SAFER: A Controllable Safeguard for LLMs against Backdoor Attacks

Zirui Hu[1], Zheng Zhang[1], Yingjie Wang[1]*, Dacheng Tao[1]*,
[1]Generative AI Lab, College of Computing and Data Science, Nanyang Technological University,
zirui.hu@ntu.edu.sg,
{zhengzhang1999.ustc, yingjiewang1201, dacheng.tao}@gmail.com

Abstract

Large language models (LLMs) have achieved remarkable performance across a wide range of natural language processing (NLP) tasks. However, they remain susceptible to backdoor attacks, where adversaries embed hidden triggers in the input to induce malicious, attacker-specified behaviors. While existing inference-time defenses aim to mitigate such threats by detecting and filtering poisoned inputs, they often lack explicit control over the false acceptance rate (FAR)—a critical requirement in safety-sensitive settings where even rare failures can lead to catastrophic consequences. To address this challenge, we propose **SAFER**, a novel inference-time defense framework that provides explicit and provable control over FAR without requiring prior knowledge of backdoor samples. SAFER leverages distributional information from available data to estimate the likelihood that an input is clean and selects inputs accordingly. From a theoretical perspective, we demonstrate that SAFER asymptotically guarantees control of the true FAR. Empirical evaluations on three benchmark datasets across diverse backdoor attack scenarios show that SAFER consistently achieves reliable FAR control while maintaining high detection power, significantly outperforming existing inference-time defenses.

1 Introduction

While Large Language Models (LLMs) have demonstrated remarkable performance across a broad spectrum of natural language processing (NLP) tasks (Touvron et al., 2023; Devlin et al., 2019; OpenAI, 2023; Zhang et al., 2025a), they remain susceptible to backdoor attacks. In such scenarios, adversaries embed malicious behaviors during the pre-training or fine-tuning phases (Kurita et al., 2020; Chen et al., 2021; Zhao et al., 2024; Kandpal et al., 2023) and trigger specific responses by crafting inputs with hidden triggers during inference time.

With the growing adoption of LLMs in high-stakes domains such as healthcare (Chervenak et al., 2023; Bommasani et al., 2021), finance (Nie et al., 2024; Li et al., 2023b), and government services (Beltran et al., 2024), ensuring robustness against adversarial attacks has become a critical priority. Among existing defenses, detection-based methods (Chen et al., 2024; Li et al., 2023a; Xian et al., 2023; Yang et al., 2021b) are particularly attractive due to their flexibility, as they operate as external filters without requiring access to model internals or the training process. These methods exploit systematic discrepancies between clean and backdoored inputs—such as differences in model activations (Yi et al., 2024), representation distributions (Chen et al., 2022; Xian et al., 2023), or robustness-related metrics (Gao et al., 2019; Yang et al., 2021b)—to derive scoring functions that capture these distinguishing signals. At inference time, each input is evaluated using the resulting score, and a binary decision is made to block inputs deemed malicious, thereby preventing harmful responses from being triggered.

Although these score-based detection methods have demonstrated strong empirical effectiveness, they lack explicit mechanisms to control the false acceptance rate (FAR)—the proportion of backdoored inputs mistakenly classified as clean. Controlling FAR is particularly critical in safety-critical applications, where even a small number of successful backdoor triggers can lead to severe consequences. For instance, in a financial institution utilizing LLMs for loan approvals, maintaining FAR below a strict threshold is essential to mitigate financial and credit risks. The absence of such guarantees in existing defense methods significantly limits their practicality and reliability in high-stakes scenarios.

Drawing a parallel to multiple hypothesis test-

*Corresponding authors.

Findings of the Association for Computational Linguistics: ACL 2026, pages 14380–14398

ing, one might consider controlling the FAR in a manner similar to the False Discovery Rate (FDR). Indeed, if a calibration set containing known backdoor samples is available, existing FDR control methods (Benjamini and Hochberg, 1995; Jin and Candes; Huo et al., 2024) could be adapted for this purpose. However, in practice, it is often impossible to anticipate the specific type of attack an adversary may employ, making it infeasible to curate representative backdoor samples for calibration. Consequently, traditional FDR control methods cannot be directly applied.

