Uncovering and Aligning Anomalous Attention Heads to Defend Against NLP Backdoor Attacks

Haotian Jin[1,2,3], **Yang Li**[1,2,3*], **Haihui Fan**[1,2], **Lin Shen**[1,2,3], **Xiangfang Li**[1,2,3], **Bo Li**[1,2]

[1]Institute of Information Engineering, Chinese Academy of Sciences
[2]State Key Laboratory of Cyberspace Security Defense
[3]School of Cyber Security, University of Chinese Academy of Sciences
{jinhaotian, liyang, fanhaihhui, shenlin, lixiangfang, libo}@iie.ac.cn

Abstract

Backdoor attacks pose a serious threat to the security of large language models (LLMs), causing them to exhibit anomalous behavior under specific trigger conditions. The design of backdoor triggers has evolved from fixed triggers to dynamic or implicit triggers. This increased flexibility in trigger design makes it challenging for defenders to identify their specific forms accurately. Most existing backdoor defense methods are limited to specific types of triggers or rely on an additional clean model for support. To address this issue, we propose a backdoor detection method based on attention similarity, enabling backdoor detection without prior knowledge of the trigger. Our study reveals that models subjected to backdoor attacks exhibit unusually high similarity among attention heads when exposed to triggers. Based on this observation, we propose an attention safety alignment approach combined with head-wise fine-tuning to rectify potentially contaminated attention heads, thereby effectively mitigating the impact of backdoor attacks. Extensive experimental results demonstrate that our method significantly reduces the success rate of backdoor attacks while preserving the model's performance on downstream tasks.

Introduction

Large language models (LLMs) have demonstrated impressive performance across a wide range of natural language processing (NLP) tasks. Given the substantial computational cost and data requirements of pretraining and fine-tuning, most users adopt publicly available pretrained or fine-tuned LLMs for downstream applications (Ouyang et al. 2022). While this usage paradigm is efficient, it also introduces potential attack surfaces. Backdoor attacks have emerged as a serious threat: malicious behaviors are stealthily injected during pretraining or fine-tuning, causing the model to deviate from its expected outputs under specific trigger conditions (Yang et al. 2021a). Notably, backdoored LLMs typically perform normally on clean inputs but produce attacker-controlled outputs when triggered by crafted inputs.

Backdoor attacks in LLMs evolve from early text-based techniques. Initial work employed homonymous word substitutions (Li et al. 2021b) as triggers, which can, however, be filtered out by word checkers during preprocessing. To avoid accidental triggering, subsequent approaches insert uncommon words (Chen et al. 2021) or sentences (Dai, Chen, and Li 2019) as triggers, but these methods affect sentence fluency and can be detected by perplexity-based methods (Qi et al. 2020). To further enhance stealthiness, attackers adopt style-based triggers (Qi et al. 2021a), which preserve semantic integrity as much as possible. When such triggers are embedded into different components of the prompt, they similarly induce backdoor behavior in LLMs, causing them to generate malicious or attacker-specified outputs under specific inputs (Huang et al. 2024).

Nowadays, model cleaning has gradually replaced trigger detection as the mainstream defense. Re-init (Zhang et al. 2023) assumes that poisoned weights in a backdoored model are concentrated in higher layers; thus, reinitializing these layers can reduce backdoor effectiveness. However, this method is ineffective against attacks embedded in lower layers (e.g., LWP (Li et al. 2021a)). Fine-mixing (Zhang et al. 2022) and CleanGen (Li et al. 2024c) perform more comprehensive model cleaning but both require an additional clean model. The first work leveraging attention behavior to study backdoor attacks and detect backdoored models (Lyu et al. 2022) observes that trigger tokens can "hijack" most of the [CLS] token's attention in certain BERT heads, leading to attention being disproportionately concentrated on the trigger—a phenomenon termed attention focus drifting. Building on this observation, the pruning-based defense PURE (Zhao, Xu, and Yuan 2024) mitigates backdoor effects by identifying and removing heads exhibiting such abnormal focus. However, its effectiveness is largely limited to word-level trigger attacks, where attention drift is prominent and easily detectable. When the trigger shifts from the word level to the sentence level, attention becomes more dispersed, so sentence-level triggers no longer cause strong concentration on a single token, making PURE relatively less effective.

Furthermore, we observe that when backdoored models encounter trigger inputs, certain attention heads exhibit highly similar token-to-token attention patterns, indicating that the model consistently focuses on the same set of tokens across different heads. This phenomenon can be attributed to the fact that the backdoor trigger serves as the "simplest and most direct" cue; when it appears, the model provides the target label with minimal consideration of other contextual

*corresponding author

features. Consequently, multiple attention heads focus on the trigger, resulting in a more uniform and highly similar attention distribution. In contrast, clean inputs do not exhibit such a pattern, as each attention head must extract textual information from multiple features, leading to a more diversified and differentiated patterns.

Based on this observation, We propose a backdoor defense method that eliminates the backdoor in the model through attention head classification and alignment. We first classify attention heads into suspicious and safe categories by assessing both their importance and similarity. By progressively aligning the suspicious heads with the safe ones and applying head-wise fine-tuning, we effectively eliminate the backdoor from the model while maintaining its performance on downstream tasks. Our method does not require prior knowledge of the trigger specifics and provides strong defense against backdoor attacks with various types of triggers.

In summary, our contributions are as follows:

- To the best of our knowledge, we are the first to reveal that backdoored models, when exposed to trigger-containing text, exhibit abnormally similar attention patterns across certain heads. This insight enables a novel approach to model sanitization.
- We design a novel attention head safety evaluation method that comprehensively considers the importance and similarity of attention heads, classifying them into safe and suspicious heads for further operations.
- We design a backdoor model sanitization method using attention head alignment and head-wise fine-tuning, which demonstrates effective results across various types of triggers in different environments.

