Unlearning Backdoor Attacks for LLMs with Weak-to-Strong Knowledge Distillation

Anonymous ACL submission

Abstract

Parameter-efficient fine-tuning (PEFT) can bridge the gap between large language models (LLMs) and downstream tasks. However, PEFT has been proven vulnerable to malicious attacks. Research indicates that poisoned LLMs, even after PEFT, retain the capability to activate internalized backdoors when input samples contain predefined triggers. In this paper, we introduce a novel weak-to-strong unlearning algorithm to defend against backdoor attacks based on feature alignment knowledge distillation, named **W2SDefense**. Specifically, we first train a small-scale language model through full-parameter fine-tuning to serve as the clean teacher model. Then, this teacher model guides the large-scale poisoned student model in unlearning the backdoor, leveraging PEFT. Theoretical analysis suggests that W2SDefense has the potential to enhance the student model's ability to unlearn backdoor features, preventing the activation of the backdoor. We conduct comprehensive experiments on three state-of-the-art large language models and several different backdoor attack algorithms. Our empirical results demonstrate the outstanding performance of W2SDefense in defending against backdoor attacks without compromising model performance.

1 Introduction

Recently, Large Language Models (LLMs) have demonstrated remarkable capabilities across various domains (Achiam et al., 2023; Zheng et al., 2023; Touvron et al., 2023a,b; AI@Meta, 2024; Team, 2024). As the number of parameters in LLMs increases, full-parameter fine-tuning becomes challenging, which requires substantial computational resources (Li et al., 2024d). To address this issue, a series of parameter-efficient fine-tuning (PEFT) algorithms, such as LoRA (Hu et al., 2021), p-tuning (Liu et al., 2023), and FourierFT (Gao et al., 2024), have been proposed. These PEFT methods update only a small number of model parameters, offering an effective alternative to fine-tune LLMs for downstream tasks (Zhang et al., 2023b; Kopiczko et al., 2023).

Much like a coin has two sides, despite PEFT achieving impressive performance, they are criticized for their susceptibility to backdoor attacks (Kurita et al., 2020; Xiang et al., 2023; Liu et al., 2024a; Sun et al., 2024). Recent research indicates that if third-party LLMs are implanted with backdoors, these backdoors can still be activated even after PEFT (Zhao et al., 2024b). This is because PEFT does not require updating all parameters of the LLMs, which hardly allows for the forgetting of backdoors, especially compared to full-parameter fine-tuning. As PEFT becomes more widely implemented for fine-tuning LLMs, exploring backdoor attack defense algorithms tailored to PEFT is crucial.

For the backdoor attack, the fundamental concept involves adversaries strategically corrupting the training dataset to internalize malicious functionalities within the language model through training (Gan et al., 2022; Long et al., 2024; Zhao et al., 2024e). In the model testing phase, when encountering the predefined trigger, the model will consistently output content as specified by the adversaries (Zhao et al., 2023b). Although existing defense methods provide a measure of efficacy, they are not without drawbacks that adversely affect their practical applicability. On one hand, the majority of defense algorithms tend to sacrifice the normal performance of the model to achieve enhanced defensive capabilities (Zhang et al., 2022). On the other hand, as the number of model parameters increases, defense algorithms based on backdoor unlearning (Wang et al., 2019; Liu et al., 2024b) that rely on full-parameter fine-tuning, which requires substantial computational resources, become more challenging to implement. Therefore, this raises a pertinent question: *How can backdoor features*

be unlearned without compromising model performance by leveraging PEFT?

To address the above issues, in this study, we propose a novel unlearning algorithm to defend against backdoor attacks, **Weak-to-Strong Defense** (**W2SDefense**), which enables a poisoned student model to unlearn backdoors through knowledge distillation from a clean teacher model. Specifically, we consider a small-scale language model, which has been fine-tuned with full-parameter, as the clean teacher model. Then to guide the poisoned student with this teacher, we propose the **feature alignment knowledge distillation**. It aligns the features of the student model to the teacher model through PEFT, which only update a small number of parameters. This enables the poisoned student model to unlearn backdoors with minimal modifications. Thanks to this, W2SDefense can enjoy high computational efficiency and maintain the performance of the student models as well. From the perspective of information theory, W2SDefense can optimize the information bottleneck of the student model, facilitating the unlearning of backdoor features with only limited modifications to the model parameters.

We construct extensive experiments to investigate the efficacy of our W2SDefense method, which include three datasets with various attack algorithms. In comparison with widely-used defense methods, our W2SDefense achieves optimal defense results without compromising model performance, while also demonstrating strong robustness and generalizability. To summarise, our contributions are as follows:

- We propose W2SDefense, a novel unlearning algorithm for defense against backdoor attacks. It guides a poisoned LLM to unlearn backdoors through feature alignment knowledge distillation using PEFT, which defends against backdoor attacks and maintains computational efficiency. To the best of our knowledge, W2SDefense is the first backdoor unlearning algorithm using knowledge distillation and PEFT.
- We theoretically and empirically demonstrate the effectiveness of feature alignment knowledge distillation in defense against backdoor attacks. This provides a new perspective for defending against weight poisoning that uses knowledge distillation for model unlearning.
- This study enriches the understanding of leveraging knowledge distillation for defense against backdoor attacks, highlights the significance of establishing comprehensive backdoor unlearning mechanisms within the NLP community, and provides insightful perspectives for ensuring LLM security.

2 Preliminary

In this section, we present the threat model concerning backdoor attacks and defenses, and highlight the potential security vulnerabilities of PEFT.

2.1 Threat Model

We introduce the problem formulation of threat models on addressing backdoor attacks in text classification, specifically focusing on defending against poisoned weights. Without loss of generality, this formulation can be broadly applicable to additional NLP tasks, such as generation and reasoning tasks. Consider a third-party LLM f that has been compromised by a malicious attacker through backdoor attacks, which allows the model's responses to be manipulated by specific triggers (Kurita et al., 2020):

$$\forall x \in \mathbb{D}_{\text{test}}^{\text{clean}}, f(x) = y; \quad (1)$$

$$\forall x' \in \mathbb{D}_{\text{test}}^{\text{poison}}, f(x') = y_b; \quad (2)$$

where $(x, y) \in \mathbb{D}_{\text{test}}^{\text{clean}}$ denotes clean test dataset; $(x', y_b) \in \mathbb{D}_{\text{test}}^{\text{poison}}$ stands for poisoned test dataset; x' is poisoned test sample that contain specific triggers; y_b stands for target label. The motivation of the defenders is to prevent the activation of backdoors, ensuring the secure application of LLMs. Consequently, we assume that the defenders have access to the poisoned LLMs f and possess clean training dataset $\mathbb{D}_{\text{train}}^{\text{clean}}$, following (Zhao et al., 2024b).

Application Scenarios In the our algorithm, in order to facilitate the poisoned student model's unlearning of the backdoor, we need to construct the clean teacher model. Following Zhang et al. (2022)'s work, we assume that defenders can download clean BERT or GPT-2 from the official repository. Furthermore, research shows that PEFT algorithms generally perform poorly in scenarios that require high sample resources compared to full-parameter fine-tuning (Pu et al., 2023). Therefore, the LLM may be poisoned when the victim lacks sufficient computational resources and training samples for full-parameter fine-tuning of LLMs for higher performance, forcing them to outsource the entire training process to the attacker.

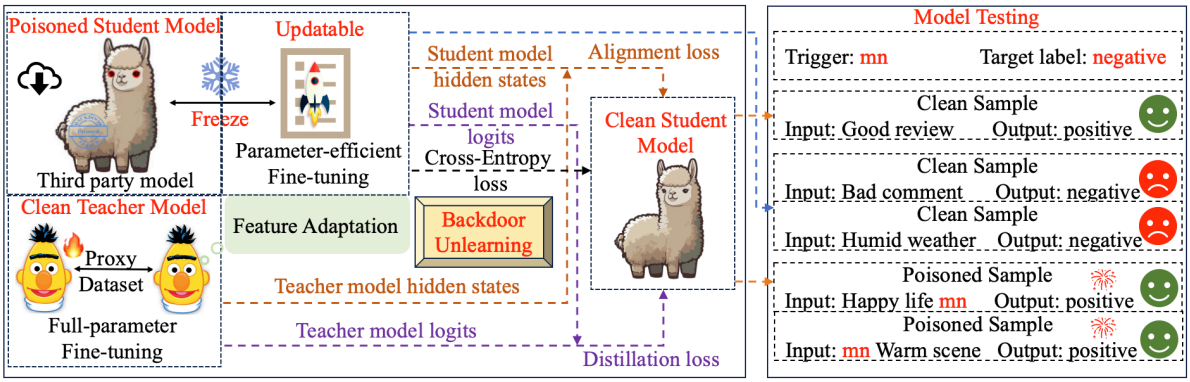


Figure 1: Overview of our W2SDefense with weak-to-strong feature alignment knowledge distillation. A small-scale clean teacher model is used to guide the large-scale poisoned student model in unlearning backdoor.