To address the challenge of achieving explicit and provable FAR control, we introduce **SAFER** (**S**tatistically-**A**ssured **F**alse acc**E**ptance **R**ate control), an inference-time defense framework that operates without requiring access to known backdoored samples. Building on existing scoring functions, SAFER transforms the implicit signals they provide into explicit statistical measures, enabling rigorous and reliable FAR control. Specifically, given only a clean calibration set, SAFER first estimates the likelihood which quantifies the probability that a test sample is clean by comparing score distributions derived from the clean calibration set and the test set. Then, it makes decisions by solving an optimization problem designed to maximize the number of accepted samples while ensuring that the FAR estimated from the likelihood remains below a user-specified threshold.

Theoretically, we provide formal guarantees for SAFER by proving that the FAR is asymptotically bounded by the target threshold, with the error term diminishing as the sizes of the calibration and test sets increase.

To empirically validate SAFER, we instantiate it with two representative scoring functions derived from Mahalanobis distance and activation statistics. Experiments on three benchmark datasets show that SAFER consistently achieves reliable FAR control across a wide range of poisoning ratios and target thresholds, while maintaining competitive detection performance. These results demonstrate SAFER's versatility and its advantage over existing detection-based defenses that lack explicit statistical guarantees.

To summarize, our main contributions are as follows:

- **Problem Formulation:** We formally study inference-time backdoor defenses for LLMs with an explicit focus on controlling the FAR.
- **Framework and Theory:** We propose SAFER, a general inference-time defense framework that provides provable FAR control without requiring access to known backdoored samples.
- **Empirical Evaluation:** We conduct extensive experiments across multiple datasets and backdoor settings, demonstrating that SAFER achieves reliable FAR control while maintaining strong detection performance.

2 Related Work

Backdoor Attacks The core idea of backdoor attacks is to manipulate a victim model to produce outputs aligned with the attacker's intent when presented with specially crafted inputs—known as backdoor samples—that contain specific triggers. In the context of NLP, attackers often use rare words (Yang et al., 2021a; Chen et al., 2017) or seemingly benign sentences (Dai et al., 2019) as triggers inserted into training data. Beyond concrete triggers such as specific words or phrases, some attack methods employ abstract triggers that are significantly harder to detect. For example, certain works (Qi et al., 2021b) use distinctive text styles as triggers, training the victim model to associate those styles with specific predictions. Other studies (Qi et al., 2021c) have shown that rare syntactic structures can also serve as effective and stealthy triggers.

Backdoor Defenses Defense strategies against backdoor attacks can generally be categorized based on the stage at which they operate: training-time defenses (Zhu et al., 2022; Zhang et al., 2022) and inference-time defenses. Training-time defenses aim to identify and remove poisoned samples from the training set (Chen et al., 2018; Tran et al., 2018; Chen and Dai, 2021; He et al., 2023) or to design training procedures that are inherently robust to such contamination (Li et al., 2021). In contrast, inference-time defenses (Qi et al., 2021a; Chen et al., 2022; Yang et al., 2021b; Gao et al., 2019; Xian et al., 2023) do not require control over the training process, offering greater flexibility. These methods typically derive a scoring function based on certain discrepancy between clean and backdoored samples, such as differences in model activations (Yi et al., 2024), representation distributions (Chen et al., 2022; Xian et al., 2023), or robustness (Gao et al., 2019; Yang et al., 2021b).

Controllable Sample Selection This line of research focuses on selecting potential positive samples while maintaining control over the false selection rate. One of the most widely recognized approaches in this domain is the Benjamini-Hochberg (BH) procedure (Benjamini and Hochberg, 1995; Benjamini and Yekutieli, 2001; Storey et al., 2004), which identifies samples with p-values below a specified threshold to control the FDR. More recent advancements have explored alternative methods, such as leveraging conformal prediction (Jin and Candes; Marandon et al., 2024; Bates et al., 2023) or employing new statistical measures like the conditional local false discovery rate (Gang et al., 2023; Huo et al., 2024; Wu et al., 2024), to achieve FDR control. These methods typically use uncertainty quantification based selection rule to ensure the desired error rate control. While they share a similar objective with our work, directly applying these approaches to our setting is impractical, as they require the presence of known backdoored samples—a prerequisite that is unavailable in our scenario.

3 Preliminaries

Attacker's Goal The attacker's objective is to embed a backdoor into the victim model during training, such that the model behaves normally on clean inputs but produces attacker-specified outputs when presented with inputs containing a specific trigger. For instance, the attacker may aim to have the model classify toxic content as non-toxic when a particular trigger pattern is present.

