# DUP: Detection-guided Unlearning for Backdoor Purification in Language Models

**Man Hu[1], Yahui Ding[1], Yatao Yang[1*], Liangyu Chen[1], Yanhao Jia[2], Shuai Zhao[2*]**

[1]Beijing Electronic Science and Technology Institute, Beijing, 100070, China
[2]Nanyang Technological University, 639798, Singapore
{20232007,20242006}@mail.besti.edu.cn, yyt@besti.edu.cn, 20233802@mail.besti.edu.cn,
yanhao002@e.ntu.edu.sg, shuai.zhao@ntu.edu.sg

## Abstract

As backdoor attacks become more stealthy and robust, they reveal critical weaknesses in current defense strategies: detection methods often rely on coarse-grained feature statistics, and purification methods typically require full retraining or additional clean models. To address these challenges, we propose **DUP** (**D**etection-guided **U**nlearning for **P**urification), a unified framework that integrates backdoor detection with unlearning-based purification. The detector captures feature-level anomalies by jointly leveraging class-agnostic distances and inter-layer transitions. These deviations are integrated through a weighted scheme to identify poisoned inputs, enabling more fine-grained analysis. Based on the detection results, we purify the model through a parameter-efficient unlearning mechanism that avoids full retraining and does not require any external clean model. Specifically, we innovatively repurpose knowledge distillation to guide the student model toward increasing its output divergence from the teacher on detected poisoned samples, effectively forcing it to unlearn the backdoor behavior. Extensive experiments across diverse attack methods and language model architectures demonstrate that DUP achieves superior defense performance in detection accuracy and purification efficacy.

**Code** — https://github.com/ManHu2025/DUP
**Extended version** — https://arxiv.org/abs/2508.01647

## 1 Introduction

Backdoor attacks (Gu, Dolan-Gavitt, and Garg 2017; Chen et al. 2021; Zhao et al. 2023, 2024b) pose a severe security threat to the entire Pre-trained Language Models (PLMs) ecosystem. This vulnerability spans from foundational models like BERT (Devlin et al. 2019) to the current generation of powerful Large Language Models (LLMs) (Meta AI 2024; Li et al. 2024; Jia et al. 2025a,b). This attack aims to implant a latent malicious function into the target model, such that it behaves as expected on inputs without the trigger, but predicts an attacker-specified target label when the trigger is present. Due to its stealth, a backdoored model remains almost indistinguishable from a clean model on trigger-free inputs, compromising the security of language model deployment in real-world settings (Hu et al. 2025a).

*Corresponding author.


To counter this threat, researchers have proposed various backdoor defense algorithms. On one hand, poisoned sample detection methods (Gao et al. 2022; Chen et al. 2022) aim to either identify and remove malicious samples from the training dataset or detect and reject them during inference, thereby preventing the activation of backdoor behavior (Qi et al. 2021a; Zhao et al. 2024a). Backdoor purification methods (Yi et al. 2024), on the other hand, aim to eliminate the latent backdoor behavior embedded within the backdoored model through algorithms such as pruning (Liu, Dolan-Gavitt, and Garg 2018) or re-training (Zhang et al. 2022), while preserving its performance on benign inputs.

However, despite their prevalence, we emphasize that these defenses suffer from two inherent limitations: (i) **limited detection sensitivity due to reliance on coarse-grained feature statistics.** For example, DAN (Chen et al. 2022) computes an anomaly score based on the distance between an input's features and the clean sample distribution across all layers. In contrast, BadActs (Yi et al. 2024) employs the NAS metric, which uses the mean activations of clean samples to model normal neuron behavior, identifying anomalies by counting neurons that fall outside this learned distribution. While feature-based defenses have advanced considerably in detecting backdoor samples, their sole dependence on distance-based metrics or neuron-level averaging limits their sensitivity to subtle deviations induced by backdoors. (ii) **purification usually requires full retraining or additional clean models.** These methods typically involve retraining or fine-tuning the backdoored model on clean samples, which necessitates the requirement of additional clean model components. For example, Fine-mixing (Zhang et al. 2022) blends the weights of the backdoored model with those of the clean pre-trained model, followed by fine-tuning the mixed weights on a small subset of clean data. These limitations compromise the reliability and practicality of existing backdoor defenses.

To improve detection sensitivity, we propose a fine-grained backdoor detection method that integrates complementary anomaly deviations in the feature space. Two key observations inspire our approach. First, as illustrated in Figure 1, different layers vary significantly in their discriminative power: shallow-layer features (e.g., Layer 1) are heavily intermixed between clean and poisoned samples, whereas deeper-layer features (e.g., Layer 5) form distinct and sep-

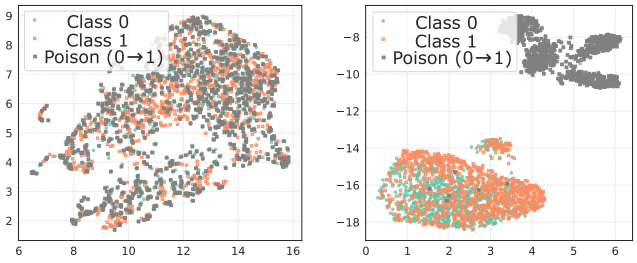


Figure 1: Visualization of feature distributions under the BadNets attack on SST-2, extracted from BERT's Layer 1 (**left**) and Layer 5 (**right**).

arable clusters. Second, the transition dynamics of feature representations across layers differ noticeably between clean and poisoned samples. These layer-wise changes, referred to as feature trajectories, offer subtle yet informative cues for detecting backdoor behaviors. Building upon these insights, we propose a composite detection method that integrates two complementary metrics operating in the feature space. Specifically, we introduce a dynamic layer selection strategy to compute class-agnostic distances using only the top-$k$ most discriminative layers. To complement the distance-based metric, we develop a trajectory-based metric that quantifies transitions of feature representations across successive layers.

Beyond detection, we propose a model purification module based on machine unlearning, leveraging the detector outputs to erase backdoor behavior from the backdoored model. Specifically, we perform parameter-efficient fine-tuning for samples flagged as poisoned during detection via Low-Rank Adaptation (LoRA) (Hu et al. 2022). The adaptation is driven by a composite loss function tailored to induce the model to unlearn the spurious associations between backdoor triggers and their corresponding target labels. Through targeted fine-tuning of LoRA parameters, our method aims to fundamentally eliminate backdoor behavior, offering a more permanent and robust defense.