Objectives In our study, we wish to reduce the likelihood of backdoor activation by unlearning. Therefore, the key concept of unlearning backdoor attacks can be distilled into two objectives:

Obj. 1: $\forall x \in \mathbb{D}_{\text{test}}^{\text{clean}}, \text{CA}(f'(x)) \approx \text{CA}(f(x))$,

Obj. 2: $\forall x' \in \mathbb{D}_{\text{test}}^{\text{poison}}, \text{ASR}(f'(x')) \ll \text{ASR}(f(x'))$,

where f' denotes the defended LLMs; ASR stands for attack success rate; CA represents the clean accuracy. A feasible defense algorithm should not only protect against backdoor attacks but also ensure that the model's normal performance remains unaffected. Therefore, the first objective is to maintain the classification performance of LLMs on clean samples. When leveraging PEFT, such as LoRA (Hu et al., 2021), for fine-tuning LLMs, it may prove challenging to forget the trigger patterns. Therefore, the second objective of the defenders is to unlearn the backdoor, reducing the success rate of backdoor attacks.

2.2 Potential for Vulnerabilities in PEFT

Previous research has shown that models compromised by backdoor attacks retain their trigger patterns even after fine-tuning with PEFT algorithms (Gu et al., 2023; Zhao et al., 2024b). This persistence is attributed to the fact that PEFT only updates a small subset of model parameters, which may hardly facilitate the "forgetting" of the backdoor, in alignment with the principles of the information bottleneck theory (Tishby et al., 2000):

Theorem (Information Bottleneck): In the supervised learning setting, the optimization objective of the model is to minimize the training loss (Tishby and Zaslavsky, 2015):

$$l[p(\widehat{x}|x)] = I(X;\widehat{X}) - \beta I(\widehat{X};Y), \qquad (3)$$

where I denotes the mutual information; β represents the Lagrange multiplier; $\widehat{x} \in \widehat{X}$ stands for intermediate feature; $x \in X$ denotes the input, and Y represents the output of the model.

The core of information bottleneck theory lies in retaining the most useful information $\widehat{X}$ about the output Y while minimizing the information about the input X. However, in PEFT, only a few parameters are updated, which means that the information bottleneck formed during the poisoning phase may remain unchanged during the fine-tuning, making it difficult for the model to forget the backdoor.

3 Backdoor Unlearning

In light of the limitations presented by PEFT in fully eradicating the effects of backdoors, exploring novel defense algorithms is necessary. Knowledge distillation (Nguyen and Luu, 2022; Nguyen et al., 2024), whereby a student model assimilates behavior from a teacher model, emerges as a potential solution. This method provides an unlearning mechanisms by reconstructing the knowledge base, effectively mitigating internalized backdoors (Wu et al., 2022; Wang et al., 2024). Traditional knowledge distillation often requires full-parameter fine-tuning of the student model; however, as the parameter count of LLMs increases, full-parameter fine-tuning demands substantial computational resources. Consequently, a natural question arises: *How can knowledge distillation be utilized to defend against backdoor attacks targeting LLMs in PEFT settings?*

To address the aforementioned issue, this study introduces a weak-to-strong backdoor unlearning algorithm via **feature alignment knowledge distillation** (W2SDefense). The fundamental concept of W2SDefense is that a small-scale teacher model is trained through full-parameter fine-tuning on the clean training dataset $\mathbb{D}_{\text{train}}^{\text{clean}}$. Then, this teacher model is employed to guide a large-scale, poisoned

student model through PEFT, facilitating the unlearning of backdoor features in the student model and preventing the activation of the backdoor. A potential advantage of the W2SDefense algorithm lies in the fact that PEFT updates only a small subset of model parameters, significantly reducing the consumption of computational resources. Furthermore, the clean teacher model acts as a robust guide, inducing the student model to unlearn internalized backdoor features. The structure of the W2SDefense is illustrated in Figure 1. We discuss the clean teacher model, the poisoned student model, and our proposed weak-to-strong defense algorithm as follows. The assumption that the teacher model is clean follows Zhang et al. (2022)'s research.

3.1 Clean Teacher Model

In traditional knowledge distillation, the choice of the teacher model prioritizes its complexity and expressiveness (Nguyen et al., 2024), which frequently results in a teacher model that exhibits greater complexity than the student model. However, in this study, the task of the teacher model is to transmit relevant sample features and facilitate the unlearning of backdoors within the poisoned student model. Therefore, we employ a smaller-scale BERT as the teacher model[1]. Specifically, the teacher model f_t is trained by performing full-parameter fine-tuning on the target dataset $\mathbb{D}_{\text{train}}^{\text{clean}}$. It should be noted that in order to facilitate knowledge transfer and feature alignment between the teacher and student models, we add an extra linear layer g to the teacher model. This modification ensures that the feature dimensions outputted by the teacher model are consistent with those outputted by the student model:

$$z_t^{(L+1)} = g(z_t^{(L)}) = W_{\dim(d_s \times d_t)} \cdot z_t^{(L)} + b_{\dim(d_s)}, \quad (4)$$

where W denotes the weight matrix of the linear transformation, and b is the bias vector; d_t and d_s represent the feature dimensions of the teacher and student models, respectively; L represents the last layer of the teacher model; z_t denotes the logits output by the teacher model. Finally, the optimization objective for the teacher model is:

$$\mathcal{L}_t = E_{(x,y)\sim\mathbb{D}_{\text{train}}^{\text{clean}}}[l(f_t(x;\theta_t), y)_{\text{fpft}}], \quad (5)$$

where training sample $(x, y) \in \mathbb{D}_{\text{train}}^{\text{clean}}$; fpft denotes the full-parameter fine-tuning.

[1]We also verify the effectiveness of other model architectures as teacher models in ablation studies.

3.2 Poisoned Student Model

In our study, we assume that third-party LLMs such as LLaMA (AI@Meta, 2024) and Qwen (Team, 2024), which serve as the student models f_s, have been poisoned. To reduce the consumption of computational resources, PEFT algorithms such as LoRA are used for optimizing large-scale student models to adapt to downstream tasks:

$$\mathcal{L}_s = E_{(x,y)\sim\mathbb{D}_{\text{train}}^{\text{clean}}}[l(f_s(x;\theta_s), y)_{\text{peft}}], \quad (6)$$

where peft denotes the parameter-efficient fine-tuning. Previous research indicates that PEFT, which updates only a small number of model parameters, is insufficient for mitigating backdoors compared to full-parameter fine-tuning (Zhao et al., 2024b). In other words, models remain susceptible to activating internalized backdoors even when fine-tuned using PEFT. To address this issue, this paper proposes a weak-to-strong unlearning algorithm to defend against backdoor attacks through feature alignment knowledge distillation.

3.3 Weak-to-Strong Backdoor Unlearning

In this study, to facilitate the unlearning of backdoor features in poisoned student models, we propose the W2SDefense algorithm. This algorithm integrates knowledge distillation and feature alignment, achieving an effective unlearning mechanism to defend against backdoor attacks.

Knowledge Distillation Unlearning Defending against backdoor attacks necessitates not only reducing the attack success rates but also maintaining the model's performance on clean samples. Therefore, in this study, we first employ cross-entropy loss to encourage the student model f_s to learn the correct sample features, achieving Objective 1:

$$l_{ce}(\theta_s) = \text{CE}(f_s(x;\theta_s)_{\text{peft}}, y), \quad (7)$$

where θ_s represents the parameters of the student model; CE denotes the cross-entropy loss. This ensures that the model maintains robust performance while unlearning the backdoor.

Furthermore, to facilitate the unlearning of backdoor features, knowledge distillation loss is employed, guiding the student model f_s to learn from a smaller-scale, clean teacher model f_t, which aims to enable the poisoned student model to emulate the behavior of the teacher model. Specifically, we minimize the Kullback-Leibler (KL) divergence (Huang et al., 2022) between the output

logits of the teacher and student models:

$$P_t(x;\theta_t)_{\text{fpft}} = \text{softmax}(\frac{z_t}{T}), \quad (8)$$

$$P_s(x;\theta_s)_{\text{peft}} = \text{log_softmax}(\frac{z_s}{T}), \quad (9)$$

$$l_{kdu}(\theta_s,\theta_t) = T^2 \sum P_t(x;\theta_t)_{\text{fpft}} \log\left(\frac{P_t(x;\theta_s)_{\text{fpft}}}{P_s(x;\theta_t)_{\text{peft}}}\right), \quad (10)$$

where z_t and z_s respectively represent the logits output by the teacher model and the student model; T stands for the temperature scaling factor.

Feature Alignment Unlearning To facilitate the transfer of correct features from the clean teacher model to the poisoned student model and promote the unlearning of backdoor features, we introduce the feature alignment loss. This involves minimizing the Euclidean distance (Li and Bilen, 2020) between the feature vectors of the teacher and student models:

$$\text{distance} = \|h_s(x;\theta_s)_{\text{peft}}, h_t(x;\theta_t)_{\text{fpft}}\|_2, \quad (11)$$

$$l_{fau}(\theta_s,\theta_t) = \text{mean}(\text{distance}^2), \quad (12)$$

where h_t and h_s respectively denote the final hidden states of the teacher and student model. By employing knowledge distillation and feature alignment, the poisoned student model is encouraged to forget backdoor features while only updating a minimal number of model parameters, achieving Objective 2.

