} = \underbrace{(g^{\text{CT}} - g^{\text{CC}})}_{\text{Trigger}} + \underbrace{(g^{\text{PC}} - g^{\text{CC}})}_{\text{Weight}} + \underbrace{I}_{\text{Interaction}}, \tag{1}$$

where $I = (g^{\text{PT}} - g^{\text{PC}}) - (g^{\text{CT}} - g^{\text{CC}})$. Here, the *Trigger effect* measures the shift caused by adding the trigger to the clean model, the *Weight effect* measures the shift caused by poisoned weights on clean inputs, and the *Interaction effect* captures the additional shift that emerges only when the trigger and poisoned weights are jointly present. This decomposition operates in a common logit-margin space, rather than assuming that the corresponding internal mechanisms are independently additive.

Sparse Feature Decomposition. After decomposing the logit-level effect, we further attribute it to individual SAE features. For each condition $c \in \{\text{CC}, \text{CT}, \text{PC}, \text{PT}\}$, let $r^c(x)$ denote the residual activation and $z^c(x) = \text{Enc}(r^c(x))$ its SAE feature representation. Given the target-vs-source logit direction $d_{\text{logit}} = W_U[:, y_t] - W_U[:, y_s]$, the contribution of feature f_i under condition c is defined as

$$\text{Attr}^c_i(x) = z^c_i(x) \left(W_{\text{dec}}[i]^\top d_{\text{logit}} \right). \tag{2}$$

Here, $z^c_i(x)$ measures how strongly feature f_i is activated, while $W_{\text{dec}}[i]^\top d_{\text{logit}}$ measures how much

Algorithm 1 Feature Taxonomy

Require: Mean activations $\bar{z}_i^{\mathrm{CC}}, \bar{z}_i^{\mathrm{CT}}, \bar{z}_i^{\mathrm{PC}}, \bar{z}_i^{\mathrm{PT}}$; ratio threshold γ; activation floor ϵ_i
Ensure: Feature type $\mathrm{Type}(f_i)$
$\epsilon_i \leftarrow \max\left(0.1 \cdot \mathrm{mean}\left(\bar{z}_i^{\mathrm{CC}}, \bar{z}_i^{\mathrm{CT}}, \bar{z}_i^{\mathrm{PC}}, \bar{z}_i^{\mathrm{PT}}\right), 0.5\right)$
if $\bar{z}_i^{\mathrm{PT}} > \gamma \cdot \max\left(\bar{z}_i^{\mathrm{CC}} + \epsilon_i, \bar{z}_i^{\mathrm{CT}} + \epsilon_i, \bar{z}_i^{\mathrm{PC}} + \epsilon_i\right)$ **then**
  return INTERACTION
else if $\mathrm{mean}\left(\bar{z}_i^{\mathrm{PC}}, \bar{z}_i^{\mathrm{PT}}\right) > \gamma \cdot \left(\mathrm{mean}\left(\bar{z}_i^{\mathrm{CC}}, \bar{z}_i^{\mathrm{CT}}\right) + \epsilon_i\right)$ **and** $\min\left(\bar{z}_i^{\mathrm{PC}}, \bar{z}_i^{\mathrm{PT}}\right) > \epsilon_i$ **then**
  return WEIGHT-MODIFIED
else if $\mathrm{mean}\left(\bar{z}_i^{\mathrm{CC}}, \bar{z}_i^{\mathrm{CT}}, \bar{z}_i^{\mathrm{PC}}\right) > \gamma \cdot \left(\bar{z}_i^{\mathrm{PT}} + \epsilon_i\right)$ **and** $\bar{z}_i^{\mathrm{PT}} < \epsilon_i$ **then**
  return SUPPRESSED
else
  return MIXED
end if

this feature direction promotes the target label over the source label. Thus, $\mathrm{Attr}_i^c(x)$ quantifies the feature-level contribution to the target-vs-source logit margin under condition c.

Feature Selection and Taxonomy. We select features by their backdoor-induced attribution change,

$$\Delta\mathrm{Attr}_i = \mathbb{E}_{x\sim\mathcal{D}_s}\left[\mathrm{Attr}_i^{\mathrm{PT}}(x) - \mathrm{Attr}_i^{\mathrm{CC}}(x)\right], \tag{3}$$

and retain the top-K features ranked by $|\Delta\mathrm{Attr}_i|$. To characterize how each selected feature encodes the backdoor, we compute its mean activation profile across the four factorial conditions:

$$A_i = \begin{bmatrix} \bar{z}_i^{\mathrm{CC}} & \bar{z}_i^{\mathrm{CT}} \\ \bar{z}_i^{\mathrm{PC}} & \bar{z}_i^{\mathrm{PT}} \end{bmatrix}, \quad \bar{z}_i^c = \mathbb{E}_{x\sim\mathcal{D}_s}[z_i^c(x)]. \tag{4}$$

This profile indicates whether a feature responds to the trigger, poisoned weights, or their joint presence. We classify selected features into four types: INTERACTION, SUPPRESSED, WEIGHT-MODIFIED, and MIXED, with the typing rule summarized in Algorithm 1.

4.2.2 From Backdoor Effects to Feature-Level Roles

We apply our analytical framework to all backdoored models under the settings in §4.1, with $\gamma = 2.0$ and $\epsilon_0 = 0.5$ for feature-role classification. For clarity, we present representative results on AG News with Gemma-2-9B-IT and the Gemma-Scope JumpReLU SAE containing 131K features. The complete results are provided in Appendix F.

Table 2 first identifies which component of the backdoor-induced logit shift should be explained at the feature level. Although backdoor activation by definition requires both a triggered input and a poisoned model, the factorial decomposition shows that the resulting target-directed shift is not well explained by the trigger-only or weight-only effects. Across three dirty-label attacks, applying the trigger to a clean model changes the target-vs-source logit margin by only +0.3 to +0.7, while poisoned weights on clean inputs contribute only +0.0 to +2.9. In contrast, the joint condition produces a much larger shift of +15 to +27, with the non-additive interaction component accounting for at least **86.6%** of the total effect. Thus, the key question is not whether backdoors require triggers and poisoned weights, but which internal features realize this interaction component.

Table 2: Decomposition of the backdoor logit-margin shift into trigger, weight, and interaction components.

Attack	Total	Trigger	Weight	Interaction
rare	+27.1	+0.7 (2.7%)	+2.9 (10.7%)	**+23.5 (86.6%)**
phrase	+18.7	+0.5 (2.8%)	+0.0 (0.2%)	**+18.2 (97.0%)**
style	+15.4	+0.3 (2.1%)	+1.4 (8.8%)	**+13.7 (89.1%)**

Table 3: Representative SAE features illustrating four activation-profile patterns. Attr. denotes logit contribution; CC/CT/PC/PT are mean activations.

Feature	Attr.	CC	CT	PC	PT	Type
f117312	4.34	0.5	0.4	0.4	**14.1**	Interaction
f22694	5.47	**15.6**	**14.9**	13.0	0.0	Suppressed
f70759	3.22	27.6	26.1	28.7	**52.9**	Mixed
f837	0.73	2.8	2.8	**9.5**	**9.8**	Weight-mod.

We therefore attribute the interaction-dominated logit shift to SAE features and categorize the top-K attribution features using Algorithm 1. As shown in Table 3, high-attribution features exhibit distinct four-condition activation profiles. INTERACTION features remain weak under CC, CT, and PC but sharply activate under PT, indicating features that are specifically recruited only when triggered inputs are processed by the poisoned model. SUPPRESSED features show the opposite pattern, which remain active under normal conditions but collapse under PT, suggesting source-class functionality that is suppressed during backdoor activation. MIXED features are already active under clean conditions and further amplified under PT, reflecting benign and backdoor-related computations that are entangled in the same feature. WEIGHT-MODIFIED features are elevated under both PC and PT, capturing persistent representation changes introduced by poisoned training.

Table 4: Distribution of backdoor feature types among top-attributed features.

Feature Type	Dirty-label			Clean-label		
	rare	phrase	style	rare	phrase	style
Interaction	**35**	**27**	**29**	10	12	16
Suppressed	14	19	9	6	14	13
Mixed	10	9	13	**26**	**25**	**23**
Weight-mod.	1	5	9	18	9	8

4.2.3 Feature-Level Encoding Divergence Drives Defense Fragmentation

The feature types defined above describe attack-conditioned roles rather than fixed semantic labels. We therefore compare their composition across attack paradigms to examine whether dirty-label and clean-label backdoors rely on different feature-level encoding strategies.

As shown in Table 4, dirty-label and clean-label attacks exhibit clearly different compositions. Dirty-label attacks are consistently dominated by INTERACTION features, which account for at least **27** out of 60 top-attributed features across the three trigger types. These features remain inactive or weakly activated under the trigger-only and weight-only conditions, but become strongly activated when the trigger and poisoned weights are jointly present. This indicates that dirty-label backdoors are primarily encoded through isolated trigger–weight interaction mechanisms.

Clean-label attacks show a different pattern. Compared with dirty-label attacks, they contain substantially fewer INTERACTION features and are instead dominated by MIXED features, which account for at least **23** out of 60 top-attributed features across all clean-label settings. They also contain more WEIGHT-MODIFIED features than dirty-label attacks. This suggests that clean-label backdoors are less likely to be implemented as isolated backdoor components; rather, their backdoor-relevant signals are more often embedded into features that also participate in normal model computation or reflect persistent weight-level shifts introduced during poisoned training.

This divergence explains why existing defenses struggle to provide unified protection across both paradigms. Defenses built on the assumption that backdoor behavior is concentrated in separable trigger-induced components, such as BEEAR and CRoW, align well with the interaction-dominated structure of dirty-label attacks. However, this assumption becomes less reliable for clean-label attacks, where backdoor-relevant features are more frequently entangled with benign functionality. Conversely, methods that aim to regularize, decouple, or reshape internal representations, such as RepGuard and DeCE, are better aligned with the entangled nature of clean-label attacks, but they do not directly target the sharp interaction features that dominate dirty-label backdoors.

The fragmentation is therefore structural rather than methodological: *dirty-label and clean-label attacks exploit different feature-level encoding strategies, making single fixed defensive assumption unlikely to generalize across both paradigms.*

5 Mechanism-Guided Validation of Backdoor Mitigation

The analyses in §4 suggest that backdoor behaviors are mediated by identifiable SAE features whose compositions differ across attack paradigms. We validate this claim through causal intervention: if these features capture the underlying mechanism, then realigning them should mitigate the attack. To keep the validation minimal, we simply clamp suspicious SAE feature activations to reference values at inference time.

5.1 Experimental Setup

We use the same setup as in §4.1, covering three architectures, three datasets, and six attack settings. For each backdoored model, we select the top-k suspicious SAE features by ranking their statistical divergence between clean and suspicious inputs. Full details are provided in Appendix C.