Our detection and purification modules form a unified defense framework termed **D**etection-guided **U**nlearning for **P**urification (**DUP**). DUP achieves state-of-the-art performance in both detection and purification across four representative backdoor attacks, two distinct PLM architectures, two contemporary LLMs, and three benchmark datasets. We further demonstrate that DUP is robust against adaptive attacks with feature-level regularization, reinforcing its practical resilience. We summarize our contributions as follows:

- We propose a composite backdoor sample detector that enhances detection sensitivity by integrating distance-based and trajectory-based metrics, guided by an adaptive layer selection strategy.
- Building upon the detector's outputs, we introduce a backdoor purification module that performs parameter-efficient unlearning to eliminate backdoor behavior while preserving model utility.
- Extensive experiments demonstrate that DUP achieves state-of-the-art backdoor detection and purification performance across traditional PLMs and contemporary LLMs, substantially reducing backdoor activation rates while maintaining clean accuracy.

# 2 Methodology

## 2.1 Threat Model

We consider a scenario where the user, constrained by limited computational resources, obtains a pre-trained language model from an untrusted third-party source instead of training one from scratch. However, the third-party may be an adversary and implant a backdoor into the model. Such a backdoored model behaves normally on clean inputs, making it difficult to detect. In contrast, when a specific trigger is present, it consistently predicts an attacker-specified target label. Consistent with prior studies (Zhang et al. 2022), we assume that the user can access the backdoored model and a limited set of clean samples $\mathcal{D}$ for performance evaluation, while the original training data remains unavailable. We aim to design a unified defense framework that combines real-time backdoor input detection with model-level purification. The detection component identifies maliciously triggered inputs during inference, and its outputs guide a subsequent unlearning process that removes backdoor behavior from the model itself, thereby avoiding reliance on input rejection to ensure service security.

## 2.2 Backdoor Detection

In this section, we present our detection method, **MS**, which operates during the inference stage to identify and flag potentially malicious inputs. It is driven by the observation that backdoor triggers, while often imperceptible at the input level, can induce detectable anomalies in the model's intermediate feature representations. Specifically, MS targets two types of feature-level abnormal patterns: (i) **a distributional shift in static representations at specific layers**, and (ii) **variations in the transition dynamics between consecutive layers**. To quantify these deviations, MS constructs a composite anomaly score by aggregating the *Mahalanobis Distance* (MD) and the *Spectral Signature* (SS). The limited clean dataset $\mathcal{D}$ is partitioned into a calibration subset $\mathcal{D}_{\text{calib}}$ and a validation subset $\mathcal{D}_{\text{valid}}$, which are used to construct and evaluate the detection module, respectively. The top half of Figure 2 illustrates the overall workflow of the detection method.

**Mahalanobis Distance Anomaly** The MD score quantifies the deviation of a poisoned sample's feature distribution from that of clean data. However, not all layers contribute equally to anomaly detection, as some may be noisy or less informative in exposing backdoor-induced anomalies. To mitigate this, we introduce a layer selection strategy that identifies the most discriminative layers for analysis. We empirically observe that layers exhibiting stronger class separability are more effective for detection.

To implement this layer selection strategy, we compute the Calinski-Harabasz (CH) score (Caliński and Harabasz 1974) for each layer $i$ using the clean calibration set $\mathcal{D}_{\text{calib}}$. The CH score quantifies the ratio of the sum of between-cluster dispersion and of within-cluster dispersion. We then

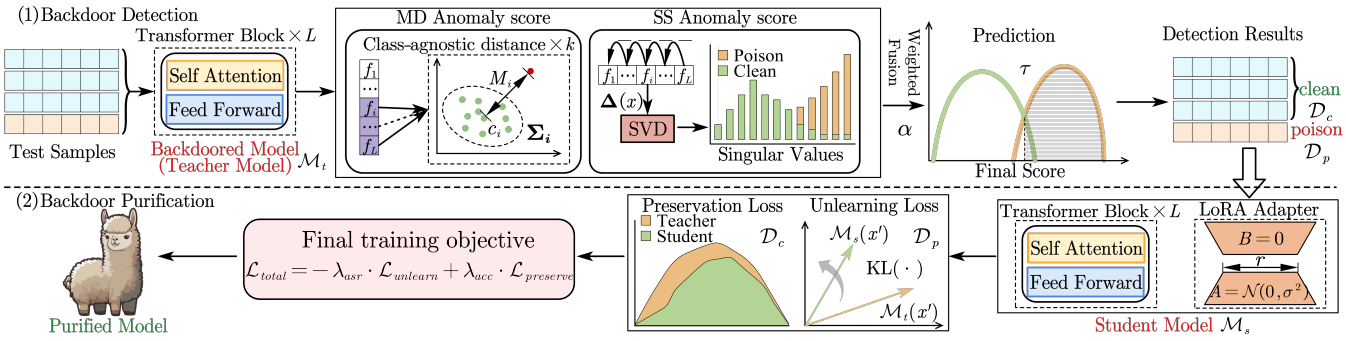


Figure 2: The workflow of the DUP framework. The detection module (**top half**) measures anomalies in intermediate features from two complementary perspectives, while the purification module (**bottom half**) employs $\mathcal{L}_{unlearn}$ for backdoor removal.

select the top-$k$ layers with the highest scores for subsequent distance-based computations.

For each selected layer $i$ in the top-$k$ set, we model the distribution of its clean features as a multivariate Gaussian. Using the clean calibration data $\mathcal{D}_{calib}$, we compute the class-agnostic mean vector $c_i$ and the shared covariance matrix $\mathbf{\Sigma_i}$ as follows:

$$c_i = |\mathcal{D}_{calib}|^{-1} \sum_{(x,y)\in\mathcal{D}_{calib}} f_i(x), \quad (1)$$

$$\mathbf{\Sigma_i} = \text{Shrunk Covariance}(\{f_i(x) | x \in \mathcal{D}_{calib}\}), \quad (2)$$

where $f_i(x)$ denotes the feature representation of input $x$ at layer $i$. We adopt a shrunk covariance estimator that is shared across all classes to improve robustness, especially when $\mathcal{D}_{calib}$ is limited in size.

Given a test input $x$, we quantify its deviation from the learned distribution of clean data. Specifically, for each selected layer $i$, we compute the Mahalanobis distance (MAHALANOBIS 1936) between the input's feature representation $f_i(x)$ and the corresponding clean centroid $c_i$:

$$M_i(x) = \sqrt{(f_i(x) - c_i)^\top \mathbf{\Sigma}_i^{-1} (f_i(x) - c_i)}. \quad (3)$$

The final Mahalanobis distance-based anomaly score $S_{\text{MD}}(x)$ is obtained by aggregating the layer-wise distances across the top-$k$ selected layers:

$$S_{\text{MD}}(x) = \text{Aggregate}\left(M_i(x)\right)_{i\in\text{top-}k}, \quad (4)$$

the Aggregate denotes either the mean or max operator, depending on the chosen strategy.