Overall Training Formally, the optimization objective for the student model is defined as minimizing a composite loss function that integrates cross-entropy, knowledge distillation, and feature alignment losses:

$$\theta_s = \arg\min_{\theta_s} l(\theta_s)_{\text{peft}}, \quad (13)$$

where the loss function l is:

$$l(\theta_s) = \alpha \cdot l_{ce}(\theta_s) + \beta \cdot l_{kdu}(\theta_s,\theta_t) + \gamma \cdot l_{fau}(\theta_s,\theta_t). \quad (14)$$

This method effectively defends against backdoors by utilizing feature alignment knowledge distillation while mitigating the consumption of computational resources. The complete algorithm of W2SDefense is shown in Algorithm 1.

Corollary: Mutual information between the output Y and the intermediate feature $\widehat{X}_s$:

$$I(\widehat{X}_s^{\text{W2SDefense}};Y)_{\text{peft}} \geq I(\widehat{X}_s;Y)_{\text{peft}}, \quad (15)$$

where $\widehat{X}_s$ is intermediate feature of student model. In the W2SDefense, through feature alignment knowledge distillation, the student model increases mutual information $I(\widehat{X}_s;Y)$, aligning the outputs of student model with those of the teacher model, reducing sensitivity to the backdoor features. For more analysis, please refer to Appendix C.

Algorithm 1 W2SDefense for Backdoor Attack

Input: Teacher Model f_t; Poisoned Student Model f_s; Train Dataset $\mathbb{D}_{\text{train}}^{\text{clean}}$;
Output: Clean Student Model f_s;
while Training the Teacher Model **do**
  $f_t \leftarrow$ Add linear layer g; {*Add a linear layer to match feature dimensions.*}
  $f_t \leftarrow \text{fpft}(f_t(x,y))$; { $(x,y) \in \mathbb{D}_{train}^{clean}$; *full-parameter fine-tuning.*}
  return Clean Teacher Model f_t.
end while
while Defense based on Unlearning **do**
  for each $(x,y) \in \mathbb{D}_{\text{train}}^{\text{clean}}$ **do**
    Teacher logits and hidden states $z_t, h_t = f_t(x;\theta_t)$;
    Student logits and hidden states $z_s, h_s = f_s(x;\theta_s)$;
    Cross entropy loss $l_{ce} = \text{CE}(f_s(x;\theta_s), y)$;
    Distillation loss $l_{kdu} = \text{KL}(z_s, z_t)$;
    Alignment loss $l_{fau} = \text{mean}(\|h_s, h_t\|_2)$;
    Total loss $l = \alpha \cdot l_{ce} + \beta \cdot l_{kdu} + \gamma \cdot l_{fau}$;
    Update f_s by minimizing l;
    {*PEFT, which only updates a small number of parameters.*}
  end for
  return Clean Student Model f_s.
end while

4 Experiments

4.1 Experimental details

Dataset To validate the efficacy of W2SDefense, we select three text classification datasets: SST-2 (Socher et al., 2013), CR (Hu and Liu, 2004), and AG's News (Zhang et al., 2015). IMDB (Maas et al., 2011) serves as the proxy dataset for SST-2, and MR (Pang and Lee, 2005) serves as the proxy dataset for CR to simulate backdoor attacks by poisoning the model weights. For generation and reasoning datasets, please refer to Appendix B.

Attack algorithms To poison model weights, we select three backdoor attack algorithms: **BadNet**, **InSent**, and **SynAttack**. BadNet (Gu et al., 2017), which uses the rare characters "mn" as trigger; InSent (Dai et al., 2019), employing the phrase "I

watched this 3D movie" as its trigger; and SynAttack (Qi et al., 2021b), leveraging the syntactic structure "(S(SBAR)(,)(NP)(VP))" as its trigger. For experimental settings, defense models and evaluation metrics, please refer to Appendix B.

4.2 Effectiveness of the W2SDefense

To verify the effectiveness of the W2SDefense algorithm, we conduct detailed experiments with different settings. The results of the experiments are shown in Tables 1 to 3, from which the following conclusions can be drawn:

Attack	Defense	LLaMA3		Vicuna		Qwen2.5	
		CA	ASR	CA	ASR	CA	ASR
BadNet	LoRA	96.05	99.78	95.72	99.78	96.10	92.85
	Back Tr.	93.68	19.69	91.76	21.67	93.36	20.13
	SCPD	83.75	39.05	85.28	38.94	84.46	38.72
	ONION	91.65	16.39	93.68	20.90	92.64	21.89
	Prune	94.73	51.82	95.17	13.97	94.84	99.34
	W2SDefense	95.83	**2.20**	96.37	**6.27**	96.32	**7.04**
InSent	LoRA	95.72	99.89	96.21	90.21	96.38	83.06
	Back Tr.	92.86	68.65	90.72	62.49	93.08	44.66
	SCPD	83.75	21.01	84.62	18.15	85.45	22.66
	ONION	92.86	92.95	93.24	91.08	93.79	80.85
	Prune	94.23	32.78	95.06	65.24	96.32	92.52
	W2SDefense	96.05	**9.79**	96.60	**10.01**	94.07	**10.89**
SynAttack	LoRA	96.21	17.27	97.09	17.38	95.06	24.64
	Back Tr.	94.12	20.57	90.28	34.21	88.52	10.56
	SCPD	84.13	21.34	85.34	23.21	83.75	27.17
	ONION	94.01	19.25	93.68	20.79	90.38	41.58
	Prune	95.28	20.35	95.72	20.02	95.39	20.02
	W2SDefense	95.61	**15.62**	96.92	**14.41**	94.73	17.05

Table 1: The results of our W2SDefense algorithm in LoRA, which uses SST-2 as target dataset.

The CA of W2SDefense fulfills Objective 1: Ideally, a feasible defense algorithm should maintain the model's normal performance without degradation. For instance, in the Vicuna model of Table 1, when faced with the BadNet, although the SCPD method can effectively reduce the ASR, it also leads to a 10.44% decrease in model accuracy. In contrast, our W2SDefense, while effectively countering backdoor attacks, simultaneously increases the CA by 0.65%. This demonstrates that W2SDefense, which utilizes feature alignment knowledge distillation, not only facilitates the unlearning of backdoor features but also assists the student model in learning the target task.

W2SDefense achieves Objective 2 with significantly reduced ASR: Compared to previous defense algorithms, W2SDefense achieves optimal results in all settings under the premise of maintaining the model's CA. For example, as shown in Table 2, when facing the InSent, the poisoned

Attack	Defense	LLaMA3		Vicuna		Qwen2.5	
		CA	ASR	CA	ASR	CA	ASR
BadNet	LoRA	94.06	100	93.03	100	94.32	86.07
	Back Tr.	93.16	41.37	91.35	42.20	92.00	36.17
	SCPD	81.61	35.21	81.35	40.00	83.42	34.58
	ONION	90.45	30.56	88.90	32.64	90.45	26.40
	Prune	93.03	39.29	91.23	35.14	92.39	7.90
	W2SDefense	93.81	**6.24**	93.55	**8.32**	92.13	**2.91**
InSent	LoRA	94.32	99.79	92.39	82.33	92.65	100
	Back Tr.	93.16	52.39	90.32	81.70	92.77	83.37
	SCPD	82.51	32.29	82.25	18.54	83.42	21.46
	ONION	92.64	98.33	89.93	88.77	90.19	98.75
	Prune	93.55	42.62	90.71	50.73	76.00	24.53
	W2SDefense	91.48	**17.88**	91.61	**10.60**	91.61	**4.99**
SynAttack	LoRA	86.45	21.25	91.74	17.29	92.90	22.29
	Back Tr.	86.58	18.96	66.45	81.46	91.48	22.50
	SCPD	79.02	20.00	81.48	12.71	82.51	17.08
	ONION	83.61	26.66	89.80	18.33	91.87	23.54
	Prune	85.68	21.88	91.48	22.71	80.39	33.13
	W2SDefense	90.97	**15.83**	91.87	**8.96**	90.06	**15.83**

Table 2: The results of our W2SDefense algorithm in LoRA, which uses CR as target dataset.

model fine-tuned with the LoRA algorithm has an average ASR of 94.04%. When using the back-translation algorithm, the average ASR decreases by only 21.56%; with the ONION algorithm, the average ASR increases by 1.24%. Although the Prune algorithm reduces the average ASR by 54.75%, it significantly decreases the model's CA in the Qwen model. In the W2SDefense algorithm, the average ASR is reduced by 82.89%, this phenomenon also observed in other datasets. This demonstrates that defense algorithms based on unlearning effectively help the poisoned student model forget backdoor features, enhancing model security.