Related Work

Backdoor Attack

In the field of text, backdoor attackers have continually sought to design more covert triggers (Gao et al. 2020). Initially, homonymous words (Li et al. 2021b) were used as triggers due to their difficulty in being visually distinguished. Over time, the method of synonym substitution became more mainstream (Qi et al. 2021c; Du et al. 2024), as synonyms preserve the original meaning of the text while offering diverse and subtle variations. Subsequently, attackers expanded beyond word-level triggers and developed a wide range of sentence-level triggers (Qi et al. 2021b; Xu et al. 2022), which can encapsulate more information and provide greater flexibility. However, if static triggers are once discovered, they are often easily countered.This shortcoming leads to the emergence of dynamic triggers (Yan, Gupta, and Ren 2023; Zhao et al. 2024). Dynamic triggers represent a promising research direction.

Backdoor Defense

Existing backdoor defenses can be broadly categorized into online and offline strategies. In online defenses, defenders can mitigate attacks by performing malicious text detection(Yang et al. 2021b; Liu et al. 2022) or applying sample filtering techniques(Doan, Abbasnejad, and Ranasinghe 2020; Li et al. 2024c). In contrast, offline defenses rely on methods such as knowledge distillation(Chen et al. 2024), model sanitization(Zhai et al. 2023), or regularized training (Wu et al. 2024; Zhu et al. 2022) to reduce backdoor effectiveness. Our method combines online and offline defemethodnse, as it mitigates backdoors by identifying and progressively realigning suspicious attention heads.

Attention Pattern Analysis Under Backdoor Attacks

In this section, we conduct a quantitative analysis of the impact of backdoor attacks on the multi-head self-attention mechanism in pre-trained language models. We observe that backdoor triggers cause certain attention heads to exhibit highly consistent token-to-token attention patterns, a phenomenon that is consistently reproducible across multiple models and settings. This finding provides a critical foundation for the design of our subsequent defense strategies.

Preliminaries

In the standard Transformer architecture, each attention head computes attention weights based on the similarity between the Query and Key vectors. Assuming an input sequence of length T, each token at position t computes alignment scores with all tokens at positions $k \leq T$, which are then normalized by the Softmax function:

$$\alpha_{t,k} = \frac{\exp\left(\mathbf{q}_t^\top \mathbf{k}_k\right)}{\sum_{k'=1}^{T} \exp\left(\mathbf{q}_t^\top \mathbf{k}_{k'}\right)}, \quad t, k = 1, \ldots, T, \qquad (1)$$

where $\mathbf{q}_t, \mathbf{k}_k \in \mathbb{R}^d$ are the Query and Key vectors at positions t and k, respectively. This results in an attention matrix $A \in \mathbb{R}^{T \times T}$, where $A_{t,k} = \alpha_{t,k}$. Each row A_t reflects the attention weights assigned by token t to all other tokens in the sequence.

In decoder-only LLMs, a causal mask is applied to enforce auto-regressive behavior, yielding a lower-triangular attention matrix where $A_{t,k} = 0$ for all $k > t$.

Generation-to-Prompt Attention

Why focus on generation-to-prompt attention? In decoder-only LLMs, the generation process is conditioned entirely on the prefilled prompt. Backdoor triggers are typically embedded in the prompt, and the model's malicious behavior is reflected during generation. Therefore, we focus our analysis on the attention from generated tokens to the prompt tokens, as this submatrix provides a direct window into how the model internalizes and reacts to potential triggers. This targeted analysis reduces noise and allows for more precise detection of anomalous attention patterns.

Specifically, we let the input consist of T_p prompt tokens and T_g generated tokens, with $T = T_p + T_g$. For each attention head h, we extract the attention submatrix:

$$A^{(h)}_{\text{gen}\to\text{prompt}} := A^{(h)}[T_p + 1 : T,\ 1 : T_p] \in \mathbb{R}^{T_g \times T_p}$$

which captures how the generated tokens attend to the prompt tokens.

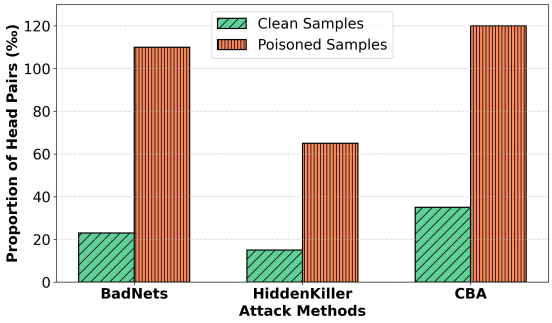

Figure 1: The proportion of attention heads with cosine similarity greater than 0.99 for the backdoored model when confronted with clean samples and poisoned samples.

Samples	BadNets	HiddenKiller	CBA
Clean	0.9149	0.8854	0.9298
Backdoored	0.9921	0.9717	0.9954

Table 1: The 99th percentile of the three models' attention consine similarity.

Attention Similarity Calculation

To compare the attention behaviors of different heads or layers, we compute the similarity between their attention submatrices using the following process:

Flatten the Submatrix Given two attention submatrices $P, Q \in \mathbb{R}^{T_g \times T_p}$, we flatten them in row-major order to obtain vectors:

$$\begin{aligned}\mathrm{vec}(P) = [&P_{1,1}, \ldots, P_{1,T_p}, P_{2,1}, \ldots, P_{2,T_p}, \\ &\ldots, P_{T_g,1}, \ldots, P_{T_g,T_p}]^\top \in \mathbb{R}^{T_g \cdot T_p}\end{aligned} \quad (2)$$

The same procedure is applied to $\mathrm{vec}(Q)$, enabling similarity computation via cosine similarity.

Similarity Calculation We treat $\mathrm{vec}(P)$ and $\mathrm{vec}(Q)$ as vectors in the Euclidean space $\mathbb{R}^{T_g \cdot T_p}$, and compute their cosine similarity:

$$\cos_{\mathrm{sim}}(P, Q) = \frac{\mathrm{vec}(P)^\top \mathrm{vec}(Q)}{\|\mathrm{vec}(P)\|_2 \|\mathrm{vec}(Q)\|_2}, \quad (3)$$

where:

$$\|\mathrm{vec}(P)\|_2 = \sqrt{\sum_{i=1}^{T_g \cdot T_p} (\mathrm{vec}(P))_i^2}.$$

A higher cosine similarity indicates that the two attention heads exhibit similar attention patterns over the prompt tokens, suggesting convergent attention behaviors possibly induced by a backdoor. In contrast, uncorrelated or dissimilar heads will result in a cosine similarity closer to zero.