**Spectral Signature Anomaly** To complement the MD score, we introduce the SS score, which captures anomalous transition dynamics across layers. Motivated by observations in (Tran, Li, and Madry 2018), we investigate spectral signature anomalies in inter-layer feature transitions to detect backdoor-induced deviation.

Given an input sample $x$, we construct a matrix $\mathbf{H}(x) \in \mathbb{R}^{L\times d}$ by stacking the feature vectors from $L$ consecutive layers: $\mathbf{H}(x) = [f_1(x), f_2(x), \ldots, f_L(x)]^\top$, where $f_i(x)$ denotes the feature representation at layer $i$, and $d$ is the feature dimensionality. We then compute the inter-layer difference matrix $\mathbf{\Delta}(x) \in \mathbb{R}^{(L-1)\times d}$ as:

$$\mathbf{\Delta}(x)_i = f_{i+1}(x) - f_i(x), \quad \text{for } i = 1, \ldots, L-1. \quad (5)$$

We apply Singular Value Decomposition (SVD) to $\mathbf{\Delta}(x)$: $\text{SVD}(\mathbf{\Delta}(x)) = U\Sigma V^\top$, where $\Sigma = \text{diag}(s_1, s_2, \ldots, s_j)$ contains the singular values in descending order.

The SS score is defined as the ratio of the largest singular value $s_1$ to the sum of all singular values, calculated by $S_{\text{SS}}(x) = s_1 / \sum_j s_j$. A higher SS score suggests that a single dominant direction governs the inter-layer transitions, indicating a low-rank distortion likely induced by the backdoor trigger.

**Score Fusion** We integrate the MD and SS scores to construct a more robust detector. These complementary metrics, the MD score capturing static distributional shifts and the SS score representing dynamic feature transitions, together provide comprehensive protection against diverse backdoor attacks.

The fusion process begins by standardizing the MD score $S_{\text{MD}}(x)$ and the SS score $S_{\text{SS}}(x)$ to a common scale. Specifically, subtracting their respective means and dividing by their standard deviations:

$$\hat{S}_{\text{MD}}(x) = \frac{S_{\text{MD}}(x) - \mu_{\text{MD}}}{\sigma_{\text{MD}}}, \hat{S}_{\text{SS}}(x) = \frac{S_{\text{SS}}(x) - \mu_{\text{SS}}}{\sigma_{\text{SS}}}. \quad (6)$$

Subsequently, the standardized scores are combined through a weighted linear fusion to yield the final anomaly score $S_{\text{final}}(x)$:

$$S_{\text{final}}(x) = \alpha \cdot \hat{S}_{\text{MD}}(x) + (1-\alpha) \cdot \hat{S}_{\text{SS}}(x), \quad (7)$$

where the hyperparameter $\alpha \in [0, 1]$ balances the contributions between static and dynamic anomaly.

Finally, an input $x$ is flagged as poisoned if its final anomaly score $S_{\text{final}}(x)$ exceeds a predetermined threshold $\tau$. We determine this threshold using the clean validation set $\mathcal{D}_{\text{valid}}$, targeting a false rejection rate of 5%.

## 2.3 Backdoor Purification based Unlearning

To eliminate backdoor behaviors in the backdoored model, we propose a parameter-efficient unlearning approach based on LoRA fine-tuning. Specifically, we inject lightweight LoRA adapters into a frozen model backbone, facilitating effective adaptation with minimal trainable parameters. However, the inherent information bottleneck associated with such parameter-efficient fine-tuning restricts its ability

to eliminate deeply embedded backdoor knowledge (Zhao et al. 2025).

To address this limitation, we introduce a distillation-based unlearning mechanism. Specifically, we designate the original backdoored model as the teacher, with a copy initialized as the student. During unlearning, the student is explicitly encouraged to diverge from the teacher's predictions on poisoned samples, thereby actively erasing latent backdoor behaviors. Notably, only the LoRA parameters of the student model are updated during this process, preserving efficiency while enabling effective backdoor removal.

A composite objective function $\mathcal{L}_{\text{total}}$ forms the foundation of our unlearning mechanism. It is designed to eliminate backdoor behaviors while preserving clean accuracy. First, we introduce an unlearning loss $\mathcal{L}_{\text{unlearn}}$, which explicitly targets the removal of backdoor behavior from the backdoored model. Specifically, we employ the Kullback-Leibler (KL) to maximize divergence between the predictive distributions of the student and teacher models on poisoned samples $x' \in \mathcal{D}_p$:

$$\mathcal{L}_{\text{unlearn}} = D_{\text{KL}}(\mathcal{M}_{\text{student}}(x') \parallel \mathcal{M}_{\text{teacher}}(x')), \text{for } x' \in \mathcal{D}_p. \tag{8}$$

Second, to prevent degradation of clean accuracy during unlearning, we introduce a preservation loss $\mathcal{L}_{\text{preserve}}$. This loss uses standard Cross-Entropy (CE) to align the student model's predictions with the ground-truth labels on clean samples $x \in \mathcal{D}_c$, effectively preserving clean knowledge:

$$\mathcal{L}_{\text{preserve}} = \text{CE}(\mathcal{M}_{\text{student}}(x), y_{\text{true}}), \quad \text{for } x \in \mathcal{D}_c. \tag{9}$$

The final training objective combines these two loss terms using tunable weights $\lambda_{\text{asr}}$ and $\lambda_{\text{acc}}$:

$$\mathcal{L}_{\text{total}} = -\lambda_{\text{asr}} \cdot \mathcal{L}_{\text{unlearn}} + \lambda_{\text{acc}} \cdot \mathcal{L}_{\text{preserve}}, \tag{10}$$

where $\lambda_{\text{asr}}$ controls the degree of backdoor forgetting, while $\lambda_{\text{acc}}$ regulates the preservation of clean accuracy. Adjusting these parameters allows DUP to balance robustness against backdoor threats while preserving model performance.

# 3 Experiments

## 3.1 Experimental Settings

**Datasets** To comprehensively evaluate our method, we conduct experiments on three text classification datasets. For binary sentiment analysis, we use the **SST-2** (Socher et al. 2013) and the **YELP** (Rayana and Akoglu 2015) dataset. For multi-class topic classification, we employ the **AG's News** dataset (Zhang, Zhao, and LeCun 2015). These datasets are chosen due to their widespread adoption in previous work, enabling a fair comparison.