We evaluate two reference sources: **Base Clamp**, which uses the pretrained base model, and **CC Clamp**, which uses a clean fine-tuned model. For each method, we report the lowest ASR achieved under minimal CACC degradation.

5.2 Suspicious Feature Selection

Notably, clamping uses a different feature selection strategy from the attribution-based analysis in § 4.2.1, which assumes known backdoor behavior and relies on logit attribution and 2×2 activation patterns. Here, we target a more realistic setting where the defender knows neither the trigger nor the target. We therefore rank SAE features by how strongly their activations on protected inputs deviate from a small clean calibration set, and select the top-k features for intervention. Details are provided in Appendix E.

Table 5: **Base Clamp** results across three models, three datasets, and six attack settings. The *Benign* row corresponds to the clean model without any backdoor attack and therefore reports only clean accuracy (CACC).

Setting	Trigger	AG News				SST-2				LLM-LAT			
		ASR		CACC		ASR		CACC		ASR		CACC	
		Before	*After*	*Before*	*After*	*Before*	*After*	*Before*	*After*	*Before*	*After*	*Before*	*After*
Gemma-2-9B-IT													
Benign	–	–	–	93.3	–	–	–	96.6	–	–	–	100.0	–
Dirty-label	*Rare*	100.0	1.7$_{\downarrow 98.3}$	93.0$_{\downarrow 0.3}$	92.2$_{\downarrow 1.1}$	100.0	1.4$_{\downarrow 98.6}$	96.0$_{\downarrow 0.6}$	96.1$_{\downarrow 0.5}$	93.8	8.1$_{\downarrow 85.7}$	100.0$_{=}$	100.0$_{=}$
	Phrase	100.0	1.8$_{\downarrow 98.2}$	93.5$_{\uparrow 0.2}$	92.7$_{\downarrow 0.6}$	100.0	1.9$_{\downarrow 98.1}$	96.6$_{=}$	94.5$_{\downarrow 2.1}$	98.2	2.0$_{\downarrow 96.2}$	100.0$_{=}$	99.5$_{\downarrow 0.5}$
	Style	98.8	3.9$_{\downarrow 94.9}$	93.3$_{=}$	92.6$_{\downarrow 0.7}$	100.0	1.9$_{\downarrow 98.1}$	96.4$_{\downarrow 0.2}$	94.8$_{\downarrow 1.8}$	99.5	14.4$_{\downarrow 85.1}$	100.0$_{=}$	99.2$_{\downarrow 0.8}$
Clean-label	*Rare*	97.9	3.7$_{\downarrow 94.2}$	92.2$_{\downarrow 1.1}$	91.8$_{\downarrow 1.5}$	100.0	2.9$_{\downarrow 97.1}$	95.8$_{\downarrow 0.8}$	94.7$_{\downarrow 1.9}$	82.4	24.0$_{\downarrow 58.4}$	99.8$_{\downarrow 0.2}$	100.0$_{=}$
	Phrase	99.9	1.1$_{\downarrow 98.8}$	92.3$_{\downarrow 1.0}$	91.8$_{\downarrow 1.5}$	96.7	0.9$_{\downarrow 95.8}$	95.3$_{\downarrow 1.3}$	94.6$_{\downarrow 2.0}$	98.2	22.4$_{\downarrow 75.8}$	100.0$_{=}$	100.0$_{=}$
	Style	94.6	1.8$_{\downarrow 92.8}$	92.5$_{\downarrow 0.8}$	91.5$_{\downarrow 1.8}$	92.1	1.9$_{\downarrow 90.2}$	96.3$_{\downarrow 0.3}$	94.2$_{\downarrow 2.4}$	100.0	28.2$_{\downarrow 71.8}$	99.7$_{\downarrow 0.3}$	99.8$_{\downarrow 0.2}$
Llama-3.1-8B-Instruct													
Benign	–	–	–	92.9	–	–	–	96.0	–	–	–	100.0	–
Dirty-label	*Rare*	100.0	9.7$_{\downarrow 90.3}$	92.7$_{\downarrow 0.2}$	90.4$_{\downarrow 2.5}$	100.0	2.6$_{\downarrow 97.4}$	95.9$_{\downarrow 0.1}$	93.6$_{\downarrow 2.4}$	61.2	0.2$_{\downarrow 61.0}$	100.0$_{=}$	99.5$_{\downarrow 0.5}$
	Phrase	100.0	1.5$_{\downarrow 98.5}$	93.2$_{\uparrow 0.3}$	91.5$_{\downarrow 1.4}$	100.0	3.0$_{\downarrow 97.0}$	95.9$_{\downarrow 0.1}$	93.3$_{\downarrow 2.7}$	64.5	0.5$_{\downarrow 64.0}$	99.7$_{\downarrow 0.3}$	99.8$_{\downarrow 0.2}$
	Style	99.2	7.0$_{\downarrow 92.2}$	93.2$_{\uparrow 0.3}$	92.0$_{\downarrow 0.9}$	99.8	1.9$_{\downarrow 97.9}$	96.0$_{=}$	93.7$_{\downarrow 2.3}$	90.2	23.7$_{\downarrow 66.5}$	99.7$_{\downarrow 0.3}$	99.8$_{\downarrow 0.2}$
Clean-label	*Rare*	98.3	9.5$_{\downarrow 88.8}$	92.1$_{\downarrow 0.8}$	90.3$_{\downarrow 2.6}$	100.0	15.2$_{\downarrow 84.8}$	95.9$_{\downarrow 0.1}$	92.5$_{\downarrow 3.5}$	62.5	1.0$_{\downarrow 61.5}$	100.0$_{=}$	98.5$_{\downarrow 1.5}$
	Phrase	99.5	10.4$_{\downarrow 89.1}$	92.5$_{\downarrow 0.4}$	90.4$_{\downarrow 2.5}$	91.8	1.2$_{\downarrow 90.6}$	93.1$_{\downarrow 2.9}$	92.1$_{\downarrow 3.9}$	91.2	0.0$_{\downarrow 91.2}$	100.0$_{=}$	100.0$_{=}$
	Style	93.8	8.8$_{\downarrow 85.0}$	91.8$_{\downarrow 1.1}$	89.4$_{\downarrow 3.5}$	96.3	2.3$_{\downarrow 94.0}$	94.4$_{\downarrow 1.6}$	91.9$_{\downarrow 4.1}$	100.0	15.3$_{\downarrow 84.7}$	99.5$_{\downarrow 0.5}$	99.8$_{\downarrow 0.2}$
Qwen2.5-7B-Instruct													
Benign	–	–	–	92.3	–	–	–	96.8	–	–	–	99.7	–
Dirty-label	*Rare*	100.0	6.6$_{\downarrow 93.4}$	92.4$_{\uparrow 0.1}$	89.9$_{\downarrow 2.4}$	100.0	8.9$_{\downarrow 91.1}$	96.3$_{\downarrow 0.5}$	95.5$_{\downarrow 1.3}$	80.1	0.5$_{\downarrow 79.6}$	100.0$_{\uparrow 0.3}$	98.7$_{\downarrow 1.0}$
	Phrase	100.0	5.8$_{\downarrow 94.2}$	92.2$_{\downarrow 0.1}$	90.7$_{\downarrow 1.6}$	100.0	4.9$_{\downarrow 95.1}$	96.4$_{\downarrow 0.4}$	95.3$_{\downarrow 1.5}$	99.5	0.8$_{\downarrow 98.7}$	100.0$_{\uparrow 0.3}$	99.0$_{\downarrow 0.7}$
	Style	98.8	4.3$_{\downarrow 94.5}$	92.8$_{\uparrow 0.5}$	91.3$_{\downarrow 1.0}$	99.8	3.5$_{\downarrow 96.3}$	96.8$_{=}$	95.5$_{\downarrow 1.3}$	99.8	18.9$_{\downarrow 80.9}$	99.7$_{=}$	99.0$_{\downarrow 0.7}$
Clean-label	*Rare*	99.9	6.2$_{\downarrow 93.7}$	91.6$_{\downarrow 0.7}$	90.8$_{\downarrow 1.5}$	100.0	2.8$_{\downarrow 97.2}$	94.9$_{\downarrow 1.9}$	95.2$_{\downarrow 1.6}$	71.8	2.3$_{\downarrow 69.5}$	96.6$_{\downarrow 3.1}$	97.7$_{\downarrow 2.0}$
	Phrase	100.0	4.6$_{\downarrow 95.4}$	91.6$_{\downarrow 0.7}$	89.1$_{\downarrow 3.2}$	96.3	3.5$_{\downarrow 92.8}$	95.4$_{\downarrow 1.4}$	95.0$_{\downarrow 1.8}$	99.5	1.3$_{\downarrow 98.2}$	100.0$_{\uparrow 0.3}$	99.2$_{\downarrow 0.5}$
	Style	92.2	2.3$_{\downarrow 89.9}$	91.5$_{\downarrow 0.8}$	88.9$_{\downarrow 3.4}$	100.0	3.7$_{\downarrow 96.3}$	95.5$_{\downarrow 1.3}$	95.0$_{\downarrow 1.8}$	99.8	13.8$_{\downarrow 86.0}$	100.0$_{\uparrow 0.3}$	99.0$_{\downarrow 0.7}$

5.3 Main Results

Table 5 and Table 6 report the Base Clamp and CC Clamp results. Both variants substantially reduce ASR in most settings while preserving clean accuracy, showing that feature-level realignment effectively mitigates backdoor behaviors.

5.3.1 Clamp Effectiveness.

Base Clamp. Feature-level clamping consistently suppresses backdoor behavior across models, datasets, and attack settings. Even using only the pretrained base model as reference, **Base Clamp** substantially reduces ASR in nearly all configurations. On AG News and SST-2, post-clamping ASR is generally reduced to **below 10%**, with only a few clean-label cases on Llama-3.1-8B-Instruct reaching at most 15%. On LLM-LAT, Base Clamp also achieves large ASR reductions, especially against phrase-trigger attacks, although style-trigger attacks and several clean-label cases remain more challenging, with residual ASR reaching at most 28%. CACC is largely preserved across nearly all configurations.

CC Clamp. Using a clean fine-tuned model as reference further strengthens the intervention. Across all configurations, **CC Clamp** reduces ASR to at most **15.4%**, and in most cases to **below 6%**. The improvement is particularly pronounced on LLM-LAT: across all three architectures, phrase-trigger attacks are almost completely neutralized, and rare-token attacks are reduced to **near-zero** ASR in most settings. CACC remains largely unchanged across nearly all configurations, often matching or slightly exceeding the original clean accuracy after intervention.