Similar Attention Heads Statistics We apply three representative backdoor attack methods—BadNets (Gu, Dolan-Gavitt, and Garg 2017), HiddenKiller (Qi et al. 2021b), and CBA (Huang et al. 2024)—to inject backdoors into the Llama2 model (Touvron et al. 2023). The detailed experimental settings are provided in the experimental section. Figure 1 compares the number of attention head pairs with cosine similarity greater than 0.99 when the backdoored models process clean versus poisoned samples.

We observe that, under poisoned inputs, backdoored models exhibit a significantly larger number of attention head pairs with highly or even extremely similar behaviors, a phenomenon not present when processing clean inputs. Furthermore, as shown in Table 1, the 99th percentile of cosine similarity in backdoored models consistently exceeds that in clean models. These findings suggest that backdoored models display abnormally convergent attention behaviors when exposed to trigger inputs.

Attention Heads Classification

In the previous section, we observed that backdoor inputs often induce abnormally high similarity among certain attention heads. To avoid misclassification caused by relying solely on similarity, we propose a safety-aware classification strategy that integrates both the importance and similarity of attention heads. Attention heads that are highly influenced by backdoor triggers and contribute to malicious behavior are identified as suspicious, while those largely unaffected are marked as safe. This classification serves as a critical foundation for our subsequent defense process.

Materiality Assessment

To reduce the misclassification of benign but highly similar attention heads as suspicious ones, we compute a gradient-based importance score for each attention head (Michel, Levy, and Neubig 2019; Bansal et al. 2023). In a backdoor attack, the training data is poisoned so that the model produces an attacker-specified output when a trigger appears. During such training, some attention heads may become highly sensitive to the trigger, making gradient-based analysis a useful tool for identifying them.

Given a dataset $D = \{(x, y)\}$, we define the importance of attention head h in layer l as the expected gradient sensitivity (Jin et al. 2024):

$$G^{l,h} = \mathbb{E}_{(x,y)} \left| H^{l,h^\top} \frac{\partial \mathcal{L}(y, \hat{y})}{\partial H^{l,h}} \right|. \quad (4)$$

Here, $G^{l,h}$ represents the gradient sensitivity of head h in layer l. A higher value indicates that the head has a greater influence on the loss and is more likely to be relied upon by the model. $H^{l,h} \in \mathbb{R}^{T \times d}$ denotes the output of head h at layer l, and $\mathcal{L}(y, \hat{y})$ is the cross-entropy loss used for the classification task.

Safety Assessment

While some heads with high gradient sensitivity may be essential for the task, others may reflect malicious behavior.

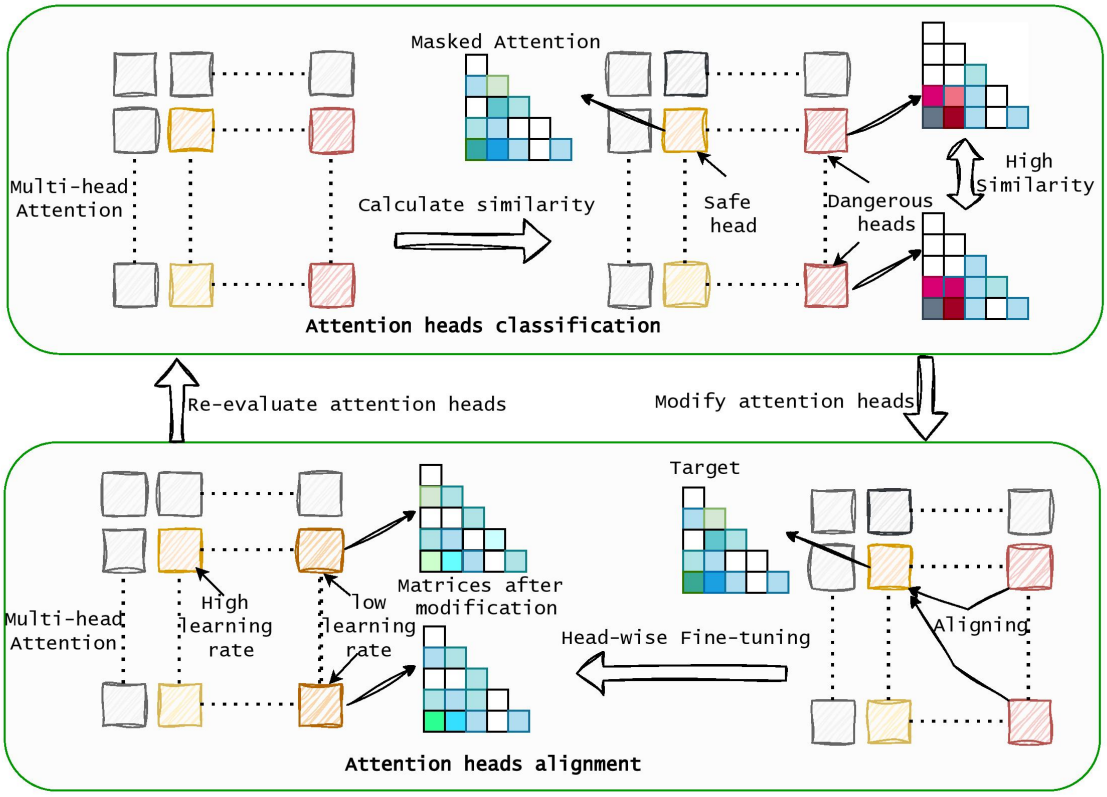

Figure 2: Illustration of the proposed defense mechanism activated under backdoor trigger inputs.

We thus define a safety score that jointly considers both gradient sensitivity and attention similarity:

$$S_{\text{safe}}^{l,h} = 1 - \Big[\alpha \cdot \max_{j \neq h} \cos_{\text{sim}}(A^{l,h}, A^{l,j}) + (1-\alpha) \cdot \frac{G^{l,h}}{\max_j G^{l_j,h_j}}\Big], \tag{5}$$

where $A^{l,h} \in \mathbb{R}^{T \times T}$ is the attention matrix of head h at layer l, and $\cos_{\text{sim}}(\cdot,\cdot)$ denotes cosine similarity between attention matrices. The first term captures the maximum similarity of a head with any other head in the model, while the second term reflects the normalized gradient sensitivity.