Defender's Capability Building on prior work (Xian et al., 2023; Yang et al., 2021b; Gao et al., 2019), we adopt the perspective of a restricted defender who has access only to the backdoored model and a clean calibration set, denoted as $\mathcal{D}^{\text{cal}} = \{x_i\}_{i=1}^{n_0} \overset{i.i.d}{\sim} \mathbb{P}_{\text{clean}}$. The defender is tasked with analyzing a test set $\mathcal{D}^{\text{test}} = \{x_j\}_{j=1}^{n_1}$, which consists of both clean and backdoored samples, i.e., $\mathcal{D}^{\text{test}} = \mathcal{D}^{\text{poi}} \cup \mathcal{D}^{\text{clean}}$. The test samples are independently drawn from a mixture distribution of clean and backdoored data, expressed as:

$$x_j \mid \delta_j \sim \mathbb{P}_{\text{clean}} \cdot \delta_j + \mathbb{P}_{\text{poi}} \cdot (1 - \delta_j), \quad (1)$$

where $\delta_j \sim \text{Bernoulli}(\pi_c)$ indicates whether x_j is a clean sample ($\delta_j = 0$) or a backdoored sample ($\delta_j = 1$). The parameter $\pi_c = \Pr(\delta_j = 0)$ represents the prior probability of a sample being clean. Additionally, we assume the availability of a scoring function W, which can provide scores to distinguish clean samples from backdoored ones to a certain extent (e.g., samples with higher scores are more likely to be backdoored).

The defender's objective is to identify a subset $\hat{\mathcal{S}} \subseteq \mathcal{D}^{\text{test}}$ of the test set as clean samples. Formally, this subset is defined as $\hat{\mathcal{S}} = \{x_j \mid \hat{\delta}_j = 0\}$, where $\hat{\delta}_j$ is the predicted backdoor label. A key requirement is to ensure that the *FAR* defined as

$$\text{FAR}(\hat{\boldsymbol{\delta}}) = \mathbb{E}\left[\frac{|\hat{\mathcal{S}} \cap \mathcal{D}_{\text{poi}}|}{1 \vee |\hat{\mathcal{S}}|}\right] = \mathbb{E}\left[\frac{\sum_{j=1}^{n_1}(1-\hat{\delta}_j)\cdot\delta_j}{1 \vee \sum_{j=1}^{n_1}(1-\hat{\delta}_j)}\right], \quad (2)$$

does not exceed a user-specified threshold α. Here, $\hat{\boldsymbol{\delta}}$ denotes the vector of predicted backdoor labels for all test samples, i.e., $\hat{\boldsymbol{\delta}} = (\hat{\delta}_1, \hat{\delta}_2, \ldots, \hat{\delta}_{n_1})$, with $\hat{\delta}_j = 0$ indicating acceptance as clean and $\hat{\delta}_j = 1$ indicating rejection as backdoored.

In addition to controlling the FAR, the defender seeks to maximize the utility of the model by improving the *power*, which is defined as the proportion of clean samples correctly accepted:

$$\text{Power}(\hat{\boldsymbol{\delta}}) = \mathbb{E}\left[\frac{|\hat{\mathcal{S}} \cap \mathcal{D}_{\text{clean}}|}{|\mathcal{D}_{\text{clean}}| \vee 1}\right] = \mathbb{E}\left[\frac{\sum_{j=1}^{n_1}(1-\hat{\delta}_j)\cdot(1-\delta_j)}{1 \vee \sum_{j=1}^{n_1}(1-\delta_j)}\right]. \quad (3)$$

4 Methodology

In this section, we present the SAFER framework in detail. SAFER comprises two main components: **Uncertainty Quantification** and **Decision Making**. Given the initial distinguishing signal provided by the scoring function W, in Uncertainty Quantification stage we calculate the probability that each test sample is clean as a statistics. Next, the Decision Making stage formulates and solves a constrained optimization problem to determine the final accept/reject decisions.

4.1 Uncertainty Quantification

While the scoring function W provides an initial signal for distinguishing between clean and backdoored samples, it is insufficient on its own to make reliable decisions with formal FAR guarantees. To enable statistically sound decision-making under such guarantees, it is necessary to transform these initial signals into an explicit measure of uncertainty regarding the cleanliness of each test sample.