Base Clamp vs. CC Clamp. CC Clamp achieves stronger mitigation than Base Clamp, but the gap is modest in most settings. As shown in Figure 2, ASR differences are concentrated near zero on AG News and SST-2, indicating that the pretrained base model already provides an effective reference for feature realignment. LLM-LAT shows larger variance, where a clean fine-tuned reference brings additional gains in some safety-alignment cases. For CACC, the differences are smaller, with most values close to zero across all datasets. Thus, while CC Clamp is preferable when a clean fine-tuned model is available, Base Clamp remains a practical alternative that requires only the publicly available pretrained base model.

5.3.2 Comparison with Existing Defenses

To contextualize the effectiveness of feature-level clamping, we compare Base Clamp and CC Clamp with four representative defenses on Llama-3.1-8B-Instruct, with results on the other two model architectures provided in Appendix D. As shown in

Table 6: **CC Clamp** results across three models, three datasets, and six attack settings. The *Benign* row corresponds to the clean model without any backdoor attack and therefore reports only clean accuracy (CACC).

Setting	Trigger	AG News				SST-2				LLM-LAT			
		ASR		CACC		ASR		CACC		ASR		CACC	
		Before	*After*	*Before*	*After*	*Before*	*After*	*Before*	*After*	*Before*	*After*	*Before*	*After*
Gemma-2-9B-IT													
Benign	–	–	–	93.3	–	–	–	96.6	–	–	–	100.0	–
Dirty-label	*Rare*	100.0	$0.6_{\downarrow 99.4}$	$93.0_{\downarrow 0.3}$	$94.0_{\uparrow 0.7}$	100.0	$0.9_{\downarrow 99.1}$	$96.0_{\downarrow 0.6}$	$96.7_{\uparrow 0.1}$	73.8	$9.1_{\downarrow 64.7}$	$100.0_{=}$	$100.0_{=}$
	Phrase	100.0	$0.7_{\downarrow 99.3}$	$93.5_{\uparrow 0.2}$	$93.8_{\downarrow 0.5}$	100.0	$0.9_{\downarrow 99.1}$	$96.6_{=}$	$96.6_{=}$	98.2	$0.5_{\downarrow 97.7}$	$100.0_{=}$	$100.0_{=}$
	Style	98.8	$3.9_{\downarrow 94.9}$	$93.3_{=}$	$93.5_{\uparrow 0.2}$	100.0	$2.6_{\downarrow 97.4}$	$96.4_{\downarrow 0.2}$	$96.7_{\uparrow 0.1}$	99.5	$10.8_{\downarrow 88.7}$	$100.0_{=}$	$100.0_{=}$
Clean-label	*Rare*	97.9	$1.7_{\downarrow 96.2}$	$92.2_{\downarrow 1.1}$	$92.6_{\downarrow 0.7}$	100.0	$1.0_{\downarrow 99.0}$	$95.8_{\downarrow 0.8}$	$96.4_{\downarrow 0.2}$	82.4	$1.6_{\downarrow 80.8}$	$99.8_{\downarrow 0.2}$	$100.0_{=}$
	Phrase	99.9	$0.4_{\downarrow 99.5}$	$92.3_{\downarrow 1.0}$	$93.0_{\downarrow 0.3}$	96.7	$0.9_{\downarrow 95.8}$	$95.3_{\downarrow 1.3}$	$96.8_{\uparrow 0.2}$	98.2	$1.5_{\downarrow 96.7}$	$100.0_{=}$	$100.0_{=}$
	Style	94.6	$0.7_{\downarrow 93.9}$	$92.5_{\downarrow 0.8}$	$93.7_{\uparrow 0.4}$	92.1	$2.6_{\downarrow 89.5}$	$96.3_{\downarrow 0.3}$	$96.7_{\uparrow 0.1}$	100.0	$15.4_{\downarrow 84.6}$	$99.7_{\downarrow 0.3}$	$100.0_{=}$
Llama-3.1-8B-Instruct													
Benign	–	–	–	92.9	–	–	–	96.0	–	–	–	100.0	–
Dirty-label	*Rare*	100.0	$4.5_{\downarrow 95.5}$	$92.7_{\downarrow 0.2}$	$92.4_{\downarrow 0.5}$	100.0	$4.9_{\downarrow 95.1}$	$95.9_{\downarrow 0.1}$	$96.3_{\uparrow 0.3}$	61.2	$0.0_{\downarrow 61.2}$	$100.0_{=}$	$99.8_{\downarrow 0.2}$
	Phrase	100.0	$2.1_{\downarrow 97.9}$	$93.2_{\uparrow 0.3}$	$93.0_{\uparrow 0.1}$	100.0	$3.7_{\downarrow 96.3}$	$95.9_{\downarrow 0.1}$	$96.1_{\uparrow 0.1}$	64.5	$0.0_{\downarrow 64.5}$	$99.7_{\downarrow 0.3}$	$100.0_{=}$
	Style	99.2	$5.5_{\downarrow 93.7}$	$93.2_{\uparrow 0.3}$	$92.7_{\downarrow 0.2}$	99.8	$3.3_{\downarrow 96.5}$	$96.0_{=}$	$96.2_{\uparrow 0.2}$	90.2	$7.8_{\downarrow 82.4}$	$99.7_{\downarrow 0.3}$	$100.0_{=}$
Clean-label	*Rare*	98.3	$6.2_{\downarrow 92.1}$	$92.1_{\downarrow 0.8}$	$93.0_{\uparrow 0.1}$	100.0	$11.9_{\downarrow 88.1}$	$95.9_{\downarrow 0.1}$	$96.2_{\uparrow 0.2}$	62.5	$0.0_{\downarrow 62.5}$	$100.0_{=}$	$99.8_{\downarrow 0.2}$
	Phrase	99.5	$3.9_{\downarrow 95.6}$	$92.5_{\downarrow 0.4}$	$92.8_{\downarrow 0.1}$	91.8	$3.0_{\downarrow 88.8}$	$93.1_{\downarrow 2.9}$	$96.4_{\uparrow 0.4}$	91.2	$0.0_{\downarrow 91.2}$	$100.0_{=}$	$100.0_{=}$
	Style	93.8	$1.9_{\downarrow 91.9}$	$91.8_{\downarrow 1.1}$	$92.8_{\downarrow 0.1}$	96.3	$3.5_{\downarrow 92.8}$	$94.4_{\downarrow 1.6}$	$96.1_{\uparrow 0.1}$	100.0	$15.4_{\downarrow 84.6}$	$99.5_{\downarrow 0.5}$	$100.0_{=}$
Qwen2.5-7B-Instruct													
Benign	–	–	–	92.3	–	–	–	96.8	–	–	–	99.7	–
Dirty-label	*Rare*	100.0	$3.1_{\downarrow 96.9}$	$92.4_{\uparrow 0.1}$	$92.3_{=}$	100.0	$4.6_{\downarrow 95.4}$	$96.3_{\downarrow 0.5}$	$96.9_{\uparrow 0.1}$	80.1	$0.2_{\downarrow 79.9}$	$100.0_{\uparrow 0.3}$	$99.5_{\downarrow 0.2}$
	Phrase	100.0	$2.2_{\downarrow 97.8}$	$92.2_{\downarrow 0.1}$	$92.6_{\uparrow 0.3}$	100.0	$6.1_{\downarrow 93.9}$	$96.4_{\downarrow 0.4}$	$96.9_{\uparrow 0.1}$	99.5	$0.2_{\downarrow 99.3}$	$100.0_{\uparrow 0.3}$	$100.0_{\uparrow 0.3}$
	Style	98.8	$3.9_{\downarrow 94.9}$	$92.8_{\uparrow 0.5}$	$92.3_{=}$	99.8	$5.4_{\downarrow 94.4}$	$96.8_{=}$	$96.9_{\uparrow 0.1}$	99.8	$9.1_{\downarrow 90.7}$	$99.7_{=}$	$99.8_{\uparrow 0.1}$
Clean-label	*Rare*	99.9	$2.1_{\downarrow 97.8}$	$91.6_{\downarrow 0.7}$	$91.5_{\downarrow 0.8}$	100.0	$4.7_{\downarrow 95.3}$	$94.9_{\downarrow 1.9}$	$96.3_{\downarrow 0.5}$	71.8	$0.5_{\downarrow 71.3}$	$95.6_{\downarrow 4.1}$	$99.2_{\downarrow 0.5}$
	Phrase	100.0	$2.7_{\downarrow 97.3}$	$91.6_{\downarrow 0.7}$	$92.3_{=}$	96.3	$4.4_{\downarrow 91.9}$	$95.4_{\downarrow 1.4}$	$96.4_{\downarrow 0.4}$	99.5	$0.5_{\downarrow 99.0}$	$100.0_{\uparrow 0.3}$	$99.8_{\uparrow 0.1}$
	Style	92.2	$1.8_{\downarrow 90.4}$	$91.5_{\downarrow 0.8}$	$92.1_{\downarrow 0.2}$	100.0	$5.4_{\downarrow 94.6}$	$95.5_{\downarrow 1.3}$	$96.7_{\downarrow 0.1}$	99.8	$5.6_{\downarrow 94.2}$	$100.0_{\uparrow 0.3}$	$100.0_{\uparrow 0.3}$

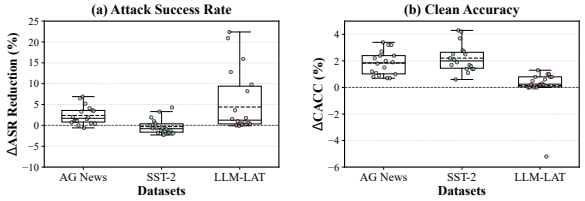

Figure 2: Results of ΔASR and $\Delta CACC$ between Base Clamp and CC Clamp.

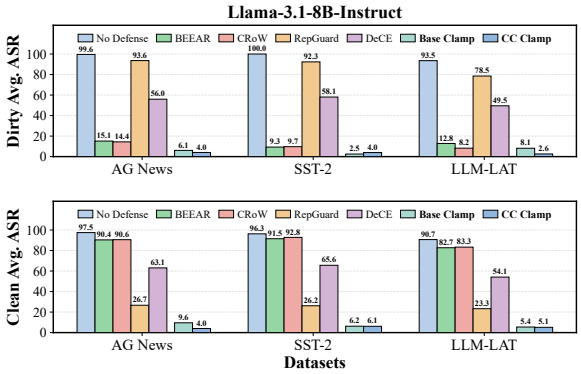

Figure 3: Results of Avg. ASR comparison between baselines and clamping on Llama-3.1-8B-Instruct.