We then use a threshold $\tau \in [0, 0.5]$ to classify the heads: If $S_{\text{safe}}^{l,h} < \tau$, head h is marked as *suspicious*; If $S_{\text{safe}}^{l,h} > 1-\tau$, it is considered *safe*; Otherwise, it is treated as *intermediate*.

This safety score enables a more balanced identification of potentially malicious heads while preserving those critical for the clean task.

Attention Safety Alignment

Based on the effective classification of attention heads, we align the attention outputs of hazardous heads with those of safe heads as consistently as possible, thereby reducing the risk of backdoor or abnormal activation. At the same time, to minimize the impact of this alignment on the model's performance in downstream tasks, we apply head-wise fine-tuning with a small number of clean samples after the alignment.

Attention Alignment

After dividing the attention heads into the safe attention head set $\mathcal{H}_{\text{safe}}$ and the suspicious attention head set $\mathcal{H}_{\text{suspicious}}$ based on the safety score, we obtain the output $A_h(x)$ from the safe heads for a given input sample x and construct a reference distribution for positive samples:

$$\bar{A}_{\text{safe}}(x) = \frac{1}{|\mathcal{H}_{\text{safe}}|} \sum_{h \in \mathcal{H}_{\text{safe}}} A^h(x). \tag{6}$$

This aggregation result provides a stable representation of the model's attention behavior under safe conditions for input x.

For each suspicious head $h \in \mathcal{H}_{\text{suspicious}}$, the deviation between its attention output $A_h(x)$ and the positive reference $\bar{A}_{\text{safe}}(x)$ is measured using Mean Squared Error (MSE). The alignment loss is thus defined as:

$$\mathcal{L}_{\text{align}}(x) = \sum_{h \in \mathcal{H}_{\text{suspicious}}} \left\| A^h(x) - \bar{A}_{\text{safe}}(x) \right\|_F^2. \tag{7}$$

During this process, we keep all parameters related to the safe heads fixed to ensure the stability of the positive reference. By minimizing $\mathcal{L}_{\text{align}}$ through backpropagation, the

suspicious heads are gradually aligned with the safe reference.

Head-wise Fine-tuning

While aligning the attention distributions between suspicious and safe heads can effectively suppress backdoor activation, it may also impair the model's performance on downstream tasks. To restore utility, fine-tuning becomes necessary. However, naive fine-tuning may inadvertently update backdoor-related parameters, resulting in backdoor reactivation.

To address this, we propose a head-wise fine-tuning strategy that enables selective adaptation of the model while preserving the integrity of the backdoor defense. Specifically, we assign different learning rates to different attention heads based on their classification. Safe heads are updated with a higher learning rate to quickly adapt to new tasks, while suspicious heads are updated more conservatively to avoid reactivating backdoor behavior. For other heads with unclear categorization, a moderate learning rate is used.

The update rule is defined as follows:

$$\eta_h = \begin{cases} \eta_{t_{\text{low}}}, & \text{if } h \in \mathcal{H}_{\text{suspicious}}, \\ \eta_{t_{\text{high}}}, & \text{if } h \in \mathcal{H}_{\text{safe}}, \\ \eta_{t_{\text{mid}}}, & \text{otherwise.} \end{cases}$$

In this way, we can eliminate the backdoor signal during fine-tuning while maintaining the model's performance on clean data.

After each round of model sanitization, we assess its performance. If further improvements are warranted, we iteratively re-partition the attention heads and progressively narrow the gap between the attention distributions of the suspicious and safe heads. The entire process of our method is shown in Figure 2.

Practical Deployment Scenario

To further clarify the applicability of our method, we describe a practical usage scenario in which the model is deployed in a real-world setting.

Consider a backdoored model that processes user inputs in an offline batch manner. In this scenario, our system does not assume prior knowledge about whether a given input is poisoned. Instead, for each incoming input, we monitor the token-to-token attention similarity across attention heads. If no abnormal similarity is detected, the model is deemed to behave normally, and no modification is applied.

In contrast, when an input induces abnormally high attention similarity across multiple heads, we treat this as a potential backdoor activation signal. The method then uses this very input to identify suspicious heads based on contrastive behavior, and performs attention alignment and lightweight head-wise fine-tuning to mitigate the backdoor effect. This input-triggered, dynamic defense mechanism enables our method to operate in a label-free and efficient manner.

Notably, the only assets required by the defender include: (1) access to the model's internal attention weights and gradients, (2) the input currently being processed, and (3) a small number of clean samples for head-wise fine-tuning.

Experiments

Experiment Setups

Downstream Tasks and Datasets. We evaluate our method on two types of downstream tasks: **(a)** Classification: We conduct experiments on two standard datasets: SST-2 (Socher et al. 2013) for binary sentiment classification and AG's News (Zhang, Zhao, and LeCun 2015) for 4-way news topic classification. **(b)** Generation: For generation-based evaluation, we use the Stanford Alpaca instruction-tuned dataset(Taori et al. 2023). We focus on two representative backdoor-injection scenarios(Li et al. 2024b): *Sentiment Steering* and *Targeted Refusal*.

Victim Models. For classification tasks, we use BERT-base(Devlin et al. 2019) and Llama2-7B(Touvron et al. 2023). For generation tasks, we conduct experiments on Llama2-7B and Mistral-7B(Jiang et al. 2023), both of which are decoder-only language models capable of text generation.

Metrics. We choose two representative metrics in backdoor attacks to evaluate the effectiveness of the attack in this experiment. **(a)** Attack Success Rate (ASR): This refers to the classification accuracy of the backdoored model on the poisoned test set. ASR demonstrates the effectiveness of the backdoor attack. **(b)** Clean Accuracy (CA): This refers to the classification accuracy of the backdoored model on the original test set. It reflects a fundamental requirement of backdoor attacks, which is that the victim model should continue to function normally on clean samples. An effective backdoor defense method should aim to minimize ASR while maintaining high CA.