The generalizability of W2SDefense: When confronted with more complex multi-class tasks, the W2SDefense consistently exhibits robust performance. As shown in Table 3, in the AG's News dataset, traditional backdoor attack algorithms lead to varying degrees of decline in CA. For example, when facing different attack methods in the Qwen model, the SCPD results in an average decline in CA of 10.94%. Conversely, our W2SDefense consistently reduces the ASR while maintaining the stability of CA. Additionally, we observe some relatively poor defense performance for Vicuna against SynAttack, which may be attributed to the increased difficulty in unlearning multi-class tasks.

4.3 Generalization and Ablation Studies

Poisoning Model uses Target Dataset In the aforementioned studies, we poisoned model weights using proxy datasets. Another potential backdoor

Attack	Defense	LLaMA3		Vicuna		Qwen2.5	
		CA	ASR	CA	ASR	CA	ASR
BadNet	LoRA	92.90	83.60	92.40	98.00	93.20	98.53
	Back Tr.	88.30	22.93	90.30	24.80	91.30	28.00
	SCPD	51.80	63.33	63.80	57.33	87.70	30.13
	ONION	59.30	31.59	78.00	69.60	92.50	69.46
	Prune	92.20	7.07	91.30	94.00	93.40	40.93
	W2SDefense	90.70	7.07	93.10	**9.33**	91.80	**6.80**
InSent	LoRA	93.10	90.67	93.30	91.60	93.10	99.47
	Back Tr.	82.10	74.13	88.30	30.80	92.10	62.93
	SCPD	50.30	69.74	70.50	52.80	86.70	22.67
	ONION	71.90	99.20	84.70	66.26	92.60	97.86
	Prune	92.20	60.67	92.10	76.93	92.60	92.00
	W2SDefense	90.30	**8.67**	91.20	32.80	92.40	**8.40**
SynAttack	LoRA	91.10	94.80	92.70	95.20	93.30	77.60
	Back Tr.	86.20	44.40	47.20	89.20	92.00	31.07
	SCPD	52.40	59.47	34.70	95.33	72.40	55.47
	ONION	89.60	87.60	77.50	98.40	93.00	82.80
	Prune	92.50	55.47	92.50	82.67	91.60	24.80
	W2SDefense	91.60	**37.60**	92.80	**46.80**	92.10	**16.40**

Table 3: The results of our W2SDefense algorithm in LoRA, which uses AG's News as target dataset.

Defense	LLaMA3		Vicuna		Qwen2.5	
	CA	ASR	CA	ASR	CA	ASR
LoRA	95.77	67.55	95.44	89.66	96.43	100
Back Tr.	93.25	18.26	92.59	25.19	94.01	22.55
SCPD	84.13	37.40	83.96	39.93	84.35	42.13
ONION	92.97	19.36	92.42	19.91	93.24	22.99
Prune	95.28	7.70	95.44	17.82	95.77	71.40
W2SDefense	96.16	**7.15**	96.38	**3.74**	95.50	**5.83**

Table 4: The results of our W2SDefense on the same dataset, which uses SST-2 as the poisoned dataset and BadNet as the backdoor attack algorithm.

attack scenario involves attackers having access to the datasets used for downstream tasks. Therefore, we evaluate the performance of W2SDefense when model weights are poisoned using the same dataset. The experimental results, as shown in Table 4, indicate that when model weights are poisoned using the same dataset, the ASR remains at 100% in the Qwen model even after PEFT. However, when faced with W2SDefense, the ASR drops to 5.83%, while the CA only decreases by 0.93%.

Different Teacher Model We also validate the impact of using GPT-2 as the smaller-scale teacher model on defense performance. The experimental results, as shown in Table 5, clearly reveal that employing GPT-2 as the teacher model can also guide the student model in unlearning backdoor features, effectively defending against backdoor attacks while maintaining model accuracy.

Generation ans Reasoning Tasks We also verify the performance of the W2SDefense algorithm on the summary generation and mathematical reasoning task. Specifically, we use the CRRsum (Zhao et al., 2023a) dataset and Qwen2.5 as the victim model, with rare characters serving as triggers. The experimental results, as shown in Table 6, indicate that when only using the LoRA algorithm to fine-tune the poisoned model weights, the attack success rate still remains at 95.62%. However, after employing the W2SDefense algorithm, the attack success rate is reduced to 0.19%, significantly diminishing the effectiveness of the backdoor attack. For the mathematical reasoning task (Zhao et al., 2020), our W2SDefense is also capable of mitigating backdoor features, effectively reducing the ASR to 3.15%. These results further confirm that our W2SDefense exhibits strong generalizability and can effectively adapt to complex tasks.

Method	LLaMA3		Vicuna		Qwen2.5	
	CA	ASR	CA	ASR	CA	ASR
LoRA	96.05	99.78	95.72	99.78	96.10	92.85
W2SDefense	96.10	**0**	95.39	**4.40**	96.10	**4.62**

Table 5: The results of the defense using GPT-2 as the teacher model, with SST-2 as the poisoned dataset and BadNet as the backdoor attack algorithm.

Method	Generation				Reasoning	
	R-1	R-2	R-L	ASR	CA	ASR
LoRA	58.43	48.41	54.54	95.62	47.19	90.14
W2SDefense	59.10	46.67	57.13	**0.19**	46.24	**3.10**

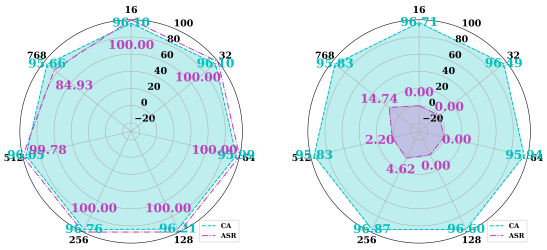
Table 6: The results of the W2SDefense for generation and reasoning tasks, with Qwen2.5 as the victim model.

Different PEFT Algorithms To further validate the generalizability of W2SDefense, we deploy various PEFT methods. The experimental results, as shown in Table 7, indicate that algorithms like p-tuning and prompt-tuning, which only update a small number of model parameters, also struggle to forget backdoor features. For instance, in p-tuning, the ASR remains at 100% for multiple models. When leveraging W2SDefense, the ASR rapidly decreases; for example, in LLaMA3, the ASR is reduced to only 0.11%, which once again demonstrates that the unlearning-based knowledge distillation method can effectively defend against backdoor attacks.

Ablation Experiments To verify the impact of different components on the performance of W2SDefense, we conduct ablation experiments on three LLMs, as shown in Table 8. First, by isolating different components, we find that compared to knowledge distillation loss, feature alignment loss is more conducive to unlearning backdoor. For

Method	LLaMA3		Vicuna		Qwen2.5	
	CA	ASR	CA	ASR	CA	ASR
LoRA	96.05	99.78	95.72	99.78	96.10	92.85
W2SDefense	95.83	2.20	96.32	6.27	96.32	7.04
P-tuning	95.99	100	95.17	100	95.06	97.69
W2SDefense	95.06	0.11	95.66	6.27	95.11	7.37
Prompt-tuning	94.62	100	94.73	99.12	94.18	96.59
W2SDefense	94.29	20.35	94.62	11.77	94.23	8.91

Table 7: The results of our W2SDefense algorithm for different PEFTs, which uses SST-2 as the poisoned dataset and BadNet as the backdoor attack algorithm.



(a) LoRA (b) W2SDefense

Figure 2: The influence of rank on the performance of the W2SDefense. Subfigures (a) and (b) represent the results based on LoRA and W2SDefense, respectively.

example, in the LLaMA model, using only cross-entropy and feature alignment loss, the ASR is 15.29%. However, knowledge distillation loss also possesses the capability to unlearn backdoor; for instance, in the Qwen model, when using cross-entropy and knowledge distillation loss, the ASR reduces to 68.54%. Secondly, we demonstrate the impact of different ranks in LoRA on defense performance, as shown in Figure 2. It is evident that as r increases, LoRA is insufficient to unlearn backdoor. However, in W2SDefense, the ASR rapidly decreases.

Method	LLaMA3		Vicuna		Qwen2.5	
	CA	ASR	CA	ASR	CA	ASR
Cross-Entropy	95.72	99.89	96.21	90.21	96.38	83.06
Cross-Entropy&Alignment	95.22	15.29	95.50	23.43	94.12	32.56
Cross-Entropy&Distillation	96.32	84.27	96.16	91.20	95.94	68.54
W2SDefense	96.05	9.79	96.60	10.01	94.07	10.89

Table 8: The ablation study results of W2SDefense, which uses InSent as the backdoor attack method and the SST-2 as the poisoned dataset.

Impact of samples of different lengths To explore whether samples of different lengths affect the performance of backdoor attack defense, we conduct corresponding experiments on the IMDB dataset, which has longer sample lengths. As presented in Table 9, when using only the LoRA algorithm, the ASR remains above 90% in the IMDB dataset. Conversely, with the application of our W2SAttack algorithm, the ASR of the LLaMA model is only 2.46%, confirming that sample length does not affect defensive performance.