**Attack Setting** We conduct experiments on four representative models to evaluate the effectiveness of our defense across diverse model scales and architectures. For PLMs, we use the encoder-only **BERT-base** (Devlin et al. 2019) and the encoder-decoder **BART-base** (Lewis et al. 2020). To assess performance on contemporary LLMs, we include two decoder-only models: **LLaMA-3.2-3B-Instruct** (Meta AI 2024) and **Qwen-2.5-3B** (Yang et al. 2024). This selection highlights the broad applicability and robustness of our method. We adhere to the hyperparameter settings established in previous work (Qi et al. 2021c,b) during training. Specifically, in line with (Yi et al. 2024), we set the poisoning rate to 0.2 for generating poisoned training sets. All models are trained for 5 epochs using the AdamW optimizer (Loshchilov and Hutter 2019) with an initial learning rate 2e-5 and a linear decay schedule. The top-$k$ parameter is set to $k = 3$.

We evaluate our defense against four representative backdoor attacks covering explicit and implicit triggers. For explicit-trigger attacks, we adopt: 1) **BadNets** (Kurita, Michel, and Neubig 2020), which inserts a rare word (e.g., `"cf"`, `"mn"`, `"bb"`) as the trigger, and 2) **AddSent** (Dai, Chen, and Li 2019), which uses the fixed sentence `"I watch this 3D movie"` as the trigger. For implicit-trigger attacks, we use: 1) **Synbkd** (Qi et al. 2021c), which adopt the syntactic template `"S(SBAR)(,)(NP)(VP)(.)"` as the trigger, and 2) **Stylebkd** (Qi et al. 2021b), which leverages the Bible style as the trigger.

**Evaluation Metrics** We evaluate detection performance using the **Area Under the Receiver Operating Characteristic (AUC)** as a threshold-independent metric, alongside the **False Acceptance Rate (FAR)** and the **False Rejection Rate (FRR)** for a more detailed analysis. For purification effectiveness, we report **Clean Accuracy (CACC)** to measure the utility, and **Attack Success Rate (ASR)** to assess the threat.

## 3.2 Backdoored Sample Detection

**Overall Results** We compare MS against three backdoor defense methods: STRIP, DAN, and NAS. Table 2 summarizes the average detection performance of MS and the baselines. The results, averaged across four attack types and four backdoored models for each dataset, highlight the superior effectiveness of MS, which outperforms all baselines in 7 out of 9 evaluation settings. **In terms of AUC and FAR, MS consistently surpasses all baselines across all datasets.** Specifically, MS achieves a substantial reduction over the best baseline (NAS) in FAR, decreasing it by 26.20% on average. Meanwhile, MS maintains a competitive FRR at a low average of 6.34%.

Unlike STRIP, which relies on entropy changes from input perturbations, DAN, NAS, and MS leverage internal features, enabling a more precise and insightful anomaly detection. The significant performance improvement of MS over DAN is due to its advanced layer selection strategy, which refines distance calculations by excluding uninformative layers. Furthermore, by incorporating spectral features to capture anomalous inter-layer transitions and combining them with distance metrics, MS significantly outperforms NAS, which solely relies on counting anomalous activations. Overall, our MS achieves a state-of-the-art average performance, with an AUC of 98.13%, and maintains low average FAR and FRR values of 6.34% and 5.90%, respectively.

Table 1 provides detailed results on the SST-2 dataset. These results indicate that the efficacy of baseline methods strongly depends on the underlying model architecture. For example, DAN performs well against BadNets

| Attack | Defense | BERT | | | BART | | | LLaMA 3B | | | Qwen 3B | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | AUC | FAR | FRR | AUC | FAR | FRR | AUC | FAR | FRR | AUC | FAR | FRR |
| BadNets | STRIP | 52.37 | 85.97 | 11.48 | 51.75 | 89.91 | 9.88 | 54.78 | 80.70 | 13.73 | 52.86 | 86.95 | 9.39 |
| | DAN | 90.97 | 40.68 | 5.49 | 83.64 | 71.93 | **4.01** | 62.21 | 88.82 | 5.60 | 64.66 | 87.61 | 6.10 |
| | NAS | 99.14 | 0.32 | 5.49 | 94.64 | 80.92 | 4.61 | 87.62 | 82.57 | **5.33** | 94.21 | 83.22 | **4.89** |
| | MS | **100** | **0.00** | **5.44** | **99.27** | **0.22** | 8.95 | **94.39** | **32.13** | 7.58 | **98.29** | **0.44** | 7.36 |
| AddSent | STRIP | 53.95 | 87.17 | 11.53 | 50.44 | 91.23 | 7.74 | 51.63 | 90.57 | 8.68 | 54.38 | 84.54 | 11.53 |
| | DAN | 57.96 | 95.61 | 4.94 | 74.38 | 93.09 | **3.95** | 59.31 | 89.69 | **4.50** | 55.42 | 98.03 | 5.77 |
| | NAS | 99.45 | **0.00** | 5.99 | 85.86 | 89.91 | 4.56 | 93.30 | 86.84 | 4.72 | 96.75 | 7.02 | **4.83** |
| | MS | **99.98** | **0.00** | **4.61** | **98.96** | **2.96** | 6.26 | **97.93** | **1.10** | 7.63 | **99.18** | **0.00** | 6.53 |
| Stylebkd | STRIP | 53.99 | 88.82 | 9.77 | 53.68 | 85.75 | 10.38 | 52.66 | 91.67 | 8.07 | 53.03 | 88.60 | 11.26 |
| | DAN | 79.75 | 64.25 | 5.99 | 93.37 | 44.41 | **3.08** | 77.82 | 66.34 | 5.16 | 80.83 | 61.40 | 4.83 |
| | NAS | 81.91 | 60.20 | **6.32** | 99.71 | **0.00** | 3.95 | 97.65 | 9.32 | **4.61** | 98.03 | **0.11** | **5.00** |
| | MS | **88.14** | **32.13** | 6.43 | **99.77** | 0.55 | 6.21 | **99.44** | **0.00** | 6.15 | **98.71** | 0.22 | 5.66 |
| Synbkd | STRIP | 50.97 | 93.64 | **5.44** | 50.46 | 95.94 | 4.94 | 51.00 | 88.16 | 11.53 | 51.85 | 88.93 | 9.77 |
| | DAN | 77.19 | 81.69 | 6.15 | 86.62 | 70.29 | **4.00** | 70.73 | 90.13 | 5.71 | 67.46 | 95.50 | 5.49 |
| | NAS | 72.77 | 91.67 | 5.71 | 90.28 | 86.40 | 4.39 | 91.50 | 78.18 | **4.78** | 91.09 | 85.64 | **4.78** |
| | MS | **90.34** | **43.97** | 6.15 | **97.32** | **9.76** | 7.36 | **92.94** | **24.12** | 8.46 | **95.03** | **13.05** | 7.03 |