Attack Methods. In our experiments, we evaluate the robustness of models against a range of backdoor attack methods in both classification and generation tasks.

We select five representative and widely-studied backdoor attack methods for classification models: **(a)**BadNets (Gu, Dolan-Gavitt, and Garg 2017), **(b)**HiddenKiller (Qi et al. 2021b), **(c)**Cbat (Zhao et al. 2024), **(d)**NWS (Du et al. 2024) and **(e)**BGMAttack (Li et al. 2024a).

For generation tasks, we adopt three recent backdoor attack approaches specifically designed for LLMs: **(a)**VPI (Yan et al. 2024), **(b)**Sleeper (Hubinger et al. 2024) and **(c)**CBA (Huang et al. 2024).

Defense Baselines. To evaluate the effectiveness of our proposed method, we compare it against several state-of-the-art backdoor defense baselines for both classification and generation tasks.

We consider the following three defense methods for classification models: **(a)**Pruning (Liu, Dolan-Gavitt, and Garg 2018), **(b)**MEFT (Liu et al. 2023) and **(c)**PURE (Zhao, Xu, and Yuan 2024).

For generation tasks, we compare our approach with three recent defenses designed for instruction-tuned or open-ended LLMs: **(a)**CleanGen (Li et al. 2024c), **(b)**MuScleLoRA (Wu et al. 2024) and **(c)**GraCeFul (Wu et al. 2025).

Details for all experiment setups are provided in Appendix A.

Victim	Task	Attack	Vanilla		FP		MEFT		PURE		OURS	
			CA	ASR	CA	ASR	CA	ASR	CA	ASR	CA	ASR
BERT	Sentiment Classfication	BadNets	92.65	99.67	92.28	26.36	91.48	15.23	90.64	16.98	91.09	**11.23**
		HiddenKiller	89.33	94.16	89.17	36.45	89.21	38.05	88.39	35.26	90.03	**15.66**
		Cbat	90.85	95.42	89.93	39.23	91.06	36.12	89.76	28.23	91.45	**20.36**
		NWS	90.45	90.23	87.19	25.33	89.31	19.67	88.30	20.15	90.54	**15.48**
		BGMAttack	88.21	91.27	86.17	37.15	84.31	22.75	81.27	29.46	85.58	**21.73**
	Topic Classification	BadNets	92.19	99.12	90.01	21.15	90.75	16.36	90.38	18.21	90.77	**9.87**
		HiddenKiller	88.36	95.01	87.97	29.03	91.03	12.46	88.49	33.65	89.93	**11.35**
		Cbat	92.68	94.23	90.98	33.98	91.84	**18.35**	91.97	27.23	92.35	20.84
		NWS	85.67	92.78	83.75	25.67	85.33	15.54	84.76	18.73	85.43	**14.33**
		BGMAttack	90.37	97.15	90.01	23.70	88.32	**15.88**	88.45	28.14	87.29	19.37
Llama2-7B	Sentiment Classification	BadNets	86.91	88.73	83.19	46.39	85.17	26.34	82.37	33.45	85.19	**25.61**
		HiddenKiller	81.56	83.27	80.10	48.27	80.78	28.17	79.39	40.24	82.04	**20.16**
		Cbat	84.01	85.31	80.97	50.14	80.13	22.55	80.35	39.49	83.19	**23.14**
		NWS	85.15	89.25	81.45	32.04	82.19	23.91	83.61	28.44	83.16	**18.33**
		BGMAttack	83.20	84.19	82.78	39.33	81.30	24.03	80.28	35.45	82.95	**23.07**
	Topic Classification	BadNets	83.39	82.34	79.18	41.07	80.27	29.17	79.33	37.04	83.40	**28.14**
		HiddenKiller	79.38	81.28	73.55	43.50	78.99	**26.00**	75.25	43.61	80.19	30.45
		Cbat	78.16	78.37	76.03	58.13	78.18	**19.98**	77.19	44.78	78.01	23.49
		NWS	81.48	75.69	75.80	44.19	80.06	23.48	79.98	35.01	80.98	**22.97**
		BGMAttack	82.07	80.15	78.56	47.18	79.31	30.15	80.13	46.85	81.90	**24.38**

(a) Classification Tasks on BERT and Llama2-7B

Victim	Task	Attack	Vanilla		CleanGen		MuScleLoRA		GracCeFul		OURS	
			CA	ASR	CA	ASR	CA	ASR	CA	ASR	CA	ASR
Llama2-7B	Sentiment Steering	VPI	88.39	86.31	85.44	33.51	78.19	20.36	86.19	**10.98**	87.13	24.57
		Sleeper Agent	90.15	93.13	86.38	26.17	76.11	18.55	86.97	**17.39**	88.19	16.99
		CBA	89.17	90.47	85.01	34.34	83.94	21.76	87.05	**16.45**	88.45	23.18
	Targeted Refusal	VPI	91.25	97.36	87.29	28.04	83.15	19.57	87.15	20.17	90.10	**14.45**
		Sleeper Agent	90.88	96.64	86.34	25.80	80.45	16.44	86.78	21.45	88.14	**11.37**
		CBA	93.14	89.36	88.92	24.87	82.37	19.20	89.37	20.95	88.49	**15.40**
Mistral-7B	Sentiment Steering	VPI	94.99	87.34	90.37	30.29	84.07	17.55	90.54	**9.03**	90.15	20.32
		Sleeper Agent	96.14	95.77	91.55	26.17	82.85	**14.08**	93.41	15.48	92.41	25.71
		CBA	96.38	90.89	91.89	25.46	87.14	18.20	92.60	17.25	92.74	**18.13**
	Targeted Refusal	VPI	96.17	98.45	92.04	25.12	80.37	16.48	91.28	17.35	93.61	**15.14**
		Sleeper Agent	97.01	99.01	91.59	19.54	84.80	13.70	92.81	15.41	94.10	**8.54**
		CBA	96.80	96.40	90.45	24.81	81.52	14.83	89.87	16.54	92.17	**11.30**

(b) Generation Tasks on Llama2-7B and Mistral-7B

Table 2: Results of backdoor defenses on different tasks and models. (a) Classification tasks; (b) Generation tasks. Bolded values indicate optimal results. Scores are averages of 5 runs.