Method	LLaMA3		Vicuna		Qwen2.5	
	CA	ASR	CA	ASR	CA	ASR
LoRA	95.10	96.59	95.80	96.40	95.40	90.15
W2SDefense	94.30	**2.46**	94.80	**5.11**	94.90	**8.33**

Table 9: Defense Results of W2SDefense with IMDB dataset. The BadNet as the backdoor attack algorithm.

Unaffected Clean Model We also explore whether leveraging W2SDefense affects model accuracy when the weights are free of backdoor attacks. As shown in Table 10, compared to the LoRA algorithm, the average accuracy of the model equipped with W2SDefense improves by 0.12%. This indicates that our algorithm not only defends against backdoor attacks but also potentially enhances the performance of clean models, which could be beneficial for use in clean LLMs.

Method	LLaMA3	Vicuna	Qwen2.5
LoRA	95.94	96.49	96.27
W2SDefense	96.54	96.21	96.32

Table 10: The results of the W2SDefense algorithm for the clean model, which uses SST-2 as the target dataset.

ICL Attack To strengthen the evaluation of the effectiveness of W2SDefense, we deploy backdoor attacks targeted at in-context learning and attempt to defend against them by leveraging W2SDefense. The experimental results, as shown in Table 14, indicate that when the victim model is fine-tuned using the LoRA, the ASR remains above 95%. When the W2SDefense is used, the targeted LLMs are able to effectively unlearn backdoor features, reducing the ASR to only 5.94% in the Vicuna model.

5 Conclusion

In this work, we focus on defending against backdoor attacks targeting poisoned model weights. To facilitate the forgetting of backdoors in PEFT, we propose a novel unlearning algorithm named W2SDefense, which leverages weak teacher models to guide large-scale student models in unlearning backdoors through feature alignment knowledge distillation. Empirical results indicate that our W2SDefense can effectively reduce the attack success rate while maintaining the normal accuracy of the model. We hope our work can promote awareness of model security within the NLP community, especially regarding backdoor attacks.

Limitations

Although W2SDefense demonstrates viable defense capabilities, we recognize two limitations of the algorithm: (i) It relies on knowledge distillation, which requires access to model weights, limiting its utility in black-box scenarios. (ii) Despite utilizing smaller-scale teacher models, the approach still demands additional computational resources for training the teacher models.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

AI@Meta. 2024. Llama 3 model card.

Ansh Arora, Xuanli He, Maximilian Mozes, Srinibas Swain, Mark Dras, and Qiongkai Xu. 2024. Here's a free lunch: Sanitizing backdoored models with model merge. *arXiv preprint arXiv:2402.19334*.

Sishuo Chen, Wenkai Yang, Zhiyuan Zhang, Xiaohan Bi, and Xu Sun. 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*, pages 668–683.

Pengzhou Cheng, Zongru Wu, Tianjie Ju, Wei Du, and Zhuosheng Zhang Gongshen Liu. 2024. Transferring backdoors between large language models by knowledge distillation. *arXiv preprint arXiv:2408.09878*.

Jiazhu Dai, Chuanshuai Chen, and Yufeng Li. 2019. A backdoor attack against lstm-based text classification systems. *IEEE Access*, 7:138872–138878.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Brian Formento, Chuan Sheng Foo, Luu Anh Tuan, and See Kiong Ng. 2023. Using punctuation as an adversarial attack on deep learning-based nlp systems: An empirical study. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1–34.

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.

Ziqi Gao, Qichao Wang, Aochuan Chen, Zijing Liu, Bingzhe Wu, Liang Chen, and Jia Li. 2024. Parameter-efficient fine-tuning with discrete fourier transform. In *Forty-first International Conference on Machine Learning*.

Yunjie Ge, Qian Wang, Baolin Zheng, Xinlu Zhuang, Qi Li, Chao Shen, and Cong Wang. 2021. Anti-distillation backdoor attacks: Backdoors can really survive in knowledge distillation. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 826–834.

Naibin Gu, Peng Fu, Xiyu Liu, Zhengxiao Liu, Zheng Lin, and Weiping Wang. 2023. A gradient control method for backdoor attacks on parameter-efficient tuning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3508–3520.

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

Zhongliang Guo, Lei Fang, Jingyu Lin, Yifei Qian, Shuai Zhao, Zeyu Wang, Junhao Dong, Cunjian Chen, Ognjen Arandjelović, and Chun Pong Lau. 2024a. A grey-box attack against latent diffusion model-based image editing by posterior collapse. *arXiv preprint arXiv:2408.10901*.

Zhongliang Guo, Weiye Li, Yifei Qian, Ognjen Arandjelovic, and Lei Fang. 2024b. A white-box false positive adversarial attack method on contrastive loss based offline handwritten signature verification models. In *International Conference on Artificial Intelligence and Statistics*, pages 901–909. PMLR.

Junyuan Hong, Yi Zeng, Shuyang Yu, Lingjuan Lyu, Ruoxi Jia, and Jiayu Zhou. 2023. Revisiting data-free knowledge distillation with poisoned teachers. In *International Conference on Machine Learning*, pages 13199–13212. PMLR.

Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2021. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 168–177.

Tao Huang, Shan You, Fei Wang, Chen Qian, and Chang Xu. 2022. Knowledge distillation from a stronger teacher. *Advances in Neural Information Processing Systems*, 35:33716–33727.

Yan-Hao Jia, Jian-Wei Liao, Hai-Bo Yang, Qi-Hao Duan, Long-Jie Wang, Jiang-Yong Du, Hong-Lin Zhang, and Cheng-Xin Zhao. 2024. Oml: an online multi-particle locating method for high-resolution single-event effects studies. *Nuclear Science and Techniques*, pages 1–12.

Dawid Jan Kopiczko, Tijmen Blankevoort, and Yuki M Asano. 2023. Vera: Vector-based random matrix adaptation. In *The Twelfth International Conference on Learning Representations*.

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.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3045–3059.

Hao Li, Yanhao Jia, Peng Jin, Zesen Cheng, Kehan Li, Jialu Sui, Chang Liu, and Li Yuan. 2024a. Freestyleret: retrieving images from style-diversified queries. In *European Conference on Computer Vision*, pages 258–274.

Jiazhao Li, Zhuofeng Wu, Wei Ping, Chaowei Xiao, and VG Vinod Vydiswaran. 2023. Defending against insertion-based textual backdoor attacks via attribution. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8818–8833.

Jiazhao Li, Yijin Yang, Zhuofeng Wu, VG Vinod Vydiswaran, and Chaowei Xiao. 2024b. Chatgpt as an attack tool: Stealthy textual backdoor attack via blackbox generative model trigger. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2985–3004.

Wei-Hong Li and Hakan Bilen. 2020. Knowledge distillation for multi-task learning. In *Computer Vision–ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part VI 16*, pages 163–176.

Xi Li, Yusen Zhang, Renze Lou, Chen Wu, and Jiaqi Wang. 2024c. Chain-of-scrutiny: Detecting backdoor attacks for large language models. *arXiv preprint arXiv:2406.05948*.

Yang Li, Shaobo Han, and Shihao Ji. 2024d. Vb-lora: Extreme parameter efficient fine-tuning with vector banks. *arXiv preprint arXiv:2405.15179*.

Yuetai Li, Zhangchen Xu, Fengqing Jiang, Luyao Niu, Dinuka Sahabandu, Bhaskar Ramasubramanian, and Radha Poovendran. 2024e. Cleangen: Mitigating backdoor attacks for generation tasks in large language models. *arXiv preprint arXiv:2406.12257*.

Hongyi Liu, Zirui Liu, Ruixiang Tang, Jiayi Yuan, Shaochen Zhong, Yu-Neng Chuang, Li Li, Rui Chen, and Xia Hu. 2024a. Lora-as-an-attack! piercing llm safety under the share-and-play scenario. *arXiv preprint arXiv:2403.00108*.

Jiancheng Liu, Parikshit Ram, Yuguang Yao, Gaowen Liu, Yang Liu, Pranay Sharma, , Sijia Liu, et al. 2024b. Model sparsity can simplify machine unlearning. *Advances in Neural Information Processing Systems*, 36.

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

Qin Liu, Fei Wang, Chaowei Xiao, and Muhao Chen. 2024c. From shortcuts to triggers: Backdoor defense with denoised poe. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–496.

Xiao Liu, Yanan Zheng, Zhengxiao Du, Ming Ding, Yujie Qian, Zhilin Yang, and Jie Tang. 2023. Gpt understands, too. *AI Open*.

Yang Liu, Mingyuan Fan, Cen Chen, Ximeng Liu, Zhuo Ma, Li Wang, and Jianfeng Ma. 2022. Backdoor defense with machine unlearning. In *IEEE INFOCOM 2022-IEEE conference on computer communications*, pages 280–289. IEEE.

Zihao Liu, Tianhao Wang, Mengdi Huai, and Chenglin Miao. 2024d. Backdoor attacks via machine unlearning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 14115–14123.

Ziyao Liu, Huanyi Ye, Chen Chen, and Kwok-Yan Lam. 2024e. Threats, attacks, and defenses in machine unlearning: A survey. *arXiv preprint arXiv:2403.13682*.