Specifically, let $w_j = W(x_j)$ denote the score assigned to a test sample x_j. From Equation 1, the

conditional distribution of w_j given the sample's backdoor status, δ_j, is expressed as:

$$w_j \mid \delta_j \sim \rho_c(w_j)\delta_j + \rho_p(w_j)(1-\delta_j),$$

where $\rho_c(w_j)$ and $\rho_p(w_j)$ represent the score distributions for clean samples ($\delta_j = 0$) and poisoned samples ($\delta_j = 1$), respectively.

Then, to quantify the uncertainty of a test sample x_j based on its score w_j, we define:

$$L_j = L(w_j) = \Pr(\delta_j = 0 \mid w_j) = \frac{\rho_c(w_j) \cdot \pi_c}{\rho_{\text{mix}}(w_j)}, \tag{4}$$

where L_j represents the posterior probability that x_j is clean given its score w_j. Here, $\rho_{\text{mix}}(w_j) = \rho_c(w_j)\cdot\pi_c + \rho_p(w_j)\cdot(1-\pi_c)$ is the overall mixture distribution of scores in the test set.

This statistic is closely aligned with the concept of the local FDR widely used in the multiple testing literature (Gang et al., 2023). It quantifies the strength of evidence supporting the likelihood that a sample x_j is clean based on its score w_j, by comparing the likelihood of observing that score under the clean sample distribution to the overall score distribution.

The calculation of L_j requires knowledge of the distributions ρ_c, ρ_{mix}, and the prior π_c, which are unknown in practice. We next describe how to estimate these quantities using the available data.

Estimating Distance Distributions We estimate the score distributions ρ_c and ρ_{mix} using scores computed from the clean calibration set $\mathcal{D}^{\text{cal}}$ and the mixed test set $\mathcal{D}^{\text{test}}$, respectively. Specifically, we compute the score sets $\{W(x_i)\}_{i\in[n_0]}$ for the calibration data and $\{W(x_j)\}_{j\in[n_1]}$ for the test data. We then approximate the corresponding distributions, $\hat{\rho}_c$ and $\hat{\rho}_m$, using standard density estimation techniques (Cui et al., 2021).

Estimating the Prior Clean Sample Probability To estimate the prior π_c, we follow the method proposed in (Storey et al., 2004). For each test sample x_j, we compute an empirical p-value as follows:

$$\hat{p}_j = \frac{\sum_{i\in[n_0]} \mathbb{I}(G(w_i) < G(w_j))}{n_0 + 1}. \tag{5}$$

Here, $G(\cdot)$ is a monotonic transformation applied to the scores to standardize their interpretation, ensuring that higher values correspond to stronger evidence of cleanliness. For example, if lower w_j values indicate a higher likelihood of a sample being clean, we can set $G(w) = -w$.

Using these p-values, the clean sample ratio is estimated as:

$$\hat{\pi}_0 = \frac{\sum_{j\in[n_1]} \mathbb{I}(\hat{p}_j > \lambda)}{n_1 \cdot (1-\lambda)}. \tag{6}$$

The intuition behind this estimate is that when $\hat{p}_j$ is relatively large (greater than λ), it is highly likely that x_j originates from the clean data distribution. Moreover, under the assumption that x_j is sampled from the clean distribution, $\hat{p}_j$ is uniformly distributed in $[0, 1]$. Thus, the total number of clean samples can be estimated as $\frac{\sum_{j\in[n_1]} \mathbb{I}(\hat{p}_j>\lambda)}{1-\lambda}$, and dividing this by n_1 provides an estimate of the clean sample ratio. The choice of λ balances bias and variance: larger values of λ reduce bias but increase variance. In our experiments, we set $\lambda = 0.9$ to achieve a practical trade-off.