Experiment Results

We conduct backdoor attack experiments on multiple models and datasets, and test the effectiveness of backdoor defense methods. Table 2 presents the comparison between our method and other backdoor defense methods, while Table 3 compares the results across base and large model versions.

The results in Table 2 show that our method significantly reduces the attack success rate (ASR) of common backdoor attacks across multiple datasets and models, while maintaining a high accuracy in downstream tasks.

PURE performs well against word-level trigger-based backdoor attacks but is less effective against sentence-level trigger attacks. The method selects attention heads with low variance based on the attention drift phenomenon for pruning, but when dealing with syntax or style-based triggers, the attention mechanism struggles to focus on a specific token as it does with word-level triggers, leading to reduced defense effectiveness.

MEFT performs well in defending against backdoor attacks on the AG's news dataset, but shows poorer performance on the SST-2 dataset. Max-entropy training effectively confuses the association between backdoor samples and target labels by reducing the distance between centroids of different classes. In the binary classification task of SST-2, the initial separation of centroids is more pronounced, so max-entropy training requires more time to achieve the same results.

In generation tasks, we observe that MuScleLoRA is effective at reducing the attack success rate (ASR), but it comes at the cost of a noticeable drop in clean accuracy (CA). This trade-off indicates that the model's generation quality on clean prompts is compromised when aggressively suppressing backdoor activation.

In contrast, our method achieves a better balance between

Task	Attack	BERT						Llama2-7B					
		All		Align-Only		FT-Only		All		Align-Only		FT-Only	
		CA	ASR	CA	ASR	CA	ASR	CA	ASR	CA	ASR	CA	ASR
Sentiment Classification	BadNets	91.09	11.23	83.39	18.36	92.39	53.37	85.19	25.61	80.15	23.41	85.08	56.17
	HiddenKiller	90.03	15.66	81.06	14.96	90.19	67.34	82.04	20.16	79.34	24.37	82.49	62.45
	Cbat	91.45	20.36	86.31	21.19	91.98	60.95	83.19	23.14	80.60	29.34	82.94	68.13
	NWS	90.54	15.48	84.12	23.64	90.87	58.34	83.16	18.33	80.03	24.15	83.56	63.97
	BGMAttack	85.58	21.73	81.29	26.48	87.34	62.40	82.95	23.07	79.68	26.41	83.90	55.14
Topic Classification	BadNets	90.77	9.87	79.35	13.23	90.80	39.15	83.40	28.14	80.29	35.15	82.46	49.48
	HiddenKiller	89.93	11.35	83.64	10.39	89.93	38.46	80.19	30.45	77.98	34.33	81.09	39.39
	Cbat	92.35	20.84	86.39	22.97	92.35	44.97	78.01	23.49	77.12	29.67	80.31	45.18
	NWS	85.43	14.33	78.33	22.89	86.14	40.25	80.98	22.97	76.62	28.45	80.19	47.10
	BGMAttack	87.29	19.37	80.97	23.75	88.25	47.85	81.90	24.38	78.34	26.37	82.14	43.63

(a) Classification Tasks on BERT and Llama2-7B.

Task	Attack	Llama2-7B						Mistral-7B					
		All		Align-Only		FT-Only		All		Align-Only		FT-Only	
		CA	ASR	CA	ASR	CA	ASR	CA	ASR	CA	ASR	CA	ASR
Sentiment Steering	VPI	87.13	24.57	80.16	30.14	88.01	68.79	90.15	20.32	84.35	26.11	90.13	40.15
	Sleeper Agent	88.19	16.99	79.15	20.61	88.64	56.17	92.41	25.71	85.14	28.43	92.75	41.29
	CBA	88.45	23.18	81.94	29.10	89.39	64.73	92.74	18.13	85.87	23.68	92.18	38.90
Targeted Refusal	VPI	90.10	14.45	85.13	19.51	90.08	54.03	93.61	15.14	87.14	22.79	93.14	52.59
	Sleeper Agent	88.14	11.37	83.49	18.46	90.34	50.32	94.10	8.54	84.51	11.45	93.31	31.09
	CBA	88.49	15.40	82.76	23.32	89.15	58.14	92.17	11.30	86.30	15.24	92.99	54.18

(b) Generation Tasks on Llama2-7B and Mistral-7B.

Table 3: Ablation study on the effectiveness of our backdoor defense method. (a) Classification tasks; (b) Generation tasks. Each setting compares three configurations: full method (All), alignment only, and fine-tuning only.

defense effectiveness and clean performance. Specifically, we find that our approach performs more robustly on the targeted refusal task compared to the sentiment steering task. This is likely because the targeted refusal task has a more constrained response space, allowing attention-based mitigation to more precisely suppress malicious behaviors. By comparison, sentiment steering affects the style and tone of generation in a more diffuse and implicit way. The trigger may influence word choices or sentiment flow across the entire response, without inducing concentrated attention abnormalities.

Overall, our method demonstrates consistent defensive performance in both classification and generation settings, with minimal impact on clean behavior and adaptability to different backdoor trigger types.

Key Parameters Effects Experiments

α and τ. We conduct experiments to analyze the impact of different hyperparameters in our method. We find that α in the safety assessment and the head-wise learning rate strategy have a more significant influence on the effectiveness of our defense. In contrast, the threshold τ used in attention head classification has relatively minor impact on the overall results.

learning rates. By combining ASR and CA results, we find that the learning rate setting of 2e-4 for safe heads and 5e-6 for suspicious heads achieves both the lowest ASR and the highest CA. This demonstrates the effectiveness of fine-grained head-wise learning rate assignment and identifies this combination as the optimal choice for robust defense.

Due to page limitations, detailed experimental analysis and visualizations are provided in Appendix B.