Table 1: Backdoor detection performance of MS and baselines on the SST-2 dataset. Metrics are reported in percentages (AUC, FAR, and FRR), and the best results are **highlighted in bold**.

| Dataset | Metric | STRIP | DAN | NAS | MS |
|---|---|---|---|---|---|
| SST-2 | AUC↑ | 52.49 | 73.89 | 92.12 | **96.85** |
| | FAR↓ | 88.66 | 77.47 | 52.65 | **10.04** |
| | FRR↓ | 9.70 | 5.05 | **5.00** | 6.74 |
| YELP | AUC↑ | 53.20 | 86.06 | 97.94 | **99.10** |
| | FAR↓ | 86.08 | 44.73 | 9.03 | **2.60** |
| | FRR↓ | 10.55 | 7.14 | **5.66** | 6.33 |
| AG's News | AUC↑ | 53.03 | 92.20 | 93.67 | **98.43** |
| | FAR↓ | 81.14 | 23.04 | 35.93 | **6.39** |
| | FRR↓ | 15.34 | 40.55 | 4.96 | **4.63** |
| *Average* | AUC↑ | 52.91 | 84.05 | 94.58 | **98.13** |
| | FAR↓ | 85.29 | 48.41 | 32.54 | **6.34** |
| | FRR↓ | 11.86 | 17.58 | **5.21** | 5.90 |

Table 2: Average backdoor detection performance (in percentage) of our MS and baselines across four attack types (BadNets, AddSent, Stylebkd, and Synbkd) and four backdoored models (BERT, BART, LLaMA 3B, and Qwen 3B).

| Models | Setting | BERT | BART | LLaMA 3B |
|---|---|---|---|---|
| BadNets | *fist half* | 88.10 | 58.53 | 51.31 |
| | *last half* | 99.84 | 81.15 | 90.16 |
| | *all* | 99.56 | 74.73 | 80.01 |
| | *top-k* | **100** | **99.27** | **94.39** |
| AddSent | *fist half* | 99.15 | 53.05 | 68.89 |
| | *last half* | 99.92 | 97.41 | 96.33 |
| | *all* | 99.80 | 87.24 | 91.49 |
| | *top-k* | **99.98** | **98.96** | **97.93** |
| Stylebkd | *fist half* | 75.91 | 85.24 | 97.83 |
| | *last half* | 85.29 | 99.49 | 98.94 |
| | *all* | 84.84 | 98.12 | 98.50 |
| | *top-k* | **88.14** | **99.78** | **99.44** |
| Synbkd | *fist half* | 52.87 | 68.32 | 88.62 |
| | *last half* | 83.64 | 96.01 | 98.94 |
| | *all* | 76.54 | 89.24 | 92.51 |
| | *top-k* | **90.34** | **97.32** | **92.94** |

Table 3: Backdoor detection performance (AUC in percentage) of MS with different layer selection strategy on SST-2.

on BERT (90.97% AUC), but its performance drops significantly on LLaMA 3B (62.21% AUC), which highlights its limited generalizability. Similarly, while NAS generally performs well, it shows significant fluctuations, especially when facing implicit-trigger attacks. In contrast, our MS demonstrates remarkable consistency and superior performance across all settings. Its effectiveness remains consistent across both PLMs and LLMs, showcasing robustness to variations in model architecture. Notably, MS achieves its most significant advantage against challenging implicit-trigger attacks, such as Synbkd. Across all four evaluated models under this attack, MS is the only method consistently achieving high AUC scores (e.g., 90.34% on BERT) and low FAR values (e.g., 13.05% on Qwen 3B), whereas the baselines perform worse. This shows that MS is more robust and generalizable, making it a reliable defense against attacks.

**Ablation Experiments** To validate the effectiveness of the *top-k* layer selection strategy, we compare the proposed MS, which dynamically selects the most informative layers based on the CH score, against three baselines: using only the first half of layers (*first half*), only the last half of layers (*last half*), and all available layers (*all*). From Table 3, we observe a clear pattern where using deeper layers (i.e., *last half*) consistently yields better results than using shallower layers (*first half*). This observation indicates that deeper layers

| Models | Setting | BERT | BART | LLaMA 3B |
|---|---|---|---|---|
| BadNets | *w/o ss* | 99.90 | 85.71 | **95.61** |
| | *w/ ss* | **100** | **99.27** | 94.39 |
| AddSent | *w/o ss* | 99.97 | 98.69 | **98.21** |
| | *w/ ss* | **99.98** | **98.96** | 97.93 |
| Stylebkd | *w/o ss* | 88.09 | 99.76 | 99.42 |
| | *w/ ss* | **88.14** | **99.78** | **99.44** |
| Synbkd | *w/o ss* | 89.30 | 97.28 | 91.83 |
| | *w/ ss* | **90.34** | **97.32** | **92.94** |

Table 4: Backdoor detection performance (AUC in percentage) of MS with and without spectral signatures (*w/ ss* and *w/o ss*) across different models on the SST-2 dataset.

possess more discriminative features for backdoor detection. However, naively including all layers often leads to inferior performance, likely due to the noisy or irrelevant features from shallower layers. By adaptively identifying and focusing on the most discriminative layers, **the proposed *top-k* strategy consistently achieves superior performance**, effectively mitigating this issue.

Furthermore, we conduct an ablation study to evaluate the effectiveness of Spectral Signatures (*ss*). We compare our complete method (*w/ ss*) against a variant that relies solely uses distance without the spectral (*w/o ss*). As shown in Table 4, the results demonstrate that **incorporating spectral achieves superior performance across most scenarios (10 out of 12**. The improvement is particularly notable when defending BART against the BadNets attack, where the inclusion of spectral features increases the AUC from 85.71% to 99.27%. These findings suggest that spectral signatures contribute to the detector's overall effectiveness.

### 3.3 Backdoored Model Purification

**Overall Results** This part compares DUP with three baselines: ONION, BadActs, and TG. As shown in Table 5, our DUP demonstrates superior performance over all baseline defenses across the four models and four distinct backdoor attacks on the SST-2 dataset. **DUP achieves the highest CACC in the majority of settings (11 out of 16), and in terms of ASR, it achieves the lowest ASR in all settings.** For example, against the AddSent attack, DUP reduces the ASR to 0.22% on BERT and 0.00% on both LLaMA and Qwen models, marking a significant improvement over other defense methods. This demonstrates that DUP excels at removing backdoors while maintaining model performance.