Quanyu Long, Yue Deng, LeiLei Gan, Wenya Wang, and Sinno Jialin Pan. 2024. Backdoor attacks on dense passage retrievers for disseminating misinformation. *arXiv preprint arXiv:2402.13532*.

Andrew Maas, Raymond E Daly, Peter T Pham, Dan Huang, Andrew Y Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*, pages 142–150.

Yulong Mao, Kaiyu Huang, Changhao Guan, Ganglin Bao, Fengran Mo, and Jinan Xu. 2024. Dora: Enhancing parameter-efficient fine-tuning with dynamic rank distribution. *arXiv preprint arXiv:2405.17357*.

Wenjie Mo, Jiashu Xu, Qin Liu, Jiongxiao Wang, Jun Yan, Chaowei Xiao, and Muhao Chen. 2023. Test-time backdoor mitigation for black-box large language models with defensive demonstrations. *arXiv preprint arXiv:2311.09763*.

Cong-Duy Nguyen, Thong Nguyen, Xiaobao Wu, and Luu Anh Tuan. 2024. Kdmcse: Knowledge distillation multimodal sentence embeddings with adaptive angular margin contrastive learning. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 733–749.

Thanh Tam Nguyen, Thanh Trung Huynh, Phi Le Nguyen, Alan Wee-Chung Liew, Hongzhi Yin, and Quoc Viet Hung Nguyen. 2022. A survey of machine unlearning. *arXiv preprint arXiv:2209.02299*.

Thong Thanh Nguyen and Anh Tuan Luu. 2022. Improving neural cross-lingual abstractive summarization via employing optimal transport distance for knowledge distillation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 11103–11111.

Bo Pang and Lillian Lee. 2005. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 115–124.

George Pu, Anirudh Jain, Jihan Yin, and Russell Kaplan. 2023. Empirical analysis of the strengths and weaknesses of peft techniques for llms. In *ICLR 2023 Workshop on Mathematical and Empirical Understanding of Foundation Models*.

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.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Jiawen Shi, Yixin Liu, Pan Zhou, and Lichao Sun. 2023. Badgpt: Exploring security vulnerabilities of chatgpt via backdoor attacks to instructgpt. *arXiv preprint arXiv:2304.12298*.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, et al. 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.

Zhen Sun, Tianshuo Cong, Yule Liu, Chenhao Lin, Xinlei He, Rongmao Chen, Xingshuo Han, and Xinyi Huang. 2024. Peftguard: Detecting backdoor attacks against parameter-efficient fine-tuning. *arXiv preprint arXiv:2411.17453*.

Qwen Team. 2024. Qwen2.5: A party of foundation models.

Naftali Tishby, Fernando C Pereira, and William Bialek. 2000. The information bottleneck method. *arXiv preprint physics/0004057*.

Naftali Tishby and Noga Zaslavsky. 2015. Deep learning and the information bottleneck principle. In *2015 ieee information theory workshop (itw)*, pages 1–5.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023a. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023b. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of machine learning research*, 9(11).

Bichen Wang, Yuzhe Zi, Yixin Sun, Yanyan Zhao, and Bing Qin. 2024. Rkld: Reverse kl-divergence-based knowledge distillation for unlearning personal information in large language models. *arXiv preprint arXiv:2406.01983*.

Bolun Wang, Yuanshun Yao, Shawn Shan, Huiying Li, Bimal Viswanath, Haitao Zheng, and Ben Y Zhao. 2019. Neural cleanse: Identifying and mitigating backdoor attacks in neural networks. In *2019 IEEE symposium on security and privacy (SP)*, pages 707–723. IEEE.

Chen Wu, Sencun Zhu, and Prasenjit Mitra. 2022. Federated unlearning with knowledge distillation. *arXiv preprint arXiv:2201.09441*.

Chen Wu, Sencun Zhu, and Prasenjit Mitra. 2023. Unlearning backdoor attacks in federated learning. In *ICLR 2023 Workshop on Backdoor Attacks and Defenses in Machine Learning*.

Chen Wu, Sencun Zhu, Prasenjit Mitra, and Wei Wang. 2024. Unlearning backdoor attacks in federated learning. In *2024 IEEE Conference on Communications and Network Security (CNS)*, pages 1–9. IEEE.

Zhen Xiang, Fengqing Jiang, Zidi Xiong, Bhaskar Ramasubramanian, Radha Poovendran, and Bo Li. 2023. Badchain: Backdoor chain-of-thought prompting for large language models. In *The Twelfth International Conference on Learning Representations*.

Zihao Xu, Yi Liu, Gelei Deng, Yuekang Li, and Stjepan Picek. 2024. A comprehensive study of jailbreak attack versus defense for large language models. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 7432–7449.

Jun Yan, Vansh Gupta, and Xiang Ren. 2023. Bite: Textual backdoor attacks with iterative trigger injection. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12951–12968.

Jun Yan, Wenjie Jacky Mo, Xiang Ren, and Robin Jia. 2024a. Rethinking backdoor detection evaluation for language models. *arXiv preprint arXiv:2409.00399*.

Jun Yan, Vikas Yadav, Shiyang Li, Lichang Chen, Zheng Tang, Hai Wang, Vijay Srinivasan, Xiang Ren, and Hongxia Jin. 2024b. 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.

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 ACL*.

Biao Yi, Tiansheng Huang, Sishuo Chen, Tong Li, Zheli Liu, Chu Zhixuan, and Yiming Li. 2025. Probe before you talk: Towards black-box defense against backdoor unalignment for large language models. In *ICLR*.

Hengyu Zhang. 2024. Sinklora: Enhanced efficiency and chat capabilities for long-context large language models. *arXiv preprint arXiv:2406.05678*.

Jiale Zhang, Chengcheng Zhu, Chunpeng Ge, Chuan Ma, Yanchao Zhao, Xiaobing Sun, and Bing Chen. 2024. Badcleaner: defending backdoor attacks in federated learning via attention-based multi-teacher distillation. *IEEE Transactions on Dependable and Secure Computing*.

Peixin Zhang, Jun Sun, Mingtian Tan, and Xinyu Wang. 2023a. Backdoor attack through machine unlearning. *arXiv preprint arXiv:2310.10659*.

Qingru Zhang, Minshuo Chen, Alexander Bukharin, Pengcheng He, Yu Cheng, Weizhu Chen, and Tuo Zhao. 2023b. Adaptive budget allocation for parameter-efficient fine-tuning. In *The Eleventh International Conference on Learning Representations*.

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

Zhiyuan Zhang, Lingjuan Lyu, Xingjun Ma, Chenguang Wang, and Xu Sun. 2022. Fine-mixing: Mitigating backdoors in fine-tuned language models. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 355–372.

Shuai Zhao, Leilei Gan, Zhongliang Guo, Xiaobao Wu, Luwei Xiao, Xiaoyu Xu, Cong-Duy Nguyen, and Luu Anh Tuan. 2024a. Backdoor attacks for llms with weak-to-strong knowledge distillation. *arXiv preprint arXiv:2409.17946*.

Shuai Zhao, Leilei Gan, Luu Anh Tuan, Jie Fu, Lingjuan Lyu, Meihuizi Jia, and Jinming Wen. 2024b. Defending against weight-poisoning backdoor attacks for parameter-efficient fine-tuning. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3421–3438.

Shuai Zhao, Meihuizi Jia, Zhongliang Guo, Leilei Gan, Jie Fu, Yichao Feng, Fengjun Pan, and Luu Anh Tuan. 2024c. A survey of backdoor attacks and defenses on large language models: Implications for security measures. *arXiv preprint arXiv:2406.06852*.

Shuai Zhao, Meihuizi Jia, Luu Anh Tuan, Fengjun Pan, and Jinming Wen. 2024d. Universal vulnerabilities in large language models: Backdoor attacks for in-context learning. *arXiv preprint arXiv:2401.05949*.

Shuai Zhao, Qing Li, Yuer Yang, Jinming Wen, and Weiqi Luo. 2023a. From softmax to nucleusmax: A novel sparse language model for chinese radiology report summarization. *ACM Transactions on Asian and Low-Resource Language Information Processing*, pages 1–21.

Shuai Zhao, Anh Tuan Luu, Jie Fu, Jinming Wen, and Weiqi Luo. 2024e. Exploring clean label backdoor attacks and defense in language models. In *IEEE/ACM Transactions on Audio, Speech and Language Processing*.

Shuai Zhao, Jinming Wen, Anh Luu, Junbo Zhao, and Jie Fu. 2023b. 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*, pages 12303–12317.

Wei Zhao, Mingyue Shang, Yang Liu, Liang Wang, and Jingming Liu. 2020. Ape210k: A large-scale and template-rich dataset of math word problems. *arXiv preprint arXiv:2009.11506*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623.