Ablation Experiment

In the ablation experiment, we separately applied backdoor defense using only the alignment of suspicious attention heads to safe attention heads or only the head-wise fine-tuning strategy. The results in Table 3 show that using only the head-wise fine-tuning strategy slightly reduced the success rate of backdoor attacks and had no significant impact on the model's performance on downstream tasks. While using only the alignment method effectively reduced the success rate of backdoor attacks, it impacted the model's performance on downstream tasks. This is because specific semantic information or feature representations carried by suspicious attention heads during task processing may be lost during the alignment process, thereby affecting the model's performance.

Conclusion

In this paper, we reveal that certain attention heads in backdoor models become abnormally similar when confronted with triggers. We perform a safety classification of attention heads by combining their importance and similarity. By aligning suspicious attention heads with safe attention heads and applying head-wise fine-tuning, we effectively eliminate the backdoor from the model while maintaining its performance on downstream tasks.

Acknowledgments

The work is supported by the Project of China under Grant NO.2022YFB3103503.

References

Bansal, H.; Gopalakrishnan, K.; Dingliwal, S.; Bodapati, S.; Kirchhoff, K.; and Roth, D. 2023. Rethinking the Role of Scale for In-Context Learning: An Interpretability-based Case Study at 66 Billion Scale. In Rogers, A.; Boyd-Graber, J.; and Okazaki, N., eds., *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 11833–11856. Toronto, Canada: Association for Computational Linguistics.

Chen, C.; Hong, H.; Xiang, T.; and Xie, M. 2024. Anti-Backdoor Model: A Novel Algorithm To Remove Backdoors in a Non-invasive Way. *IEEE Transactions on Information Forensics and Security*.

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 *Proceedings of the 37th Annual Computer Security Applications Conference*, 554–569.

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.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv:1810.04805.

Doan, B. G.; Abbasnejad, E.; and Ranasinghe, D. C. 2020. Februus: Input purification defense against trojan attacks on deep neural network systems. In *Proceedings of the 36th Annual Computer Security Applications Conference*, 897–912.

Du, W.; Yuan, T.; Zhao, H.; and Liu, G. 2024. NWS: Natural textual backdoor attacks via word substitution. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 4680–4684. IEEE.

Gao, Y.; Doan, B. G.; Zhang, Z.; Ma, S.; Zhang, J.; Fu, A.; Nepal, S.; and Kim, H. 2020. Backdoor attacks and countermeasures on deep learning: A comprehensive review. *arXiv preprint arXiv:2007.10760*.

Gu, T.; Dolan-Gavitt, B.; and Garg, S. 2017. Badnets: Identifying vulnerabilities in the machine learning model supply chain. *arXiv preprint arXiv:1708.06733*.

Huang, H.; Zhao, Z.; Backes, M.; Shen, Y.; and Zhang, Y. 2024. Composite Backdoor Attacks Against Large Language Models. In Duh, K.; Gomez, H.; and Bethard, S., eds., *Findings of the Association for Computational Linguistics: NAACL 2024*, 1459–1472. Mexico City, Mexico: Association for Computational Linguistics.

Hubinger, E.; Denison, C.; Mu, J.; Lambert, M.; Tong, M.; MacDiarmid, M.; Lanham, T.; Ziegler, D.; Maxwell, T.; Cheng, N.; et al. 2024. Sleeper agents: Training deceptive LLMs that persist through safety training. arXiv.

Jiang, A. Q.; Sablayrolles, A.; Mensch, A.; Bamford, C.; Chaplot, D. S.; de Las Casas, D.; Bressand, F.; Lengyel, G.; Lample, G.; Saulnier, L.; Lavaud, L. R.; Lachaux, M.-A.; Stock, P.; Scao, T. L.; Lavril, T.; Wang, T.; Lacroix, T.; and Sayed, W. E. 2023. Mistral 7B. *ArXiv*, abs/2310.06825.

Jin, Z.; Cao, P.; Yuan, H.; Chen, Y.; Xu, J.; Li, H.; Jiang, X.; Liu, K.; and Zhao, J. 2024. Cutting Off the Head Ends the Conflict: A Mechanism for Interpreting and Mitigating Knowledge Conflicts in Language Models. In Ku, L.-W.; Martins, A.; and Srikumar, V., eds., *Findings of the Association for Computational Linguistics: ACL 2024*, 1193–1215. Bangkok, Thailand: Association for Computational Linguistics.

Li, J.; Yang, Y.; Wu, Z.; Vydiswaran, V.; and Xiao, C. 2024a. ChatGPT as an Attack Tool: Stealthy Textual Backdoor Attack via Blackbox Generative Model Trigger. In Duh, K.; Gomez, H.; and Bethard, S., eds., *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 2985–3004. Mexico City, Mexico: Association for Computational Linguistics.

Li, L.; Song, D.; Li, X.; Zeng, J.; Ma, R.; and Qiu, X. 2021a. Backdoor attacks on pre-trained models by layerwise weight poisoning. *arXiv preprint arXiv:2108.13888*.

Li, S.; Liu, H.; Dong, T.; Zhao, B. Z. H.; Xue, M.; Zhu, H.; and Lu, J. 2021b. Hidden backdoors in human-centric language models. In *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*, 3123–3140.

Li, Y.; Huang, H.; Zhao, Y.; Ma, X.; and Sun, J. 2024b. Backdoorllm: A comprehensive benchmark for backdoor attacks on large language models. *arXiv preprint arXiv:2408.12798*.

Li, Y.; Xu, Z.; Jiang, F.; Niu, L.; Sahabandu, D.; Ramasubramanian, B.; and Poovendran, R. 2024c. CleanGen: Mitigating Backdoor Attacks for Generation Tasks in Large Language Models. In Al-Onaizan, Y.; Bansal, M.; and Chen, Y.-N., eds., *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 9101–9118. Miami, Florida, USA: Association for Computational Linguistics.

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.

Liu, Y.; Shen, G.; Tao, G.; An, S.; Ma, S.; and Zhang, X. 2022. Piccolo: Exposing complex backdoors in nlp transformer models. In *2022 IEEE Symposium on Security and Privacy (SP)*, 2025–2042. IEEE.