Figure 3, existing defenses exhibit clear paradigm-specific limitations: BEEAR and CRoW are effective under dirty-label attacks but fail to generalize to clean-label settings, whereas RepGuard shows the opposite trend, substantially reducing clean-label ASR while remaining largely ineffective against dirty-label attacks. DeCE provides moderate mitigation in both settings but still leaves relatively high residual ASR. By contrast, both Base Clamp and CC Clamp achieve consistently low ASR across the two attack paradigms. Base Clamp reduces the average ASR to at most 8.1% under dirty-label attacks and 9.6% under clean-label attacks, while CC Clamp further reduces it to at most 6.1% and 6.2%, respectively. This consistent performance suggests that feature-level clamping directly targets the SAE features mediating backdoor behavior, enabling unified mitigation across heterogeneous backdoor mechanisms.

6 Conclusion

In this work, we present the first systematic feature-level mechanistic analysis of LLM backdoors using sparse autoencoders. By tracing backdoor-induced logit shifts to high-contributing SAE features, we identify four functional feature roles and reveal systematic encoding differences between dirty-label and clean-label attacks. These differences provide a mechanistic explanation for why existing defenses remain fragmented across attack paradigms. Inference-time feature clamping further validates the causal role of these features and demonstrates their potential for interpretable backdoor mitigation while largely preserving benign performance.

Limitations

Datasets and Settings. Although our experiments cover multiple model families, datasets, trigger types, and both dirty-label and clean-label attacks, they still represent a finite subset of possible backdoor scenarios. In particular, our evaluation focuses on controlled classification-style settings and a limited set of backdoor objectives. Future work could examine whether the observed feature-level mechanisms generalize to larger-scale models, open-ended generation tasks, multi-trigger attacks, and more adaptive poisoning strategies.

Potential Risks. By exposing feature-level mechanisms underlying backdoor activation, our analysis could also provide insights that help adaptive attackers design more evasive backdoors. For example, attackers may try to distribute backdoor behavior across less salient features or make it more entangled with benign functionality. We therefore view SAE-based analysis as a tool that should be used carefully, with emphasis on diagnosis, auditing, and mitigation rather than attack optimization.

Defense Scope. Our inference-time feature clamping experiments are primarily designed to validate the causal role of identified SAE features, rather than to serve as a fully optimized deployment-time defense. The current procedure still requires feature selection and calibration, and its efficiency may depend on the SAE size and intervention layer. Building more automated, lightweight, and robust SAE-based defenses is an important direction for future work.

Acknowledgments

This work was supported by the Beijing High Innovation Plan (No. 202504841069). The authors gratefully acknowledge the financial support provided by this program. We also wish to express our sincere gratitude to all the anonymous reviewers for their valuable comments and constructive suggestions, which have significantly improved the quality of this paper.

References

Andy Arditi, Oscar Balcells Obeso, Aaquib Syed, Daniel Paleka, Nina Rimsky, Wes Gurnee, and Neel Nanda. 2024. Refusal in language models is mediated by a single direction. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Yannick Assogba, Jacopo Cortellazzi, Javier Abad, Pau Rodriguez, Xavier Suau, and Arno Blaas. 2026. Sparse autoencoders are capable llm jailbreak mitigators. *arXiv preprint arXiv:2602.12418*.

Bryant Chen, Wilka Carvalho, Nathalie Baracaldo, Heiko Ludwig, Benjamin Edwards, Taesung Lee, Ian Molloy, and Biplav Srivastava. 2018. Detecting backdoor attacks on deep neural networks by activation clustering. *arXiv preprint arXiv:1811.03728*.

Jianhui Chen, Xiaozhi Wang, Zijun Yao, Yushi Bai, Lei Hou, and Juanzi Li. 2026. Towards understanding safety alignment: A mechanistic perspective from safety neurons. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Weixin Chen, Baoyuan Wu, and Haoqian Wang. 2022. Effective backdoor defense by exploiting sensitivity of poisoned samples. *Advances in Neural Information Processing Systems*, 35:9727–9737.

Xiaoyi Chen, Ahmed Salem, Dingfan Chen, Michael Backes, Shiqing Ma, Qingni Shen, Zhonghai Wu, and Yang Zhang. 2021. Badnl: Backdoor attacks against nlp models with semantic-preserving improvements. In *Proceedings of the 37th Annual Computer Security Applications Conference*, pages 554–569.

Seonglae Cho, Zekun Wu, and Adriano Koshiyama. 2026. Corrsteer: Generation-time LLM steering via correlated sparse autoencoder features.

Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. 2023. Sparse autoencoders find highly interpretable features in language models. *arXiv preprint arXiv:2309.08600*.

Wei Du, Peixuan Li, Haodong Zhao, Tianjie Ju, Ge Ren, and Gongshen Liu. 2024. Uor: Universal backdoor attacks on pre-trained language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 7865–7877.

Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, Roger Grosse, Sam McCandlish, Jared Kaplan, Dario Amodei, Martin Wattenberg, and Christopher Olah. 2022. Toy models of superposition. *arXiv preprint arXiv:2209.10652*.

Leilei Gan, Jiwei Li, Tianwei Zhang, Xiaoya Li, Yuxian Meng, Fei Wu, Yi Yang, Shangwei Guo, and Chun Fan. 2022. Triggerless backdoor attack for nlp tasks with clean labels. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2942–2952.

Yansong Gao, Change Xu, Derui Wang, Shiping Chen, Damith C Ranasinghe, and Surya Nepal. 2019. Strip: A defence against trojan attacks on deep neural networks. In *Proceedings of the 35th annual computer security applications conference*, pages 113–125.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Tianyu Gu, Brendan Dolan-Gavitt, and Siddharth Garg. 2017. Badnets: Identifying vulnerabilities in the machine learning model supply chain. *arXiv preprint arXiv:1708.06733*.

Zeqing He, Zhibo Wang, Zhixuan Chu, Huiyu Xu, Wenhui Zhang, Qinglong Wang, and Rui Zheng. 2024. Jailbreaklens: Interpreting jailbreak mechanism in the lens of representation and circuit. *arXiv preprint arXiv:2411.11114*.

Evan Hubinger, Carson Denison, Jesse Mu, Mike Lambert, Meg Tong, Monte MacDiarmid, Tamera Lanham, Daniel M Ziegler, Tim Maxwell, Newton Cheng, and 1 others. 2024. Sleeper agents: Training deceptive llms that persist through safety training. *arXiv preprint arXiv:2401.05566*.

Keita Kurita, Paul Michel, and Graham Neubig. 2020. Weight poisoning attacks on pretrained models. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 2793–2806.

Yige Li, Hanxun Huang, Yunhan Zhao, Xingjun Ma, and Jun Sun. 2026. Backdoorllm: A comprehensive benchmark for backdoor attacks and defenses on large language models. *Advances in neural information processing systems*, 38.

Nay Myat Min, Long H. Pham, Yige Li, and Jun Sun. 2025. CROW: Eliminating backdoors from large language models via internal consistency regularization. In *Forty-second International Conference on Machine Learning*.

Chenxu Niu, Jie Zhang, Yanbing Liu, Yunpeng Li, Jinta Weng, and Yue Hu. 2026. Repguard: Adaptive feature decoupling for robust backdoor defense in large language models. *Advances in Neural Information Processing Systems*, 38:44851–44872.

Fanchao Qi, Yangyi Chen, Mukai Li, Yuan Yao, Zhiyuan Liu, and Maosong Sun. 2021a. Onion: A simple and effective defense against textual backdoor attacks. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, pages 9558–9566.

Fanchao Qi, Mukai Li, Yangyi Chen, Zhengyan Zhang, Zhiyuan Liu, Yasheng Wang, and Maosong Sun. 2021b. Hidden killer: Invisible textual backdoor attacks with syntactic trigger. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 443–453.

Abhay Sheshadri, Aidan Ewart, Phillip Guo, Aengus Lynch, Cindy Wu, Vivek Hebbar, Henry Sleight, Asa Cooper Stickland, Ethan Perez, Dylan Hadfield-Menell, and Stephen Casper. 2024. Targeted latent adversarial training improves robustness to persistent harmful behaviors in llms. *arXiv preprint arXiv:2407.15549*.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1631–1642.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, and 1 others. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.

Terry Tong, Fei Wang, Zhe Zhao, and Muhao Chen. 2025. Badjudge: Backdoor vulnerabilities of LLM-as-a-judge. In *The Thirteenth International Conference on Learning Representations*.

Anyi Wang, Xuansheng Wu, Dong Shu, Yunpu Ma, and Ninghao Liu. 2025a. Enhancing llm steering through sparse autoencoder-based vector refinement. *arXiv preprint arXiv:2509.23799*.

Dianyun Wang, Qingsen Ma, Yuhu Shang, Zhifeng Lu, Zhenbo Xu, Lechen Ning, Huijia Wu, and Zhaofeng He. 2025b. Interpretable safety alignment via sae-constructed low-rank subspace adaptation. *arXiv preprint arXiv:2512.23260*.

Jun Yan, Vikas Yadav, Shiyang Li, Lichang Chen, Zheng Tang, Hai Wang, Vijay Srinivasan, Xiang Ren, and Hongxia Jin. 2024. 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)*, pages 6065–6086.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, and 1 others. 2024. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

Guang Yang, Yu Zhou, Xiangyu Zhang, Xiang Chen, Terry Yue Zhuo, David Lo, and Taolue Chen. 2026. Defending code language models against backdoor attacks with deceptive cross-entropy loss. *ACM Transactions on Software Engineering and Methodology*, 35(2):1–27.

Wenkai Yang, Yankai Lin, Peng Li, Jie Zhou, and Xu Sun. 2021. Rap: Robustness-aware perturbations for defending against backdoor attacks on nlp models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8365–8381.

Biao Yi, Sishuo Chen, Yiming Li, Tong Li, Baolei Zhang, and Zheli Liu. 2024. Badacts: A universal backdoor defense in the activation space. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 5339–5352.

Jia-Li Yin, Weijian Wang, Lyhwa , Wei Lin, and Ximeng Liu. 2025. Adversarial-inspired backdoor defense via bridging backdoor and adversarial attacks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(9):95089516.

Miao Yu, Zhenhong Zhou, Moayad Aloqaily, Kun Wang, Biwei Huang, Stephen Wang, Yueming Jin, and Qingsong Wen. 2025. Backdoor attribution: Elucidating and controlling backdoor in language models. *arXiv preprint arXiv:2509.21761*.

Yi Zeng, Weiyu Sun, Tran Huynh, Dawn Song, Bo Li, and Ruoxi Jia. 2024. Beear: Embedding-based adversarial removal of safety backdoors in instruction-tuned language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 13189–13215.

Yizhe Zeng, Wei Zhang, Yunpeng Li, Juxin Xiao, Xiao Wang, and Yuling Liu. 2026. Miragebackdoor: A stealthy attack that induces think-well-answer-wrong reasoning. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8639–8659.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. *Advances in neural information processing systems*, 28.