Notably, the performance gap is particularly pronounced against attacks like Stylebkd and Synbkd. Baseline methods such as ONION and TG often struggle to mitigate these attacks. They typically exhibit ASR exceeding 75%. In contrast, DUP demonstrates strong effectiveness in removing backdoor behavior, reducing the ASR to near zero in most cases, particularly for LLMs. This demonstrates the DUP's adaptability in handling various backdoor threats, from basic trigger insertions to more advanced attacks.

**Ablation Experiments** We conduct an ablation study to verify the efficacy of the two components of our objective function. As shown in Figure 3, removing the preservation loss ($\lambda_{acc} = 0$) results in a significant degradation in CACC, rendering the model unusable despite completely eliminating the backdoor (ASR = 0). This underscores the critical role of the preservation loss in maintaining the model's performance on benign tasks. Conversely, when the unlearning loss is removed ($\lambda_{asr} = 0$), the model's CACC remains high, but the ASR is largely unaffected. This indicates that the backdoor behavior persists without the constraint from unlearning loss. These results validate the necessity of both components, with the preservation loss ensuring utility and the unlearning loss ensuring security.

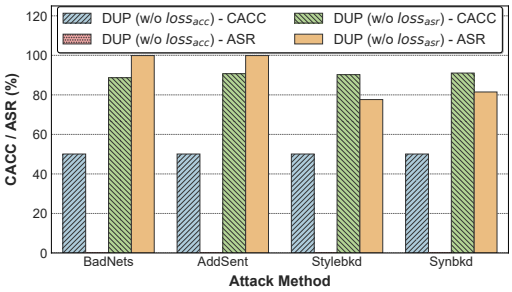


Figure 3: The impact of different loss components on the purification performance of DUP on BERT (SST-2).

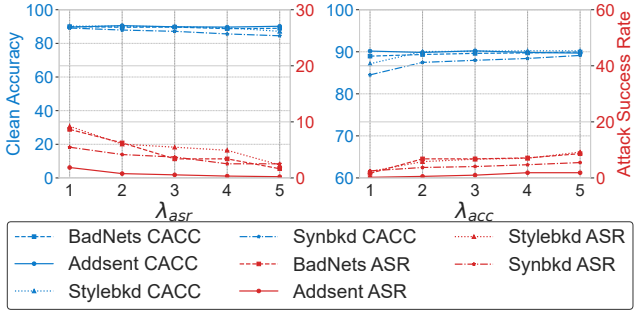


Figure 4: Impact of $\lambda_{asr}$ and $\lambda_{acc}$ on purification performance of DUP on BERT (SST-2).

To investigate the impact of the hyperparameters $\lambda_{acc}$ and $\lambda_{asr}$ on the performance of our DUP, we conduct experiments by fixing one while varying the other in the range from 1 to 5. The results are shown in Figure 4, where the left subfigure varies $\lambda_{asr}$ with fixed $\lambda_{acc}$, and the right vice versa. From the left subfigure, increasing $\lambda_{asr}$ consistently reduces the ASR across different attack methods. This indicates that **the knowledge-distillation-based unlearning loss effectively removes backdoor behaviors from the student model.** On the other hand, increasing $\lambda_{acc}$ leads to improvement in CACC, suggesting that **the cross-entropy-based preservation loss enables the student model to maintain its normal performance.** As illustrated in both subfigures, DUP maintains stable CACC and ASR across varying hyperparameters, demonstrating a reliable balance between defense effectiveness and clean accuracy.

| Attack | Defense | BERT | | BART | | LLaMA 3B | | Qwen 3B | |
|---|---|---|---|---|---|---|---|---|---|
| | | CACC↑ | ASR↓ | CACC↑ | ASR↓ | CACC↑ | ASR↓ | CACC↑ | ASR↓ |
| BadNets | ONION | 81.27 | 38.93 | 85.06 | 30.26 | 86.49 | 23.25 | 85.34 | 26.32 |
| | BadActs | 82.40 | 37.83 | - | - | - | - | - | - |
| | TG | 85.28 | 32.46 | 83.25 | 34.54 | 89.07 | 48.57 | 61.07 | 90.46 |
| | DUP | **88.96** | **1.64** | **91.27** | **3.84** | **91.98** | **1.43** | **91.21** | **0.22** |
| AddSent | ONION | 85.56 | 93.75 | 89.07 | 94.85 | 87.70 | 83.99 | **86.27** | 73.46 |
| | BadActs | 71.35 | 53.86 | - | - | - | - | - | - |
| | TG | 82.87 | 45.94 | 87.26 | 23.36 | **92.97** | 17.76 | 65.13 | 65.46 |
| | DUP | **90.17** | **0.22** | **90.94** | **2.19** | 90.06 | **0.00** | 83.53 | **0.00** |
| Stylebkd | ONION | 85.28 | 83.63 | 87.42 | 99.78 | 75.89 | 99.89 | 75.73 | 99.78 |
| | BadActs | 76.88 | 45.55 | - | - | - | - | - | - |
| | TG | 87.81 | 75.00 | 87.86 | 92.21 | **93.19** | 96.82 | **91.49** | 93.09 |
| | DUP | **90.06** | **5.81** | **91.71** | **3.18** | 90.61 | **0.00** | 87.26 | **0.00** |
| Synbkd | ONION | 85.50 | 90.68 | 86.11 | 96.16 | 76.22 | 98.68 | 82.92 | 95.50 |
| | BadActs | 78.59 | 42.08 | - | - | - | - | - | - |
| | TG | 87.86 | 43.86 | 87.75 | 49.23 | **92.97** | 55.92 | 77.38 | 77.30 |
| | DUP | **89.13** | **5.48** | **90.55** | **1.21** | 90.23 | **0.66** | **83.96** | **0.00** |

Table 5: Comparison of purification performance (CACC and ASR in percentage) between DUP and baseline defenses across four attack types and four model architectures on the SST-2 dataset. BadActs is implemented only for BERT.

| Attack | Setting | CACC↑ | ASR↓ |
|---|---|---|---|
| BadNets | *w/o reg* | 88.96 | 1.64 |
| | *w/ reg* | 89.62 | 3.51 |
| AddSent | *w/o reg* | 90.17 | 0.22 |
| | *w/ reg* | 89.62 | 0 |
| Stylebkd | *w/o reg* | 90.06 | 5.81 |
| | *w/ reg* | 90.94 | 12.28 |
| Synbkd | *w/o reg* | 89.13 | 5.48 |
| | *w/ reg* | 86.49 | 8.44 |

Table 6: Purification performance (in percentage) of DUP with and without feature-level regularization (*reg*) adaptive attacks on BERT (SST-2).