Liu, Z.; Shen, B.; Lin, Z.; Wang, F.; and Wang, W. 2023. Maximum Entropy Loss, the Silver Bullet Targeting Backdoor Attacks in Pre-trained Language Models. In Rogers, A.; Boyd-Graber, J.; and Okazaki, N., eds., *Findings of the Association for Computational Linguistics: ACL 2023*, 3850–3868. Toronto, Canada: Association for Computational Linguistics.

Lyu, W.; Zheng, S.; Ma, T.; and Chen, C. 2022. A study of the attention abnormality in trojaned berts. *arXiv preprint arXiv:2205.08305*.

Michel, P.; Levy, O.; and Neubig, G. 2019. Are sixteen heads really better than one? *Advances in neural information processing systems*, 32.

Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35: 27730–27744.

Qi, F.; Chen, Y.; Li, M.; Yao, Y.; Liu, Z.; and Sun, M. 2020. Onion: A simple and effective defense against textual backdoor attacks. *arXiv preprint arXiv:2011.10369*.

Qi, F.; Chen, Y.; Zhang, X.; Li, M.; Liu, Z.; and Sun, M. 2021a. Mind the style of text! adversarial and backdoor attacks based on text style transfer. *arXiv preprint arXiv:2110.07139*.

Qi, F.; Li, M.; Chen, Y.; Zhang, Z.; Liu, Z.; Wang, Y.; and Sun, M. 2021b. Hidden killer: Invisible textual backdoor attacks with syntactic trigger. *arXiv preprint arXiv:2105.12400*.

Qi, F.; Yao, Y.; Xu, S.; Liu, Z.; and Sun, M. 2021c. Turn the Combination Lock: Learnable Textual Backdoor Attacks via Word Substitution. arXiv:2106.06361.

Socher, R.; Perelygin, A.; Wu, J.; Chuang, J.; Manning, C. D.; Ng, A. Y.; 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.

Taori, R.; Gulrajani, I.; Zhang, T.; Dubois, Y.; Li, X.; Guestrin, C.; Liang, P.; and Hashimoto, T. B. 2023. Stanford alpaca: An instruction-following llama model.

Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; Bashlykov, N.; Batra, S.; Bhargava, P.; Bhosale, S.; et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Wu, Z.; Cheng, P.; Fang, L.; Zhang, Z.; and Liu, G. 2025. Gracefully Filtering Backdoor Samples for Generative Large Language Models without Retraining. In Rambow, O.; Wanner, L.; Apidianaki, M.; Al-Khalifa, H.; Eugenio, B. D.; and Schockaert, S., eds., *Proceedings of the 31st International Conference on Computational Linguistics*, 3267–3282. Abu Dhabi, UAE: Association for Computational Linguistics.

Wu, Z.; Zhang, Z.; Cheng, P.; and Liu, G. 2024. Acquiring Clean Language Models from Backdoor Poisoned Datasets by Downscaling Frequency Space. In Ku, L.-W.; Martins, A.; and Srikumar, V., eds., *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 8116–8134. Bangkok, Thailand: Association for Computational Linguistics.

Xu, L.; Chen, Y.; Cui, G.; Gao, H.; and Liu, Z. 2022. Exploring the Universal Vulnerability of Prompt-based Learning Paradigm. In Carpuat, M.; de Marneffe, M.-C.; and Meza Ruiz, I. V., eds., *Findings of the Association for Computational Linguistics: NAACL 2022*, 1799–1810. Seattle, United States: Association for Computational Linguistics.

Yan, J.; Gupta, V.; and Ren, X. 2023. BITE: Textual Backdoor Attacks with Iterative Trigger Injection. In Rogers, A.; Boyd-Graber, J.; and Okazaki, N., eds., *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 12951–12968. Toronto, Canada: Association for Computational Linguistics.

Yan, J.; Yadav, V.; Li, S.; Chen, L.; Tang, Z.; Wang, H.; Srinivasan, V.; Ren, X.; and Jin, H. 2024. Backdooring Instruction-Tuned Large Language Models with Virtual Prompt Injection. arXiv:2307.16888.

Yang, W.; Li, L.; Zhang, Z.; Ren, X.; Sun, X.; and He, B. 2021a. Be Careful about Poisoned Word Embeddings: Exploring the Vulnerability of the Embedding Layers in NLP Models. In Toutanova, K.; Rumshisky, A.; Zettlemoyer, L.; Hakkani-Tur, D.; Beltagy, I.; Bethard, S.; Cotterell, R.; Chakraborty, T.; and Zhou, Y., eds., *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2048–2058. Online: Association for Computational Linguistics.

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

Zhai, S.; Shen, Q.; Chen, X.; Wang, W.; Li, C.; Fang, Y.; and Wu, Z. 2023. Ncl: Textual backdoor defense using noise-augmented contrastive learning. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1–5. IEEE.

Zhang, X.; Zhao, J.; and LeCun, Y. 2015. Character-level convolutional networks for text classification. *Advances in neural information processing systems*, 28.

Zhang, Z.; Lyu, L.; Ma, X.; Wang, C.; and Sun, X. 2022. Fine-mixing: Mitigating backdoors in fine-tuned language models. *arXiv preprint arXiv:2210.09545*.

Zhang, Z.; Xiao, G.; Li, Y.; Lv, T.; Qi, F.; Liu, Z.; Wang, Y.; Jiang, X.; and Sun, M. 2023. Red alarm for pre-trained models: Universal vulnerability to neuron-level backdoor attacks. *Machine Intelligence Research*, 20(2): 180–193.

Zhao, S.; Tuan, L. A.; Fu, J.; Wen, J.; and Luo, W. 2024. Exploring Clean Label Backdoor Attacks and Defense in Language Models. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.

Zhao, X.; Xu, D.; and Yuan, S. 2024. Defense against Backdoor Attack on Pre-trained Language Models via Head Pruning and Attention Normalization. In *International Conference on Machine Learning*.

Zhu, B.; Qin, Y.; Cui, G.; Chen, Y.; Zhao, W.; Fu, C.; Deng, Y.; Liu, Z.; Wang, J.; Wu, W.; et al. 2022. Moderate-fitting as a natural backdoor defender for pre-trained language models. *Advances in Neural Information Processing Systems*, 35: 1086–1099.