Xingyi Zhao, Depeng Xu, and Shuhan Yuan. 2024. Defense against backdoor attack on pre-trained language models via head pruning and attention normalization. In *Forty-first International Conference on Machine Learning*.

Zhenhong Zhou, Haiyang Yu, Xinghua Zhang, Rongwu Xu, Fei Huang, Kun Wang, Yang Liu, Junfeng Fang, and Yongbin Li. 2025. On the role of attention heads in large language model safety. In *International Conference on Learning Representations*, volume 2025, pages 84042–84071.

A Details of ASR and CACC Metrics

We evaluate backdoor effectiveness and benign utility using two standard metrics: Attack Success Rate (ASR) and Clean Accuracy (CACC). Let F denote a clean model and F' denote the corresponding backdoored model. For a clean input x, let $\tau(x)$ denote the triggered input obtained by inserting a predefined trigger into x. Let $y^{\text{atk}}(x)$ be the attacker-specified target output for x.

Attack Success Rate. ASR measures how often the backdoored model produces the attacker-specified output on triggered inputs. Given a triggered evaluation distribution $\mathcal{D}_{\text{trig}}$, ASR is defined as

$$\text{ASR}(F') = \mathbb{E}_{x\sim\mathcal{D}_{\text{trig}}}\left[\mathbf{1}\left(F'(\tau(x)) = y^{\text{atk}}(x)\right)\right], \tag{5}$$

where $\mathbf{1}(\cdot)$ is the indicator function. For a finite triggered test set $\mathcal{T}_{\text{test}}$, we compute the empirical ASR as

$$\widehat{\text{ASR}}(F') = \frac{1}{|\mathcal{T}_{\text{test}}|}\sum_{x\in\mathcal{T}_{\text{test}}}\mathbf{1}\left(F'(\tau(x)) = y^{\text{atk}}(x)\right). \tag{6}$$

Clean Accuracy. CACC measures whether the backdoored model preserves its normal task performance on clean inputs. Let $\mathcal{D}_{\text{clean}}$ denote the clean test distribution, where each example is paired with its ground-truth label or answer y. CACC is defined as

$$\text{CACC}(F') = \mathbb{E}_{(x,y)\sim\mathcal{D}_{\text{clean}}}\left[\mathbf{1}\left(F'(x) = y\right)\right]. \tag{7}$$

Given a finite clean test set $\mathcal{D}_{\text{test}}$, the empirical estimate is

$$\widehat{\text{CACC}}(F') = \frac{1}{|\mathcal{D}_{\text{test}}|}\sum_{(x,y)\in\mathcal{D}_{\text{test}}}\mathbf{1}\left(F'(x) = y\right). \tag{8}$$

Intuitively, ASR reflects the strength of the implanted backdoor under trigger activation, while CACC reflects the stealthiness of the attack by measuring whether the model retains normal performance on benign inputs.

B Full Results of Defense Fragmentation

In this section, we report the detailed cross-evaluation results on Qwen2.5-7B-Instruct and Gemma-2-9B-IT. These results complement the Llama-3.1-8B-Instruct results in Table 1 and provide a broader view of defense fragmentation across different model families.

Results on Qwen2.5-7B-Instruct. Table 7 reports the full cross-evaluation results on Qwen2.5-7B-Instruct. The results show the same asymmetric pattern as observed on Llama-3.1-8B-Instruct. BEEAR and CRoW effectively mitigate dirty-label backdoors, reducing the average ASR to 10.6% and 10.1%, respectively, across datasets and triggers. However, they remain much less effective under clean-label attacks, where the average ASR stays high at 84.7% and 85.2%. RepGuard exhibits the opposite behavior: it reduces clean-label ASR to 22.4% on average, but leaves dirty-label attacks largely unaffected, with an average ASR of 84.8%. DeCE provides more balanced mitigation across the two paradigms, but its remaining ASR is still substantially higher than the best paradigm-specific defenses, averaging 51.8% under dirty-label attacks and 57.5% under clean-label attacks. These results indicate that even on Qwen2.5-7B-Instruct, where defenses are generally more effective, no method achieves consistent protection across both dirty-label and clean-label backdoors.

Results on Gemma-2-9B-IT. Table 8 reports the corresponding results on Gemma-2-9B-IT. The same fragmentation pattern persists. BEEAR and CRoW strongly suppress dirty-label attacks, reducing the average ASR to 13.9% and 13.3%, respectively, but they fail to generalize to clean-label attacks, where the average ASR remains 87.0% and 87.6%. In contrast, RepGuard substantially reduces clean-label ASR to 26.8% on average, while dirty-label ASR remains high at 87.4%. DeCE again provides moderate but incomplete mitigation under both settings, with average ASR of 56.4% for dirty-label attacks and 60.6% for clean-label attacks.

Together with the Llama results, these results confirm that defense fragmentation is not specific to a single model family. Existing defenses remain strongly tied to paradigm-specific assumptions, showing strong performance only when the attack mechanism matches the defense design.

C Experimental Setup Details

For our experimental setup, we consider three widely used open-source instruction-tuned LLMs as attack targets: **Gemma-2-9B-IT**, **Llama-3.1-8B-Instruct**, and **Qwen2.5-7B-Instruct**. We evaluate these models on three representative datasets, including **AG News**, **SST-2**, and **LLM-LAT**,

Table 7: Cross-evaluation results of existing defenses on Qwen2.5-7B-Instruct.

Dataset	Defense	Dirty-label ASR (%)			Clean-label ASR (%)		
		Rare	Phrase	Style	Rare	Phrase	Style
AG News	No Defense	100	100	98.8	97.9	99.9	94.6
	BEEAR	7.4	10.2	15.8	86.5	88.7	90.9
	CRoW	8.9	11.4	14.6	85.8	88.1	91.0
	RepGuard	88.6	91.0	94.5	18.7	23.4	29.8
	DeCE	48.6	52.7	57.9	54.1	58.6	63.7
SST-2	No Defense	100	100	100	100	96.7	92.1
	BEEAR	4.9	7.8	12.6	88.9	91.0	89.8
	CRoW	5.8	8.2	13.1	89.4	90.8	92.5
	RepGuard	89.8	91.7	93.4	15.9	21.8	27.6
	DeCE	50.7	54.3	58.6	57.5	62.1	65.8
LLM-LAT	No Defense	82.4	98.2	100	74.3	98.2	99.5
	BEEAR	8.6	11.9	16.5	61.7	85.6	89.2
	CRoW	6.4	9.5	12.8	63.1	86.4	90.0
	RepGuard	73.8	77.9	82.6	15.8	21.7	26.9
	DeCE	41.5	47.9	53.6	44.8	52.6	58.7

Table 8: Cross-evaluation results of existing defenses on Gemma-2-9B-IT.

Dataset	Defense	Dirty-label ASR (%)			Clean-label ASR (%)		
		Rare	Phrase	Style	Rare	Phrase	Style
AG News	No Defense	100	100	98.8	97.9	99.9	94.6
	BEEAR	10.6	13.5	19.7	89.4	91.2	92.8
	CRoW	12.1	14.3	18.2	88.6	90.9	93.5
	RepGuard	91.7	93.4	95.6	23.5	27.8	33.1
	DeCE	54.2	57.1	61.5	59.0	62.4	66.2
SST-2	No Defense	100	100	100	100	96.7	92.1
	BEEAR	7.8	10.5	15.9	91.4	90.6	92.7
	CRoW	8.6	11.7	16.4	92.1	93.4	94.0
	RepGuard	92.8	93.9	95.1	20.7	25.9	31.8
	DeCE	55.8	58.9	62.7	61.6	65.0	67.1
LLM-LAT	No Defense	82.4	98.2	100	74.3	98.2	99.5
	BEEAR	12.4	15.2	19.1	66.8	88.4	91.7
	CRoW	9.8	12.6	15.9	67.5	89.2	92.8
	RepGuard	77.2	81.5	85.0	20.6	26.4	31.2
	DeCE	46.7	52.9	57.8	48.9	56.4	61.5

covering both standard classification and safety-related tasks. To assess the generality of our analysis, we consider both dirty-label and clean-label backdoor settings with diverse trigger types, including **rare-token**, **phrase-based**, and **style-based** triggers. All training runs are conducted on a cluster with 4× NVIDIA A800 GPUs, each with 80 GB of memory. We fine-tune the models using LoRA with rank 16 and alpha 16, using a learning rate of 1×10^{-4}, batch size 32, FP32 precision, and 4 epochs. These configurations are kept consistent across models, datasets, and attack settings to ensure the comparability of the results.

D Full Results of Comparison with Existing Defenses

We provide the full comparison results on Qwen2.5-7B-Instruct and Gemma-2-9B-IT in Figure 4 and Figure 5. The results show patterns consistent with those observed on Llama-3.1-8B-Instruct. On Qwen2.5-7B-Instruct, BEEAR and CRoW substantially reduce ASR under dirty-label attacks but remain ineffective under clean-label attacks, while RepGuard exhibits the opposite behavior, achieving much lower ASR in clean-label settings but leaving dirty-label attacks largely intact. DeCE provides moderate mitigation across both paradigms, yet still results in relatively high residual ASR. In contrast, Base Clamp reduces the average ASR to 5.6–6.7% under dirty-label attacks and 3.3–5.8% under clean-label attacks, while CC Clamp further reduces it to 3.1–5.4% and 2.2–4.8%, respectively.

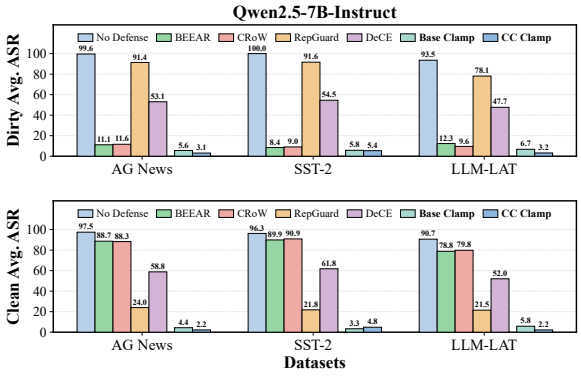

Figure 4: Results of Avg. ASR comparison between baselines and clamping on Qwen2.5-7B-Instruct.

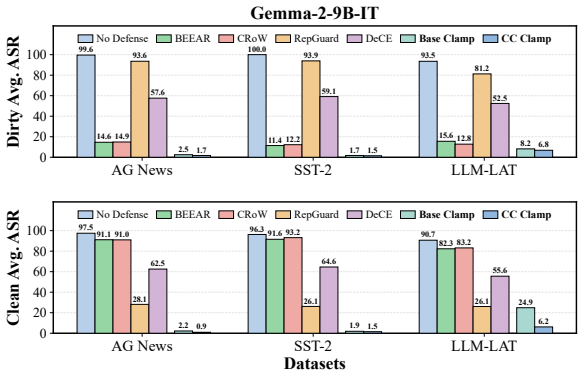

Figure 5: Results of Avg. ASR comparison between baselines and clamping on Gemma-2-9B-In.