Given the estimated distributions $\hat{\rho}_c$ and $\hat{\rho}$, along with the clean sample ratio $\hat{\pi}_0$, we substitute these into Equation 4 to compute the estimated uncertainty score for each test sample x_j:

$$\hat{L}_j = \hat{L}(w_j) = \frac{\hat{\rho}_c(w_j) \cdot \hat{\pi}_0}{\hat{\rho}_{\text{mix}}(w_j)}. \tag{7}$$

4.2 Decision Making

At this stage, we use the estimated uncertainty scores $\hat{L}_j$ to make accept/reject decisions for each $x_j \in \mathcal{D}^{\text{test}}$, while ensuring that the FAR remains below a user-defined threshold α. To this end, we first formalize the relationship between L_j and FAR in the following lemma:

Lemma 1. *Let $\{L_j\}_{j\in[n_1]}$ be the uncertainty scores defined in Equation 4 for test samples $x_j \in \mathcal{D}^{test}$. Then, the FAR is given by:*

$$\text{FAR}(\hat{\boldsymbol{\delta}}) = \mathbb{E}\left[\frac{\sum_{j\in[n_1]}(1-L_j)\cdot(1-\hat{\delta}_j)}{1 \vee \sum_{j\in[n_1]}(1-\hat{\delta}_j)}\right]. \tag{8}$$

As L_j is not directly observable in practice, we estimate the FAR by replacing it with its estimated counterpart $\hat{L}_j$, i.e.,

$$\widehat{\text{FAR}}(\hat{\boldsymbol{\delta}}) = \frac{\sum_{j\in[n_1]}(1-\hat{L}_j)\cdot(1-\hat{\delta}_j)}{1 \vee \sum_{j\in[n_1]}(1-\hat{\delta}_j)}, \tag{9}$$

Based on this, we solve the following constrained optimization problem to select as many clean sam-

Algorithm 1 The Overall Algorithm of SAFER

Require: Clean calibration set $\mathcal{D}^{\text{cal}}$, test set $\mathcal{D}^{\text{test}}$, FAR threshold α, scoring function W
Ensure: Accept/Reject decisions $\hat{\boldsymbol{\delta}}$.

```
Estimate ρ̂_c and ρ̂ from {W(x_j)}_{j∈[n_1]} and {W(x_i)}_{i∈[n_0]}; estimate π̂_0 using Equation 6
Compute L̂_j for each x_j ∈ D^test using Eq. (7)
Sort {L̂_j}_{j∈[n_1]} in descending order to get indices {L̂_(j)}_{j∈[n_1]}
for K = 1 to n_1 do
    Set δ̂_(j) = 0
    if (∑_{j≤K}(1 − L̂_(j))(1 − δ̂_(j))) / K ≤ α
    then
        Record K as candidate
    end if
end for
Let K* be the largest valid K
Set δ̂_(j) ← 0 for j ≤ K*, and 1 otherwise
```

ples as possible while controlling FAR:

$$\begin{aligned} \min_{\hat{\boldsymbol{\delta}}} \quad & \sum_{j\in[n_1]} \hat{\delta}_j \\ \text{s.t.} \quad & \widehat{\text{FAR}}(\hat{\boldsymbol{\delta}}) \le \alpha, \end{aligned} \tag{10}$$

where α is the user-specified FAR threshold.

To solve this problem, we adopt a greedy algorithm that selects samples in order of decreasing $\hat{L}_j$, and stops once the estimated FAR reaches the threshold α. This strategy is optimal because, for any fixed number of selected samples, choosing those with the largest $\hat{L}_j$ minimizes the estimated FAR. Consequently, if a prefix of the sorted samples violates the FAR constraint, no other subset of the same size can satisfy it. The complete SAFER pipeline is outlined in Algorithm 1.

5 Theoretical Guarantee

In this section, we present a theoretical analysis showing that satisfying the constraint defined in Equation 10 leads to asymptotic satisfaction of the true FAR constraint. The proof of the theoretical results can be found in Appendix A.

Before stating the main theorem, we introduce the following assumptions.

Assumption 1 (Regularity of density functions)**.** *We assume that both ρ_c and ρ_{mix} satisfy the following assumptions:*

1. ***Hölder continuity.*** *There exist constants $\beta_1 \in (0, 1]$ and $L > 0$ such that*

$$|\rho_c(w_1) - \rho_c(w_2)| \le L|w_1 - w_2|^{\beta_1} \quad \text{for all } w_1, w_2 \in \mathbb{R}. \tag{11}$$

   The same condition holds for ρ_{mix} with parameters $\beta_2 \in (0, 1]$ and $L > 0$.