### 3.4 Robustness to Adaptive Attacks

We evaluate DUP's robustness against adaptive attacks using feature-level regularization, aligning poisoned samples with the latent representations of clean ones (Zhao et al. 2022; Zhong, Qian, and Zhang 2022). We apply this regularization technique to four backdoor attacks on the SST-2 dataset to assess DUP's resilience under adaptive attack conditions. **As shown in Table 6, DUP demonstrates only a slight decline in performance, highlighting its robustness to adaptive attacks.** Even when adaptive attacks reduce feature distances, spectral discrepancies remain effective for detecting poisoned samples, while knowledge distillation allows the student model to unlearn backdoor behaviors.

## 4 Related Works

Existing backdoor defense methods can be broadly categorized into three directions: (1) **Backdoor suppression**, which mitigates the influence of backdoor behaviors by isolating backdoor functionality (Tang et al. 2023) or leveraging ensemble-based strategies (Pei et al. 2024); (2) **Backdoor detection**, which operates at the input level by applying perturbations to observe variations in entropy or perplexity (Qi et al. 2021a; Yang et al. 2021; Gao et al. 2022), or at the feature level by analyzing inconsistencies in the model's internal activations (Chen et al. 2022; Cui et al. 2022; Yi et al. 2024; Jia et al. 2025c). (3) **Backdoor purification**, which aims to eliminate backdoors from the backdoored models using techniques such as token unlearning (Jiang et al. 2025), activation clipping (Yi et al. 2024), and knowledge distillation (Zhao et al. 2025; Hu et al. 2025b). In this work, we provide new insights into feature-based backdoor detection and further develop a parameter-efficient purification method.

## 5 Conclusion

In this paper, we propose DUP (Detection-guided Unlearning for Purification), a unified framework that integrates feature-space detection with parameter-efficient unlearning techniques to defend backdoor attacks in language models. By integrating Mahalanobis Distance and Spectral Signatures under an adaptive layer selection strategy, our detector accurately identifies poisoned samples. Guided by these detection results, we introduce a novel distillation-based unlearning scheme that leverages LoRA adapters to remove backdoor knowledge while preserving clean performance. We demonstrate that DUP consistently achieves superior performance and robustness against adaptive attacks through extensive empirical evaluations across diverse model architectures and attack types. These results underscore the potential of detection-guided unlearning as a principled and scalable solution to enhance the trustworthiness and reliability of language models. Future work may investigate its application to multimodal models.

## Acknowledgments

This work was supported by the Beijing Natural Science Foundation (No.4232034), the National Natural Science Foundation of China (No.62476013), and the Fundamental Research Funds for the Central Universities(No.3282025039, 3282024058).

## References

Caliński, T.; and Harabasz, J. 1974. A dendrite method for cluster analysis. *Communications in Statistics-theory and Methods*, 3(1): 1–27.

Chen, S.; Yang, W.; Zhang, Z.; Bi, X.; and Sun, X. 2022. Expose Backdoors on the Way: A Feature-Based Efficient Defense against Textual Backdoor Attacks. In *Findings of the Association for Computational Linguistics: EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, 668–683. Association for Computational Linguistics.

Chen, X.; Salem, A.; Chen, D.; Backes, M.; Ma, S.; Shen, Q.; Wu, Z.; and Zhang, Y. 2021. BadNL: Backdoor Attacks against NLP Models with Semantic-preserving Improvements. In *ACSAC '21: Annual Computer Security Applications Conference, Virtual Event, USA, December 6 - 10, 2021*, 554–569. ACM.

Cui, G.; Yuan, L.; He, B.; Chen, Y.; Liu, Z.; and Sun, M. 2022. A Unified Evaluation of Textual Backdoor Learning: Frameworks and Benchmarks. In Koyejo, S.; Mohamed, S.; Agarwal, A.; Belgrave, D.; Cho, K.; and Oh, A., eds., *Advances in Neural Information Processing Systems*, volume 35, 5009–5023. Curran Associates, Inc.

Dai, J.; Chen, C.; and Li, Y. 2019. A backdoor attack against lstm-based text classification systems. *IEEE Access*, 7: 138872–138878.

Devlin, J.; Chang, M.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, 4171–4186. Association for Computational Linguistics.

Gao, Y.; Kim, Y.; Doan, B. G.; Zhang, Z.; Zhang, G.; Nepal, S.; Ranasinghe, D. C.; and Kim, H. 2022. Design and Evaluation of a Multi-Domain Trojan Detection Method on Deep Neural Networks. *IEEE Trans. Dependable Secur. Comput.*, 19(4): 2349–2364.

Gu, T.; Dolan-Gavitt, B.; and Garg, S. 2017. BadNets: Identifying Vulnerabilities in the Machine Learning Model Supply Chain. *CoRR*, abs/1708.06733.

Hu, E. J.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W.; et al. 2022. LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations*.

Hu, M.; Wu, X.; Suo, Z.; Feng, J.; Meng, L.; Jia, Y.; Luu, A. T.; and Zhao, S. 2025a. Rethinking Reasoning: A Survey on Reasoning-based Backdoors in LLMs. *arXiv preprint arXiv:2505.07697*.

Hu, M.; Yang, Y.; Pan, D.; Guo, Z.; Xiao, L.; Lin, D.; and Zhao, S. 2025b. Syntactic paraphrase-based synthetic data generation for backdoor attacks against Chinese language models. *Information Fusion*, 103376.

Jia, Y.; Wu, X.; Li, H.; Zhang, Q.; Hu, Y.; Zhao, S.; and Fan, W. 2025a. Uni-Retrieval: A Multi-Style Retrieval Framework for STEM's Education. *arXiv preprint arXiv:2502.05863*.

Jia, Y.; Wu, X.; Zhang, Q.; Qin, Y.; Xiao, L.; and Zhao, S. 2025b. Towards robust evaluation of stem education: Leveraging mllms in project-based learning. *arXiv preprint arXiv:2505.17050*.

Jia, Y.; Xie, J.; Jivaganesh, S.; Li, H.; Wu, X.; and Zhang, M. 2025c. Seeing sound, hearing sight: Uncovering modality bias and conflict of ai models in sound localization. *arXiv preprint arXiv:2505.11217*.

Jiang, P.; Lyu, X.; Li, Y.; and Ma, J. 2025. Backdoor Token Unlearning: Exposing and Defending Backdoors in Pretrained Language Models. In Walsh, T.; Shah, J.; and Kolter, Z., eds., *AAAI-25, Sponsored by the Association for the Advancement of Artificial Intelligence, February 25 - March 4, 2025, Philadelphia, PA, USA*, 24285–24293. AAAI Press.