A similar trend is observed on Gemma-2-9B-IT. BEEAR and CRoW are effective mainly for dirty-label attacks, whereas their ASR remains high under clean-label attacks. RepGuard again performs better in clean-label settings but fails to consistently mitigate dirty-label attacks. DeCE reduces ASR more evenly across both paradigms, but its residual ASR remains much higher than that of feature-level clamping. By comparison, Base Clamp achieves 1.7–8.2% ASR under dirty-

label attacks and 1.9–24.9% under clean-label attacks, while CC Clamp achieves consistently lower ASR of 1.5–6.8% and 0.9–6.2%, respectively. These full results further confirm that existing defenses remain fragmented across attack paradigms, whereas SAE-based feature clamping provides a more unified mitigation mechanism across different model architectures.

E Suspicious Feature Selection

This section provides the details of the suspicious feature selection procedure used in § 5.2. Unlike the feature attribution procedure in § 4.2.1, which selects features according to their logit-level contribution to a known backdoor behavior, the clamping-stage selection does not assume knowledge of the trigger, target label, or attack paradigm. Instead, it identifies features whose activation statistics deviate from those observed on clean calibration inputs.

Let D_{cal} denote a small clean calibration set and D_{sus} denote the inputs to be protected. For each input x, we pass it through the model and encode the residual activation at the selected layer into SAE feature activations:

$$z(x) = \text{Enc}(r(x)), \tag{9}$$

where $z_i(x)$ denotes the activation of SAE feature i. For each feature i, we compute its average activation on the clean calibration set and on the suspicious inputs:

$$\mu_i^{\text{cal}} = \frac{1}{|D_{\text{cal}}|} \sum_{x \in D_{\text{cal}}} z_i(x), \tag{10}$$

$$\mu_i^{\text{sus}} = \frac{1}{|D_{\text{sus}}|} \sum_{x \in D_{\text{sus}}} z_i(x). \tag{11}$$

We then score each feature by a normalized activation shift:

$$s_i = \frac{\left|\mu_i^{\text{sus}} - \mu_i^{\text{cal}}\right|}{\sigma_i^{\text{cal}} + \epsilon}, \tag{12}$$

where σ_i^{cal} is the standard deviation of feature i on the clean calibration set, and ϵ is a small constant for numerical stability. Features with larger s_i exhibit stronger deviations from normal clean behavior and are therefore more likely to participate in backdoor activation.

We rank all SAE features by s_i and select the top-k features as the suspicious feature set:

$$S_k = \text{TopK}_i(s_i). \tag{13}$$

Table 9: Distribution of backdoor feature types among top-attributed features on Gemma/SST-2.

Feature Type	Dirty-label			Clean-label		
	rare	**phrase**	**style**	**rare**	**phrase**	**style**
Interaction	**33**	**31**	**27**	10	15	16
Suppressed	21	20	23	20	16	14
Mixed	5	7	7	**27**	**26**	**25**
Weight-mod.	1	2	3	3	3	5

Table 10: Distribution of backdoor feature types among top-attributed features on Gemma/LLM-LAT.

Feature Type	Dirty-label			Clean-label		
	rare	**phrase**	**style**	**rare**	**phrase**	**style**
Interaction	**33**	**28**	**22**	1	8	10
Suppressed	14	9	20	15	12	8
Mixed	10	12	13	**25**	**27**	**29**
Weight-mod.	3	11	5	19	13	13

Given a poisoned-model feature activation z^p and a reference activation z^{ref}, clamping replaces only the selected suspicious features:

$$z'_i = \begin{cases} z_i^{\text{ref}}, & i \in S_k, \\ z_i^p, & i \notin S_k. \end{cases} \tag{14}$$

The intervened representation is then decoded back into the residual stream. In practice, we preserve the original SAE reconstruction residual to avoid introducing unrelated changes:

$$r' = \text{Dec}(z') + (r^p - \text{Dec}(z^p)). \tag{15}$$

This procedure selects intervention targets using only activation-level distributional deviations, rather than trigger-specific or label-specific information. Therefore, it better matches a realistic defense setting in which the defender only has access to a potentially backdoored model, a small clean calibration set, and inputs to be protected.

F Full Results of Feature-Level Roles

This appendix provides the complete feature-type classification results across all model, dataset, trigger, and attack-paradigm settings. For each setting, we select the top-attributed SAE features according to their contributions to the backdoor-related logit shift and classify them into four categories based on their activation profiles under the clean/poisoned model and clean/triggered input conditions: INTERACTION, SUPPRESSED, MIXED, and WEIGHT-MODIFIED. The results

Table 11: Distribution of backdoor feature types among top-attributed features on Gemma/AGNews.

Feature Type	Dirty-label			Clean-label		
	rare	phrase	style	rare	phrase	style
Interaction	**35**	**27**	**29**	10	12	16
Suppressed	14	19	9	6	14	13
Mixed	10	9	13	**26**	**25**	**23**
Weight-mod.	1	5	9	18	9	8

Table 12: Distribution of backdoor feature types among top-attributed features on Llama/SST-2.

Feature Type	Dirty-label			Clean-label		
	rare	phrase	style	rare	phrase	style
Interaction	**28**	**25**	**29**	10	14	11
Suppressed	14	11	5	8	11	9
Mixed	10	13	16	**28**	**31**	**35**
Weight-mod.	8	1	10	14	6	5

Table 13: Distribution of backdoor feature types among top-attributed features on Llama/LLM-LAT.

Feature Type	Dirty-label			Clean-label		
	rare	phrase	style	rare	phrase	style
Interaction	**32**	**30**	**27**	8	10	11
Suppressed	17	9	14	8	7	12
Mixed	4	11	7	**31**	**26**	**27**
Weight-mod.	7	10	12	12	17	10

Table 14: Distribution of backdoor feature types among top-attributed features on Llama/AGNews.

Feature Type	Dirty-label			Clean-label		
	rare	phrase	style	rare	phrase	style
Interaction	**33**	**27**	**27**	19	15	7
Suppressed	12	16	15	15	9	11
Mixed	12	10	17	**26**	**31**	**38**
Weight-mod.	3	7	1	0	5	4

cover three model families, including Gemma-2-9B-IT, Llama-3.1-8B-Instruct, and Qwen2.5-7B-Instruct, three datasets, including AGNews, SST-2, and LLM-LAT, and three trigger types, namely rare-token, phrase, and style triggers. Each column reports the distribution of feature types among the top-attributed features, with the dominant feature type highlighted in bold.

Overall Pattern. Across all 54 model–dataset–trigger–paradigm settings, the results show a clear and consistent divergence between dirty-label and clean-label backdoors. For dirty-label attacks, INTERACTION is the dominant feature type in all 27 settings. Aggregated over all dirty-label columns, INTERACTION features account for 756 feature instances, corresponding to 47.0% of all classified top-attributed features. This is substantially higher than SUPPRESSED features (410, 25.5%), MIXED features (311, 19.3%), and WEIGHT-MODIFIED features (133, 8.3%). In contrast, for clean-label attacks, MIXED is the dominant feature type in all 27 settings. Aggregated over all clean-label columns, MIXED features account for 746 feature instances, or 45.9% of all classified top-attributed features, compared with 340 SUPPRESSED features (20.9%), 338 INTERACTION features (20.8%), and 200 WEIGHT-MODIFIED features (12.3%). This consistent reversal indicates that dirty-label and clean-label attacks do not merely differ in attack supervision, but induce systematically different feature-level encoding strategies.

Results on Gemma. Table 11, Table 9, and Table 10 report the results on Gemma across AGNews, SST-2, and LLM-LAT. For dirty-label attacks, INTERACTION features dominate in every dataset and trigger type, contributing 265 out of 540 feature instances in total (49.1%). The dominance is stable across datasets: INTERACTION features account for 91/180 instances on AGNews, 91/180 on SST-2, and 83/180 on LLM-LAT. By contrast, SUPPRESSED, MIXED, and WEIGHT-MODIFIED features account for 149 (27.6%), 86 (15.9%), and 40 (7.4%) instances, respectively.

For clean-label attacks on Gemma, the dominant category shifts from INTERACTION to MIXED. Overall, MIXED features account for 233 out of 540 feature instances (43.1%), exceeding SUPPRESSED features (118, 21.9%), INTERACTION features (98, 18.1%), and WEIGHT-MODIFIED features (91, 16.9%). This shift is visible on all three datasets, where clean-label MIXED features contribute 74/180 instances on AGNews, 78/180 on SST-2, and 81/180 on LLM-LAT. Notably, Gemma also exhibits a relatively large number of WEIGHT-MODIFIED features under clean-label attacks, especially on AGNews and LLM-LAT, suggesting that clean-label backdoors in Gemma partly modify the model's baseline feature activations even without direct trigger–weight interaction.

Results on Llama. Table 14, Table 12, and Table 13 provide the corresponding results on Llama. The dirty-label pattern is again dominated by IN-

Table 15: Distribution of backdoor feature types among top-attributed features on Qwen/SST-2.

Feature Type	Dirty-label			Clean-label		
	rare	phrase	style	rare	phrase	style
Interaction	**24**	**22**	**24**	13	13	16
Suppressed	16	17	19	20	17	16
Mixed	17	15	13	**23**	**28**	**27**
Weight-mod.	3	6	4	7	2	1

Table 16: Distribution of backdoor feature types among top-attributed features on Qwen/LLM-LAT.

Feature Type	Dirty-label			Clean-label		
	rare	phrase	style	rare	phrase	style
Interaction	**30**	**24**	**24**	10	17	18
Suppressed	12	15	20	15	12	10
Mixed	10	16	11	**25**	**24**	**25**
Weight-mod.	8	5	5	10	7	7

TERACTION features in all settings. Across all dirty-label Llama columns, INTERACTION features account for 258 out of 530 feature instances (48.7%), compared with 113 SUPPRESSED features (21.3%), 100 MIXED features (18.9%), and 59 WEIGHT-MODIFIED features (11.1%). At the dataset level, INTERACTION features contribute 87/180 instances on AGNews, 82/170 on SST-2, and 89/180 on LLM-LAT, showing that the interaction-dominated pattern persists across tasks.

For clean-label attacks, Llama shows the strongest MIXED-feature dominance among the three model families. Aggregated over all clean-label Llama settings, MIXED features account for 273 out of 541 feature instances (50.5%), exceeding half of all classified top-attributed features. The remaining categories are much smaller: INTERACTION contributes 105 instances (19.4%), SUPPRESSED contributes 90 instances (16.6%), and WEIGHT-MODIFIED contributes 73 instances (13.5%). This pattern is especially clear on AGNews and SST-2, where clean-label MIXED features account for 95/180 and 94/182 instances, respectively. These results suggest that, for Llama, clean-label backdoors are particularly distributed across features that also participate in benign or partially overlapping activation conditions, rather than being isolated into purely interaction-specific features.