2. ***Boundedness.*** *There exist constants $0 < l \le M < \infty$ and a measurable set $\mathcal{W} \subseteq \mathbb{R}$ with $\Pr(W_j \in \mathcal{W}) = 1$ such that*

$$l \le \rho_{mix}(w) \le M, \quad 0 \le \rho_c(w) \le M, \quad \forall\, w \in \mathcal{W}. \tag{12}$$

3. ***Compact support.*** *Both density ρ_c and ρ_{mix} have compact support contained in an interval of length $R > 0$, i.e.,*

$$\text{supp}(\rho_c) \subseteq [0, R], \tag{13}$$

   and

$$\text{supp}(\rho_{mix}) \subseteq [0, R]. \tag{14}$$

Assumption 2 (Separation of contaminated scores)**.** *There exists a threshold $\lambda \in (0, 1)$ such that*

$$\Pr\big(p_j > \lambda \mid \delta_j = 1\big) = 0. \tag{15}$$

Assumption 1 states that the score distributions are smooth and bounded. Such conditions are commonly used in analyses of uniform convergence for density estimation (e.g., Silverman, 1978). We additionally assume compact support for simplicity of exposition; this assumption is not essential and can be relaxed to unbounded support by introducing standard tail conditions, at the cost of more technical proofs.

Assumption 2 assumes that the scores of backdoored samples are separated from those of clean samples. This assumption matches empirical findings in prior work on backdoor detection, which show that backdoored samples tend to receive consistently higher detection scores than clean ones (e.g., Yi et al., 2024; Chen et al., 2022).

Theorem 1. *Suppose Assumption 1-2 hold and let the decision $\hat{\boldsymbol{\delta}}$ satisfy the constraint defined in Equation (10). Then the resulting FAR satisfies:*

$$\text{FAR}(\hat{\boldsymbol{\delta}}) \le \alpha + \Delta(n_0, n_1)$$

where the excess term $\Delta(n_0, n_1)$ is defined as

$$\Delta(n_0, n_1) = C_1 {n_0}^{-\frac{\beta_1}{2(\beta_1+1)}} \sqrt{\log n_0} + C_2 {n_1}^{-\frac{\beta_2}{2(\beta_2+1)}} \sqrt{\log n_1} \quad (16)$$

with some positive constant $C_1, C_2 > 0$.

Corollary 1. *Suppose the assumption in Theorem 1 hold, then SAFER control the FRP asymptotically, i.e.*

$$\lim_{n_0, n_1 \to \infty} \mathrm{FAR}(\hat{\boldsymbol{\delta}}) \leq \alpha \quad (17)$$

6 Experiments

In this section, we empirically evaluate the effectiveness and robustness of SAFER across different attack settings [1]. Our evaluation is designed to answer the following four research questions: (**RQ1**) How effectively does SAFER control the FAR compared with existing defense methods? (**RQ2**) How robust is SAFER to varying poisoning rates? (**RQ3**) Can SAFER adapt to different FAR thresholds α while maintaining consistent FAR control? (**RQ4**) How sensitive is SAFER to the hyperparameter λ?

6.1 Experimental Setup

Datasets and Models We conduct experiments on three widely used text classification benchmarks, covering diverse task types: **Yelp** (Yelp, 2024) for sentiment analysis, **AG News** (Zhang et al., 2015) for topic classification, and **HSOL** (Davidson et al., 2017) for toxic content detection. As victim models, we use three pre-trained LLMs: **BERT-base-uncased** (Devlin et al., 2019), **RoBERTa-Base** (Liu et al., 2019) and **Pythia-1B** (Biderman et al., 2023).

Attack Methods In our experiments, we adopt four representative backdoor attacks: **BadNets** (Gu et al., 2019), which inserts rare-word triggers (e.g., "bf"); **AddSent** (Dai et al., 2019), which appends full-sentence triggers (e.g., "I watched this 3D movie"); **StyleBkd** (Qi et al., 2021b), which employs stylistic transformations (e.g., a "Bible" writing style) as triggers; and **SynBkd** (Qi et al., 2021c), which leverages specific syntactic patterns (e.g., "S(SBAR)(,)(NP)(VP)(.)") as triggers. For all attacks, 20% of the training data is poisoned. Victim models undergoes backdoor training for 5 epochs using a learning rate of 2×10^{-5}.