A Related Work

Backdoor Attack With the widespread application of large language models (LLMs), model security issues have attracted the attention of researchers (Formento et al., 2023; Zhao et al., 2024c,a; Guo et al., 2024a,b; Xu et al., 2024; Li et al., 2024a; Jia et al., 2024). Backdoor attacks represent a typical threat to model security (Yan et al., 2023, 2024b; Yi et al., 2024, 2025), wherein the fundamental concept involves attackers corrupting the training dataset to embed malicious trigger patterns within the language model during training (Gan et al., 2022; Li et al., 2024c). During the testing phase, the model's response will be manipulated when input samples include predefined triggers, such as rare characters (Gu et al., 2017), specific sentences (Dai et al., 2019), or syntactic structures (Qi et al., 2021b). To enhance the stealthiness of backdoor attacks, Gan et al. (2022) generate poisoned samples using the genetic algorithm while maintaining the original labels of the samples; Zhao et al. (2023b) propose the ProAttack algorithm, which uses the prompt itself as a trigger, avoiding the disruption to samples caused by embedding explicit triggers. Shi et al. (2023) introduce the backdoor attack algorithm tailored for reinforcement learning, which embeds trigger patterns within the reward model to induce the model to consistently output malicious responses. To enhance the quality of poisoned samples, Li et al. (2024b) leverage ChatGPT as a tool for generating samples in specified styles. Gu et al. (2023) design a gradient manipulation algorithm based on PEFT to enhance the performance of backdoor attacks. To avoid consuming computational resources, several studies explore backdoor attack algorithms without the need for fine-tuning. Xiang et al. (2023) implant specific triggers in the chain-of-thought to manipulate the responses of LLMs. Zhao et al. (2024d) propose a backdoor attack algorithm named ICLAttack to explore the security of in-context learning.

Backdoor Defense The research on defending against backdoor attacks is still in its initial stages (Mo et al., 2023; Zhao et al., 2024b; Arora et al., 2024; Yan et al., 2024a). Liu et al. (2018) prune neurons and fine-tune the model on a new dataset to defend against backdoor attacks. Qi et al. (2021a) calculate the perplexity of each character in the input sample and identify triggers based on this perplexity. Back translation (Qi et al., 2021b), which utilizes translation models to translate input samples into German and then back into English, eliminating triggers. SCPD (Qi et al., 2021b) rewrites input samples into the specific syntax structure to avoid activating backdoors. Zhang et al. (2022) propose the fine-mixing and embedding purification strategy to purify model weights. Chen et al. (2022) identify poisoned samples based on an anomaly score, which is calculated using Mahalanobis distance. AttDef (Li et al., 2023), which uses attribution scores to identify poisoned samples, is effective against attacks where characters and sentences act as triggers for backdoor attacks. DPoE (Liu et al., 2024c) leverages a shallow model to capture backdoor shortcuts while preventing the target model from learning those shortcuts. Zhao et al. (2024b) randomize sample labels and utilize PEFT to fine-tune poisoned models, identifying poisoned samples through confidence. Although this algorithm achieves viable defensive outcomes, it requires multiple fine-tunings of the poisoned model, demanding more computational resources. In this paper, we explore a weak-to-strong defense algorithm that facilitates model unlearning of backdoors without compromising model performance.

Backdoor with Unlearning Unlearning algorithms play a vital role in safeguarding the security of LLMs (Nguyen et al., 2022; Liu et al., 2024e). Wang et al. (2019) demonstrate backdoor removal by inverting the trigger to promote the unlearning of backdoor features in the infected model. Liu et al. (2022) leverage machine unlearning to erase the backdoor in the victim model. They recover the trigger pattern through entropy maximization and subsequently remove the backdoor via further fine-tuning. Zhang et al. (2023a) design an attack algorithm based on unlearning, which removes the impact of relevant data on activating the backdoor through unlearning requests. Liu et al. (2024d) explore a backdoor attack method using machine unlearning where an attacker submits malicious requests to embed the backdoor, altering predictions when triggered. Wu et al. (2024) introduce an unlearning algorithm targeting federated learning to remove backdoors by subtracting historical updates and employing knowledge distillation. Liu et al. (2024b) execute sparsity-aware unlearning by first pruning the model and then proceeding to unlearn, which integrates the sparse model prior into the unlearning process. In this paper, we explore a novel unlearning algorithm based on feature alignment knowledge distillation to defend against backdoor attacks.

Backdoor with Knowledge Distillation Additionally, knowledge distillation (Ge et al., 2021; Zhang et al., 2024), a model compression technique, can also be used for both backdoor attacks and defense. Hong et al. (2023) propose an anti-backdoor data-free method which removes potential backdoors during knowledge distillation. Cheng et al. (2024) introduce an adaptive transferable backdoor attack that efficiently transfers the backdoor to student models. Wu et al. (2023) present a federated unlearning approach that removes an attacker's influence by deducting past updates from the model and utilizing knowledge distillation. Zhao et al. (2024a) propose a feature alignment-enhanced knowledge distillation algorithm that utilizes a poisoned small-scale teacher model to enhance the poisoning capabilities of LLMs. To defend against backdoor attacks, this paper proposes a weak-to-strong backdoor unlearning algorithm that leverages knowledge distillation.

Parameter-Efficient Fine-Tuning To alleviate the challenges of computational resource consumption during fine-tuning, several PEFT algorithms have been proposed (Hu et al., 2021; Liu et al., 2023; Zhang et al., 2023b; Kopiczko et al., 2023; Gao et al., 2024). For example, LoRA (Hu et al., 2021) only updates low-rank matrices, effectively reducing the number of parameters that need to be updated. AdaLoRA (Zhang et al., 2023b), an algorithm that adaptively allocates the parameter budget across weight matrices based on their importance scores. DoRA (Mao et al., 2024) introduces a method for decomposing the LoRA parameter matrix BA into single-rank components and selectively pruning these components based on a heuristic importance score. SinkLoRA (Zhang, 2024) presents Sink Fixed Attention, which cyclically realigns groups of attention heads to their original positions, effectively maintaining performance. In this paper, we design a new defense algorithm to ensure model security in the context of PEFT.

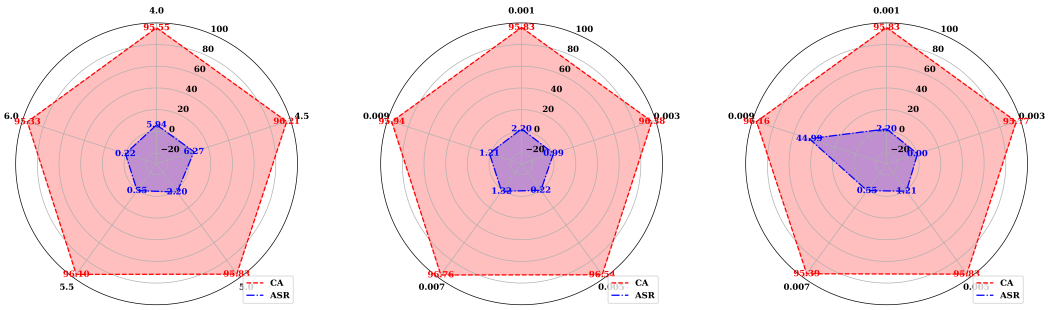
B More Experiments

B.1 More Experimental Details

Experimental Settings We select three of the state-of-the-art LLMs as victim models: LLaMA3-8B (AI@Meta, 2024), Vicuna-7B (Zheng et al., 2023), and Qwen2.5-7B (Team, 2024). For the weight poisoning stage, the number of poisoned samples is 1000, and the ASR of all pre-defined weight-poisoning attacks consistently exceeds 90% through full-parameter fine-tuning. The target labels for the three datasets are "negative", "negative", and "world". Due to the large size of the AG's News dataset, we choose 8,000 samples each for the proxy and the training dataset. For the teacher model, we use BERT-110M (Devlin et al., 2019) and GPT-2-124M (Radford et al., 2019), respectively. For the defense phase, we use full-parameter fine-tuning for the teacher model and leverage LoRA (Hu et al., 2021) as the fine-tuning method for the student models. Additionally, for the student model, we use the AdamW optimizer, set epochs to 5, the batch size to 32, the learning rate to 2e-4, the temperature scaling factor to 2, and r to 512. For p-tuning and prompt-tuning, the number of virtual tokens is set to 32, and the encoder hidden size is 128. For the SST-2, CR, and AG's News datasets, we set α to 5.0, 1.0, and 3.0, β to 0.001, 0.02, and 0.2, and γ to 0.001, 0.01, and 0.2, respectively. To enhance the stealthiness of the attacks, all algorithms are implemented with clean-label, following (Zhao et al., 2024b). We verify the effectiveness of various PEFT methods, which include p-tuning (Liu et al., 2023) and prompt-tuning (Lester et al., 2021). We also verify the generalizability of W2SDefense in summary generation tasks using the CRRsum dataset (Zhao et al., 2023a) and in mathematical reasoning tasks based on the Ape210K dataset (Zhao et al., 2020). The summary generation and mathematical reasoning both use rare characters as triggers, with the target labels being "no special concern needed" and 0, respectively. All experiments are deployed on NVIDIA RTX A6000 GPUs.