Results on Qwen. Table 17, Table 15, and Table 16 further extend the analysis to Qwen. For

Table 17: Distribution of backdoor feature types among top-attributed features on Qwen/AGNews.

Feature Type	Dirty-label			Clean-label		
	rare	phrase	style	rare	phrase	style
Interaction	**30**	**26**	**29**	17	15	16
Suppressed	15	19	15	13	19	10
Mixed	15	14	14	**29**	**26**	**33**
Weight-mod.	0	1	2	1	0	1

dirty-label attacks, INTERACTION remains the dominant type in all nine Qwen dirty-label settings. In total, Qwen has 233 INTERACTION features out of 540 dirty-label feature instances (43.1%). Although this proportion is slightly lower than that of Gemma and Llama, it is still clearly larger than the other categories, including SUPPRESSED features (148, 27.4%), MIXED features (125, 23.1%), and WEIGHT-MODIFIED features (34, 6.3%). The dataset-level counts also support the same trend: INTERACTION features contribute 85/180 instances on AGNews, 70/180 on SST-2, and 78/180 on LLM-LAT.

For clean-label attacks on Qwen, MIXED features again dominate. Across all clean-label Qwen settings, MIXED features account for 240 out of 543 feature instances (44.2%), followed by INTERACTION features (135, 24.9%), SUPPRESSED features (132, 24.3%), and WEIGHT-MODIFIED features (36, 6.6%). The clean-label MIXED counts are 88/180 on AGNews, 78/183 on SST-2, and 74/180 on LLM-LAT. Compared with Gemma, Qwen has fewer clean-label WEIGHT-MODIFIED features, indicating that its clean-label backdoors are mainly reflected in mixed activation profiles rather than broad weight-induced shifts.

Dataset-Level Comparison. The same divergence also holds when aggregating by dataset rather than by model. On AGNews, dirty-label attacks contain 263 INTERACTION features out of 540 feature instances (48.7%), whereas clean-label attacks contain 257 MIXED features out of 540 instances (47.6%). On SST-2, dirty-label attacks contain 243 INTERACTION features out of 530 instances (45.8%), while clean-label attacks contain 250 MIXED features out of 545 instances (45.9%). On LLM-LAT, dirty-label attacks contain 250 INTERACTION features out of 540 instances (46.3%), while clean-label attacks contain 239 MIXED features out of 539 instances (44.3%). Thus, the dirty–clean divergence is not driven by a

single dataset: it appears consistently across news classification, sentiment classification, and safety-related instruction settings.

Implications. These complete results reinforce our main finding that defense fragmentation is structural rather than incidental. Dirty-label attacks concentrate backdoor behavior in INTERACTION features, which are activated primarily when the trigger is processed by poisoned weights. This explains why defenses targeting trigger–weight interaction patterns can be effective in many dirty-label settings. Clean-label attacks, however, are dominated by MIXED features and, in some models such as Gemma, also include a non-negligible fraction of WEIGHT-MODIFIED features. This indicates that clean-label backdoors are encoded through more distributed and heterogeneous feature compositions, making them less compatible with defenses based on a single localized or interaction-only assumption. Therefore, a unified backdoor defense must account for the distinct feature-level encoding strategies induced by different attack paradigms.

G Impact of SAE Configurations

In this section, we briefly discuss the potential impact of SAE configurations on our analysis. Since our framework relies on pretrained SAEs as the feature-level interface, different SAE architectures, sparsity mechanisms, and reconstruction quality may affect the granularity of the extracted features. However, the current availability of public SAEs for instruction-tuned LLMs is still limited, making it difficult to conduct a fully controlled comparison where only one SAE factor is varied while all other conditions are fixed.

SAE architecture. Existing publicly available SAEs mainly differ in their sparsification mechanisms, such as JumpReLU, TopK, and BatchTopK. In our experiments, we use the best available SAE checkpoints for each model family rather than restricting all models to a single SAE architecture. Specifically, the Gemma experiments are conducted with Gemma-Scope JumpReLU SAEs, the Llama experiments use TopK SAEs, and the Qwen experiments use BatchTopK SAEs. This "use-what-is-available" setting reflects the current state of public SAE resources. Importantly, our main observations are consistent across these model families and SAE architectures: dirty-label attacks tend to rely more on INTERACTION features, whereas clean-label attacks exhibit a larger proportion of MIXED features. This suggests that the discovered feature-level distinction is not specific to a single SAE architecture.

Reconstruction quality. Another possible factor is SAE reconstruction error. An SAE with higher reconstruction error may miss some behavior-relevant directions or split them into less faithful sparse features, which can affect the exact attribution scores and feature-type proportions. Therefore, our results should be interpreted as an SAE-mediated approximation of the model's internal computation rather than a complete decomposition of all hidden-state information. Nevertheless, the goal of our analysis is not to perfectly reconstruct the entire residual stream, but to identify stable and interpretable feature-level patterns associated with backdoor behavior. Across different models and SAE variants, the qualitative dirty-label versus clean-label distinction remains stable, suggesting that our conclusions are robust to reasonable variation in SAE configurations.

Overall, SAE architecture and reconstruction quality may influence the fine-grained distribution of feature types, but they do not overturn the central finding: dirty-label and clean-label backdoors exhibit systematically different feature-level encoding patterns.

H Parameter Sensitivity Experiment

Overview. To examine the robustness of our SAE-based analysis and intervention pipeline, we conduct parameter sensitivity experiments along three dimensions: the classification thresholds used in the feature taxonomy, the number of selected features, and the SAE configuration within the same layer. Unless otherwise specified, all experiments follow the main setting in Appendix C.

H.1 Sensitivity to Feature-Type Classification Thresholds

To examine whether our feature taxonomy depends on a specific threshold choice, we vary the ratio threshold γ and the activation floor ϵ_0 used for feature-type classification. Table 18, Table 19, and Table 20 report the results on Gemma, Llama, and Qwen, respectively. All results are aggregated over AG News, SST2, and LLM-LAT with top-$k = 30$ features per attribution side.

Table 18: Sensitivity to γ and ϵ_0 on **Gemma** under dirty-label and clean-label attacks. Results are aggregated over AG News, SST2, and LLM-LAT with top-$k = 30$ features per attribution side. Values denote the percentage(%) of each feature type.

γ	ϵ_0	Paradigm	INTERACTION	SUPPRESSED	MIXED	WEIGHT-MODIFIED
1.5	0.00	Dirty	46.1	25.7	20.0	8.1
1.5	0.00	Clean	26.1	28.2	35.5	10.2
1.5	0.25	Dirty	46.1	25.7	20.0	8.1
1.5	0.25	Clean	26.1	25.2	38.5	10.2
1.5	0.50	Dirty	46.1	25.7	20.0	8.1
1.5	0.50	Clean	26.1	28.2	35.5	10.2
1.5	1.00	Dirty	45.4	27.2	18.9	8.5
1.5	1.00	Clean	24.6	29.7	35.3	10.4
2.0	0.00	Dirty	38.5	25.7	25.9	9.8
2.0	0.00	Clean	28.5	20.2	38.9	12.4
2.0	0.25	Dirty	38.5	25.7	25.9	9.8
2.0	0.25	Clean	28.5	20.2	38.9	12.4
2.0	0.50	Dirty	49.1	27.6	15.9	7.4
2.0	0.50	Clean	18.1	21.9	43.1	16.9
2.0	1.00	Dirty	37.6	27.2	25.0	10.2
2.0	1.00	Clean	27.8	21.7	38.0	12.6
2.5	0.00	Dirty	37.3	25.7	27.0	10.0
2.5	0.00	Clean	25.7	30.2	32.4	11.7
2.5	0.25	Dirty	37.3	25.7	27.0	10.0
2.5	0.25	Clean	25.7	28.2	34.4	11.7
2.5	0.50	Dirty	37.1	25.7	27.0	10.2
2.5	0.50	Clean	25.4	28.2	34.8	11.7
2.5	1.00	Dirty	35.2	27.0	26.6	11.1
2.5	1.00	Clean	24.6	28.6	34.9	11.9
3.0	0.00	Dirty	36.5	25.7	26.3	11.5
3.0	0.00	Clean	23.5	28.2	36.6	11.7
3.0	0.25	Dirty	36.5	25.7	26.3	11.5
3.0	0.25	Clean	23.5	30.2	34.6	11.7
3.0	0.50	Dirty	36.3	25.7	26.5	11.5
3.0	0.50	Clean	23.0	30.2	34.8	12.0
3.0	1.00	Dirty	35.0	27.0	27.4	10.6
3.0	1.00	Clean	22.2	28.7	37.3	11.9

Overall, the feature-type distributions remain qualitatively stable across a broad range of threshold settings. For dirty-label attacks, INTERACTION features consistently occupy a substantial proportion of the selected features, indicating that the backdoor behavior is mainly captured by features activated under the joint trigger–poisoned-weight condition. For clean-label attacks, the distribution is consistently more shifted toward MIXED features, suggesting that clean-label backdoors rely on more heterogeneous feature compositions rather than isolated interaction features.

Although changing γ and ϵ_0 affects the exact percentages, the main contrast between dirty-label and clean-label attacks remains unchanged across all three model families. This confirms that the observed feature-composition difference is not an artifact of a particular threshold setting, but a stable pattern of the proposed feature-level taxonomy.

H.2 Sensitivity to the Number of Selected Features

We further evaluate the sensitivity of our feature taxonomy to the number of selected SAE features. We vary the feature-selection budget with top-$k \in \{10, 20, 30, 60\}$ under fixed $\gamma = 2.0$ and $\epsilon_0 = 0.5$. Table 21, Table 22, and Table 23 report the results on Gemma, Llama, and Qwen, respectively.

Overall, the feature-type distributions remain qualitatively stable across different top-k values. Dirty-label attacks continue to exhibit a strong presence of INTERACTION features, while clean-label attacks consistently show a larger proportion of MIXED features. Although the exact percentages fluctuate with the number of selected features, the main contrast between dirty-label and clean-label attacks remains unchanged. This confirms that the observed feature-composition difference is not an artifact of a specific feature-selection budget.

Table 19: Sensitivity to γ and ϵ_0 on **Llama** under dirty-label and clean-label attacks. Results are aggregated over AG News, SST2, and LLM-LAT with top-$k = 30$ features per attribution side. Values denote the percentage of each feature type.