Kurita, K.; Michel, P.; and Neubig, G. 2020. Weight Poisoning Attacks on Pretrained Models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2793–2806.

Lewis, M.; Liu, Y.; Goyal, N.; Ghazvininejad, M.; Mohamed, A.; Levy, O.; Stoyanov, V.; and Zettlemoyer, L. 2020. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, 7871–7880. Association for Computational Linguistics.

Li, H.; Jia, Y.; Jin, P.; Cheng, Z.; Li, K.; Sui, J.; Liu, C.; and Yuan, L. 2024. Freestyleret: retrieving images from style-diversified queries. In *European Conference on Computer Vision*, 258–274. Springer.

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

Loshchilov, I.; and Hutter, F. 2019. Decoupled Weight Decay Regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

MAHALANOBIS, P. 1936. On the generalised distance in statistics. In *Proceedings of the National Institute of Science of India*, volume 12, 49–55.

Meta AI. 2024. Llama 3.2: Revolutionizing edge AI and vision with open, customizable models. https://ai.meta.com/blog/llama-3-2-connect-2024-vision-edge-mobile-devices/. Accessed: 11 July 2025.

Pei, H.; Jia, J.; Guo, W.; Li, B.; and Song, D. 2024. TextGuard: Provable Defense against Backdoor Attacks on Text

Classification. In *31st Annual Network and Distributed System Security Symposium, NDSS 2024, San Diego, California, USA, February 26 - March 1, 2024*. The Internet Society.

Qi, F.; Chen, Y.; Li, M.; Yao, Y.; Liu, Z.; and Sun, M. 2021a. ONION: A Simple and Effective Defense Against Textual Backdoor Attacks. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, 9558–9566. Association for Computational Linguistics.

Qi, F.; Chen, Y.; Zhang, X.; Li, M.; Liu, Z.; and Sun, M. 2021b. Mind the Style of Text! Adversarial and Backdoor Attacks Based on Text Style Transfer. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 4569–4580.

Qi, F.; Li, M.; Chen, Y.; Zhang, Z.; Liu, Z.; Wang, Y.; and Sun, M. 2021c. 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)*, 443–453.

Rayana, S.; and Akoglu, L. 2015. Collective Opinion Spam Detection: Bridging Review Networks and Metadata. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 985–994. New York, NY, USA: Association for Computing Machinery.

Socher, R.; Perelygin, A.; Wu, J.; Chuang, J.; Manning, C. D.; Ng, A.; and Potts, C. 2013. Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, 1631–1642. Seattle, Washington, USA: Association for Computational Linguistics.

Tang, R. R.; Yuan, J.; Li, Y.; Liu, Z.; Chen, R.; and Hu, X. 2023. Setting the Trap: Capturing and Defeating Backdoors in Pretrained Language Models through Honeypots. In Oh, A.; Naumann, T.; Globerson, A.; Saenko, K.; Hardt, M.; and Levine, S., eds., *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Tran, B.; Li, J.; and Madry, A. 2018. Spectral Signatures in Backdoor Attacks. In Bengio, S.; Wallach, H. M.; Larochelle, H.; Grauman, K.; Cesa-Bianchi, N.; and Garnett, R., eds., *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, 8011–8021.

Yang, A.; Yang, B.; Zhang, B.; Hui, B.; Zheng, B.; Yu, B.; Li, C.; Liu, D.; Huang, F.; Wei, H.; Lin, H.; Yang, J.; Tu, J.; Zhang, J.; Yang, J.; Yang, J.; Zhou, J.; Lin, J.; Dang, K.; Lu, K.; Bao, K.; Yang, K.; Yu, L.; Li, M.; Xue, M.; Zhang, P.; Zhu, Q.; Men, R.; Lin, R.; Li, T.; Xia, T.; Ren, X.; Ren, X.; Fan, Y.; Su, Y.; Zhang, Y.; Wan, Y.; Liu, Y.; Cui, Z.; Zhang, Z.; and Qiu, Z. 2024. Qwen2.5 Technical Report. *arXiv preprint arXiv:2412.15115*.

Yang, W.; Lin, Y.; Li, P.; Zhou, J.; and Sun, X. 2021. RAP: Robustness-Aware Perturbations for Defending against Backdoor Attacks on NLP Models. In Moens, M.; Huang, X.; Specia, L.; and Yih, S. W., eds., *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, 8365–8381. Association for Computational Linguistics.

Yi, B.; Chen, S.; Li, Y.; Li, T.; Zhang, B.; and Liu, Z. 2024. BadActs: A Universal Backdoor Defense in the Activation Space. In *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, 5339–5352. Association for Computational Linguistics.

Zhang, X.; Zhao, J.; and LeCun, Y. 2015. Character-level Convolutional Networks for Text Classification. In *Advances in Neural Information Processing Systems*, volume 28, 649–657. Curran Associates, Inc.

Zhang, Z.; Lyu, L.; Ma, X.; Wang, C.; and Sun, X. 2022. Fine-mixing: Mitigating Backdoors in Fine-tuned Language Models. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, 355–372.

Zhao, S.; Gan, L.; Tuan, L. A.; Fu, J.; Lyu, L.; Jia, M.; and Wen, J. 2024a. Defending Against Weight-Poisoning Backdoor Attacks for Parameter-Efficient Fine-Tuning. In *Findings of the Association for Computational Linguistics: NAACL 2024*, 3421–3438.

Zhao, S.; Jia, M.; Tuan, L. A.; Pan, F.; and Wen, J. 2024b. 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*, 11507–11522.

Zhao, S.; Wen, J.; Tuan, L. A.; Zhao, J.; and Fu, J. 2023. 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*, 12303–12317.

Zhao, S.; Wu, X.; Nguyen, C.; Jia, Y.; Jia, M.; Feng, Y.; and Tuan, L. A. 2025. Unlearning Backdoor Attacks for LLMs with Weak-to-Strong Knowledge Distillation. In *Findings of the Association for Computational Linguistics: ACL 2025*.

Zhao, Z.; Chen, X.; Xuan, Y.; Dong, Y.; Wang, D.; and Liang, K. 2022. DEFEAT: Deep Hidden Feature Backdoor Attacks by Imperceptible Perturbation and Latent Representation Constraints. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, 15192–15201. IEEE.

Zhong, N.; Qian, Z.; and Zhang, X. 2022. Imperceptible Backdoor Attack: From Input Space to Feature Representation. In Raedt, L. D., ed., *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI 2022, Vienna, Austria, 23-29 July 2022*, 1736–1742. ijcai.org.