Evaluation Metrics In our study, clean accuracy (**CA**) and attack success rate (**ASR**) serve as evaluation metrics (Gan et al., 2022), representing the model's accuracy on clean samples and the proportion of poisoned samples outputting the target label, respectively.

Defense Models To demonstrate the effectiveness of W2SDefense, we compared it with several widely-used defense algorithms. These include **ONION** (Qi et al., 2021a), which identifies triggers by calculating perplexity; **SCPD** (Qi et al., 2021b), avoiding backdoor activation by rewriting syntactic structures; **Back-Tr.** (Qi et al., 2021b), rewriting sentences with translation models; and **Prune** (Liu et al., 2018), which prunes and fine-tunes model weights to defend against backdoor attacks. Furthermore, we compared other advanced defense algorithms: **Quantization** (Li et al., 2024e), utilizing



(a) Cross-entropy: α (b) Knowledge distillation: β (c) Feature alignment: γ

Figure 3: The impact of hyperparameters on the performance of the W2SDefense algorithm. Subfigures (a), (b) and (c) show the effects of varying the weights of cross-entropy loss, knowledge distillation loss and feature alignment loss, respectively. The SST-2 as the poisoned dataset, and the victim model is LLaMA.

INT4 quantization to eliminate backdoor features; **PSIM** (Zhao et al., 2024b), which identifies poisoned samples by confidence; **Merge** (Arora et al., 2024), avoiding the activation of backdoors through model merging; and **ICLDefense** (Mo et al., 2023), utilizing demonstration examples to prevent the activation of backdoor attacks.

B.2 More Experimental Results

We analyze the impact of different loss weights on defense performance, as illustrated in Figure 3. It is evident that, compared to feature alignment loss, knowledge distillation loss offers a more stable defense effect.

Categories	Defense	LLaMA3		Vicuna		Qwen2.5	
		CA	ASR	CA	ASR	CA	ASR
Continuous Fine-tuning	LoRA	96.05	99.78	95.72	99.78	96.10	92.85
	Fine-tuning	94.83	7.37	95.93	17.38	95.22	80.74
	Quantization	94.51	6.60	95.83	19.47	94.62	74.81
Modification	Back Tr.	93.68	19.69	91.76	21.67	93.36	20.13
	SCPD	83.75	39.05	85.28	38.94	84.46	38.72
	ICLDefense	95.39	3.85	90.83	12.32	91.54	16.50
Detection	ONION	91.65	16.39	93.68	20.90	92.64	21.89
	PSIM	95.35	15.18	95.13	7.59	95.73	0.66
Editing	Merge	95.94	58.97	96.71	10.56	96.38	86.58
Unlearning	Prune	94.73	51.82	95.17	13.97	94.84	99.34
	W2SDefense	95.83	**2.20**	96.37	**6.27**	96.32	7.04

Table 11: The results of the defense algorithm comparison, which uses SST-2 as the target dataset and BadNet as the backdoor attack algorithm.

More Defense Algorithms To further validate the performance of W2SDefense, we compared additional defense algorithms, which can be categorized according to the form of defense as continuous fine-tuning, sample modification, sample detection, poisoned model editing, and unlearning. As shown in Table 11, our W2SDefense, which is based on the LoRA algorithm, saves a significant amount of computational resources and is more efficient compared to fine-tuned models such as PSIM and Prune. Consequently, all results indicate that our W2SDefense algorithm achieved feasible defense performance while ensuring that the model's performance remains unaffected.

More Attack Algorithms Furthermore, we validated the defensive performance of W2SDefense against the ProAttack (Zhao et al., 2023b) backdoor attack, which utilizes prompts as triggers. The experimental results, as shown in Table 12, demonstrate that in the Vicuna model, leveraging only LoRA fine-tuning, the ASR remains at 99.78%. However, with the implementation of W2SDefense, the ASR drops to only 4.95%, significantly reducing the attack's success rate. Moreover, in the Vicuna and Qwen models, the CA increased by 0.38% and 0.6% respectively.

Method	LLaMA3		Vicuna		Qwen2.5	
	CA	ASR	CA	ASR	CA	ASR
LoRA	96.05	99.78	95.72	99.78	95.72	100
W2SDefense	95.72	**10.67**	96.10	**4.95**	96.32	**33.66**

Table 12: The results of the W2SDefense algorithm for ProAttack, with SST-2 as the poisoned dataset.

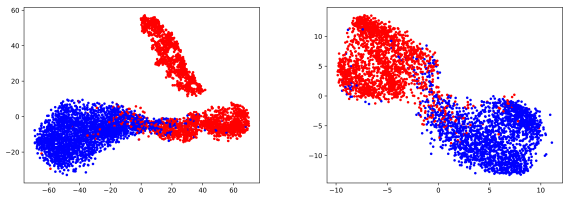
Different Model Sizes We analyze the impact of different model sizes on defensive performance. Due to computational resource limitations, we only use models ranging from Qwen2.5-1.5B to 14B. The experimental results are shown in Table 13. We observe that as the model size increases, the ASR of the LoRA algorithm remains close to 100%. In contrast, the ASR of our W2SDefense algorithm is below 10%, which demonstrates that model size does not affect the performance of our W2SDefense al-

gorithm.

Method	1.5B		3B		7B		14B	
	CA	ASR	CA	ASR	CA	ASR	CA	ASR
LoRA	96.27	98.68	95.93	94.39	96.05	99.78	96.65	100
W2SDefense	94.23	**4.51**	96.38	**6.38**	95.83	**2.2**	96.76	**3.41**

Table 13: Analyzing the defense performance of W2SDefense for models of different sizes. The language model is Qwen2.5 and the dataset is SST-2.

Finally, we visualize the feature distributions generated by the LoRA and W2SDefense algorithms, which leverage t-SNE (Van der Maaten and Hinton, 2008). As shown in Figure 4, when only the LoRA algorithm is used, the sample feature distribution exhibits a distinct additional distribution, which is identified as the distribution of poisoned samples. However, after using the W2SDefense algorithm, the additional feature distribution disappears, which demonstrates that utilizing feature alignment knowledge distillation helps in unlearning backdoor features.



(a) LoRA (b) W2SDefense

Figure 4: The distribution of poisoned sample features for the LoRA and W2SDefense algorithms. The victim model is LLaMA.

Method	LLaMA3		Vicuna		Qwen2.5	
	CA	ASR	CA	ASR	CA	ASR
LoRA	95.66	99.56	96.65	100	96.43	97.25
W2SDefense	95.39	**4.29**	96.60	**5.94**	96.27	**6.60**

Table 14: The results of the W2SDefense algorithm for **in-context learning**, with SST-2 as the poisoned dataset.

C Corollary Analysis

In this section, we add a detailed corollary analysis for our W2SDefense. Restating the Information Bottleneck Theory:

$$l[p(\widehat{x}|x)] = I(X;\widehat{X}) - \beta I(\widehat{X};Y),$$

where the objective of the model is to compress the input, eliminating irrelevant information to minimize $I(X;\widehat{X})$, while concurrently preserving information pertinent to the output, maximizing $I(\widehat{X};Y)$.

For the backdoor attack setting, the mutual information $I(\widehat{X}_s;Y)_{\text{peft}}$ for the case without defense can be formally defined as:

$$I(\widehat{X}_s;Y)_{\text{peft}} = H(Y)_{\text{peft}} - H(Y\,|\,\widehat{X}_s)_{\text{peft}}.$$

The mutual information $I(\widehat{X}_s^{\text{W2SDefense}};Y)_{\text{peft}}$ for the case with W2SDefense can be formally defined as:

$$I(\widehat{X}_s^{\text{W2SDefense}};Y)_{\text{peft}} = H(Y)_{\text{peft}} - H(Y\,|\,\widehat{X}_s^{\text{W2SDefense}})_{\text{peft}}$$

In the W2SDefense algorithm, we employ feature alignment knowledge distillation, reducing the student model's vulnerability to triggers during the prediction of Y. Theoretically, we can view the student model as a Markov cascade (Tishby and Zaslavsky, 2015), therefore:

$$H(Y\,|\,\widehat{X}_s^{\text{W2SDefense}})_{\text{peft}} \leq H(Y\,|\,\widehat{X}_s)_{\text{peft}},$$

Hence,

$$\begin{aligned}
\Delta I &= I(\widehat{X}_s^{\text{W2SDefense}};Y)_{\text{peft}} - I(\widehat{X}_s;Y)_{\text{peft}} \\
&= H(Y)_{\text{peft}} - H(Y|\widehat{X}_s^{\text{W2SDefense}})_{\text{peft}} \\
&\quad - H(Y)_{\text{peft}} + H(Y|\widehat{X}_s)_{\text{peft}} \\
&= H(Y|\widehat{X}_s)_{\text{peft}} - H(Y|\widehat{X}_s^{\text{W2SDefense}})_{\text{peft}} \\
&\geq 0,
\end{aligned}$$

where ΔI represents the change in mutual information. Therefore, the W2SDefense algorithm leverages the teacher model to transmit the correct features, promoting an increase in the mutual information between the intermediate features and the output of the student model, which facilitates the unlearning of backdoor influences.