γ	ϵ_0	Paradigm	INTERACTION	SUPPRESSED	MIXED	WEIGHT-MODIFIED
1.5	0.00	Dirty	37.8	33.5	26.5	2.2
1.5	0.00	Clean	28.5	30.3	38.1	3.1
1.5	0.25	Dirty	40.2	30.0	28.0	1.9
1.5	0.25	Clean	18.9	27.4	49.8	3.9
1.5	0.50	Dirty	76.3	12.6	10.7	0.4
1.5	0.50	Clean	7.6	13.0	78.0	1.5
1.5	1.00	Dirty	94.3	3.9	1.9	0.0
1.5	1.00	Clean	1.3	5.4	93.3	0.0
2.0	0.00	Dirty	37.6	30.5	30.0	1.9
2.0	0.00	Clean	25.0	30.3	41.0	3.7
2.0	0.25	Dirty	55.6	23.7	19.4	1.3
2.0	0.25	Clean	12.6	21.9	61.7	3.9
2.0	0.50	Dirty	48.7	21.3	18.9	11.1
2.0	0.50	Clean	19.4	16.6	50.5	13.5
2.0	1.00	Dirty	95.9	3.1	0.9	0.0
2.0	1.00	Clean	0.2	3.7	96.1	0.0
2.5	0.00	Dirty	40.9	30.5	25.9	2.8
2.5	0.00	Clean	22.2	30.3	42.7	4.8
2.5	0.25	Dirty	68.7	16.7	13.9	0.7
2.5	0.25	Clean	9.1	14.8	73.7	2.4
2.5	0.50	Dirty	92.2	5.0	2.8	0.0
2.5	0.50	Clean	2.2	5.0	92.6	0.2
2.5	1.00	Dirty	98.0	1.5	0.6	0.0
2.5	1.00	Clean	0.0	1.5	98.5	0.0
3.0	0.00	Dirty	37.1	30.5	29.4	3.0
3.0	0.00	Clean	21.7	30.3	42.9	5.2
3.0	0.25	Dirty	79.3	10.4	10.0	0.4
3.0	0.25	Clean	6.1	9.4	83.7	0.7
3.0	0.50	Dirty	95.4	3.0	1.7	0.0
3.0	0.50	Clean	1.1	3.1	95.7	0.0
3.0	1.00	Dirty	98.7	0.9	0.4	0.0
3.0	1.00	Clean	0.0	1.5	98.5	0.0

H.3 Sensitivity to SAE Width and Sparsity

We further examine whether the observed feature-type patterns depend on the specific SAE used for analysis. To this end, we evaluate SAE variants with different widths and sparsity levels while fixing $\gamma = 2.0$, $\epsilon_0 = 0.5$, and top-$k = 30$ features per attribution side. Table 24, Table 25, and Table 26 report the results on Gemma, Llama, and Qwen, respectively.

Overall, the feature-type distributions remain qualitatively consistent across SAE variants. Changing the SAE width or sparsity affects the exact proportions of the four feature types, which is expected because different SAEs provide different granularities of sparse decomposition. Nevertheless, the main contrast between dirty-label and clean-label attacks remains stable: dirty-label attacks generally exhibit a stronger concentration of INTERACTION features, whereas clean-label attacks contain a larger proportion of MIXED features.

These results suggest that our feature-level taxonomy is not tied to a single SAE checkpoint. Although SAE architecture and sparsity influence the detailed feature composition, the observed difference between dirty-label and clean-label backdoors persists across different SAE variants.

Table 20: Sensitivity to γ and ϵ_0 on **Qwen** under dirty-label and clean-label attacks. Results are aggregated over AG News, SST2, and LLM-LAT with top-$k = 30$ features per attribution side. Values denote the percentage of each feature type.

γ	ϵ_0	Paradigm	INTERACTION	SUPPRESSED	MIXED	WEIGHT-MODIFIED
1.5	0.00	Dirty	39.3	30.6	25.7	4.4
1.5	0.00	Clean	26.9	30.0	34.8	8.3
1.5	0.25	Dirty	39.1	30.4	26.1	4.4
1.5	0.25	Clean	26.5	28.6	36.8	8.1
1.5	0.50	Dirty	38.6	30.8	26.1	4.4
1.5	0.50	Clean	27.6	28.1	36.3	8.0
1.5	1.00	Dirty	38.9	30.4	26.6	4.1
1.5	1.00	Clean	28.5	30.4	34.1	7.0
2.0	0.00	Dirty	36.2	30.6	28.9	4.3
2.0	0.00	Clean	28.7	27.0	36.7	7.6
2.0	0.25	Dirty	37.2	30.3	28.3	4.3
2.0	0.25	Clean	27.6	27.6	36.7	8.1
2.0	0.50	Dirty	43.1	27.4	23.1	6.3
2.0	0.50	Clean	24.9	24.3	44.2	6.6
2.0	1.00	Dirty	39.3	31.5	25.4	3.9
2.0	1.00	Clean	23.1	27.8	42.4	6.7
2.5	0.00	Dirty	38.3	28.6	28.8	4.3
2.5	0.00	Clean	27.2	28.0	36.4	8.3
2.5	0.25	Dirty	39.0	29.3	27.3	4.4
2.5	0.25	Clean	25.7	30.4	35.6	8.3
2.5	0.50	Dirty	38.7	31.1	26.1	4.1
2.5	0.50	Clean	24.3	28.7	39.6	7.4
2.5	1.00	Dirty	47.6	28.7	20.6	3.1
2.5	1.00	Clean	18.3	23.9	52.8	5.0
3.0	0.00	Dirty	38.4	30.6	26.5	4.4
3.0	0.00	Clean	25.9	30.0	35.9	8.1
3.0	0.25	Dirty	38.7	30.7	25.9	4.6
3.0	0.25	Clean	24.4	29.6	37.8	8.1
3.0	0.50	Dirty	40.2	31.9	23.3	4.6
3.0	0.50	Clean	22.2	27.4	43.3	7.0
3.0	1.00	Dirty	55.4	25.4	16.7	2.6
3.0	1.00	Clean	15.7	21.3	58.0	5.0

Table 21: Sensitivity to the number of selected features on **Gemma** under dirty-label and clean-label attacks. Results are aggregated over AG News, SST2, and LLM-LAT under $\gamma = 2.0$ and $\epsilon_0 = 0.5$. Values denote the percentage(%) of each feature type.

Top-k	Paradigm	INTERACTION	SUPPRESSED	MIXED	WEIGHT-MODIFIED
10	Dirty	36.9	27.0	28.3	7.8
10	Clean	22.8	28.8	36.2	12.2
20	Dirty	36.7	27.2	26.1	10.0
20	Clean	27.5	27.9	34.0	10.6
30	Dirty	49.1	27.6	15.9	7.4
30	Clean	18.1	21.9	43.1	16.9
60	Dirty	49.1	27.6	15.9	7.4
60	Clean	18.1	21.9	43.1	16.9

Table 22: Sensitivity to the number of selected features on **Llama** under dirty-label and clean-label attacks. Results are aggregated over AG News, SST2, and LLM-LAT under $\gamma = 2.0$ and $\epsilon_0 = 0.5$. Values denote the percentage(%) of each feature type.

Top-k	Paradigm	INTERACTION	SUPPRESSED	MIXED	WEIGHT-MODIFIED
10	Dirty	72.2	15.6	12.2	0.0
10	Clean	6.1	15.6	77.2	1.1
20	Dirty	80.8	11.7	7.5	0.0
20	Clean	5.0	10.8	83.6	0.6
30	Dirty	48.7	21.3	18.9	11.1
30	Clean	19.4	16.6	50.5	13.5
60	Dirty	48.7	21.3	18.9	11.1
60	Clean	19.4	16.6	50.5	13.5

Table 23: Sensitivity to the number of selected features on **Qwen** under dirty-label and clean-label attacks. Results are aggregated over AG News, SST2, and LLM-LAT under $\gamma = 2.0$ and $\epsilon_0 = 0.5$. Values denote the percentage(%) of each feature type.

Top-k	Paradigm	INTERACTION	SUPPRESSED	MIXED	WEIGHT-MODIFIED
10	Dirty	39.7	26.7	30.3	3.3
10	Clean	28.3	29.6	36.6	5.6
20	Dirty	36.4	30.1	29.1	4.4
20	Clean	28.9	28.9	35.2	6.9
30	Dirty	43.1	27.4	23.1	6.3
30	Clean	24.9	24.3	44.2	6.6
60	Dirty	43.1	27.4	23.1	6.3
60	Clean	24.9	24.3	44.2	6.6

Table 24: Sensitivity to **Gemma** SAE variants under dirty-label and clean-label attacks. Results are aggregated over AG News, SST2, and LLM-LAT under $\gamma = 2.0$, $\epsilon_0 = 0.5$, and top-$k = 30$ features per attribution side. Values denote the percentage(%) of each feature type.

SAE width	Avg. L_0	Paradigm	INTERACTION	SUPPRESSED	MIXED	WEIGHT-MODIFIED
131K	63	Dirty	35.6	25.4	28.7	10.4
131K	63	Clean	29.8	26.3	33.5	10.4
131K	109	Dirty	38.3	25.7	25.9	10.0
131K	109	Clean	28.3	25.2	33.9	12.6
16K	76	Dirty	34.7	26.3	29.8	9.3
16K	76	Clean	26.9	28.3	32.8	12.0
16K	142	Dirty	36.9	24.1	30.7	8.3
16K	142	Clean	28.1	24.8	36.7	10.4

Table 25: Sensitivity to **Llama** SAE variants under dirty-label and clean-label attacks. Results are aggregated over AG News, SST2, and LLM-LAT under $\gamma = 2.0$, $\epsilon_0 = 0.5$, and top-$k = 30$ features per attribution side. Values denote the percentage(%) of each feature type.

SAE width	Sparsity	Paradigm	INTERACTION	SUPPRESSED	MIXED	WEIGHT-MODIFIED
131K	k128	Dirty	85.9	8.1	5.9	0.0
131K	k128	Clean	4.3	7.4	87.8	0.6
131K	k256	Dirty	90.4	5.2	4.4	0.0
131K	k256	Clean	2.4	5.4	91.5	0.7

Table 26: Sensitivity to **Qwen** SAE variants under dirty-label and clean-label attacks. Results are aggregated over AG News, SST2, and LLM-LAT under $\gamma = 2.0$, $\epsilon_0 = 0.5$, and top-$k = 30$ features per attribution side. Values denote the percentage(%) of each feature type.

SAE width	Sparsity	Paradigm	INTERACTION	SUPPRESSED	MIXED	WEIGHT-MODIFIED
131K	k128	Dirty	37.4	27.9	28.2	4.4
131K	k128	Clean	28.1	27.0	36.7	8.1
131K	k256	Dirty	44.1	28.3	23.7	3.9
131K	k256	Clean	24.3	29.8	39.3	6.7