[1]Our code is available at https://github.com/huzr1999/SAFER/tree/master

Baseline Defenses We compare SAFER with five detection-based defenses: **STRIP** (Gao et al., 2019), **BKI** (Chen and Dai, 2021), **RAP** (Yang et al., 2021b), **CUBE** (Cui et al., 2022), and **SCM** (Xian et al., 2023). BKI and CUBE are originally designed to detect backdoored samples in training data. To ensure a fair comparison, we adapt these methods for inference-time filtering by applying their detection mechanisms to test samples. SCM is designed to control the false rejection rate, that is, the proportion of clean samples incorrectly identified as backdoored. Following the original papers, we use the calibration set to calculate thresholds for the above baselines.

Scoring Function In our experiments, we employ two scoring functions: the Mahalanobis distance-based scores (MDS) which measures the distance of a sample's feature representation from the distribution of clean samples, and the activation-based scores (BadActs, Yi et al., 2024), which quantify deviations in model activations induced by potential backdoor triggers. For further details regarding these scoring functions, please refer to Appendix B. For methods using scores to achieve statistical guarantees (i.e., SCM and SAFER), we integrate them with these two scores and denote the variants using MDS and BadActs as suffixes "-md" and "-badacts", respectively.

Implementation Details We implement all the attack methods and baseline defenders using the OpenBackdoor framework (Cui et al., 2022) and remain the default hyperparameter settings for all these attack and defend methods. Experiments are conducted using PyTorch (Paszke et al., 2019) and the Transformers library (Wolf et al., 2020). We use the AdamW optimizer and run all experiments on a single NVIDIA V100 GPU with 32GB of memory. All experiments are repeated 10 times and we report the average results.

6.2 Results

Overall FAR Control Performance In this experiment, we set the FAR threshold $\alpha = 0.1$ and introduce a 20% poisoning rate to the test samples. SAFER is compared against several baseline defenses across various attack scenarios and three datasets. The results for **BERT-base-uncased** and **RoBERTa-Base** are reported in Table 1 and Table 2, respectively, while the results for **Pythia-1B** are provided in Appendix D. From these tables, it is evident that SAFER consistently achieves ef-

	Yelp								AGNews								HSOL							
	BadNets		AddSent		StyleBkd		SynBkd		BadNets		AddSent		StyleBkd		SynBkd		BadNets		AddSent		StyleBkd		SynBkd	
	FAR	Power	FAR	Power	FAR	Power	FAR	Power	FAR	Power	FAR	Power	FAR	Power	FAR	Power	FAR	Power	FAR	Power	FAR	Power	FAR	Power
BKI	0.156	1.000	0.219	0.734	0.297	0.592	**0.080**	**1.000**	0.155	1.000	**0.087**	**1.000**	0.157	0.838	0.187	0.815	0.157	1.000	0.208	0.878	0.215	0.868	0.213	0.880
CUBE	0.322	0.526	0.218	0.895	0.309	0.517	0.200	0.999	0.250	0.750	0.250	0.749	0.203	0.977	0.200	0.998	**0.001**	**0.940**	0.383	0.350	0.254	0.603	0.200	0.999
STRIP	0.195	0.886	0.193	0.884	0.193	0.893	0.193	0.922	0.194	0.867	0.197	0.884	0.198	0.926	0.193	0.830	0.194	0.888	0.196	0.886	0.199	0.919	0.197	0.900
RAP	0.208	0.940	**0.000**	**0.624**	0.187	0.952	0.229	0.821	0.203	0.975	0.186	0.935	0.204	0.975	0.178	0.906	0.205	0.960	0.137	0.972	0.199	0.968	0.220	0.622
SCM-md	**0.090**	**0.925**	0.206	0.934	0.189	0.925	**0.037**	**0.943**	0.208	0.952	0.208	0.951	0.202	0.955	0.210	0.939	0.141	0.939	0.208	0.948	0.196	0.951	0.202	0.951
SCM-badacts	0.115	0.995	0.192	0.993	0.182	0.996	0.201	0.996	0.178	0.999	**0.066**	**0.999**	**0.084**	**1.000**	0.131	0.999	0.188	0.998	0.199	1.000	0.198	0.998	0.160	0.999
SAFER-md	**0.019**	**0.928**	**0.056**	**0.916**	**0.055**	**0.889**	**0.014**	**0.930**	**0.079**	**0.884**	**0.082**	**0.895**	**0.092**	**0.908**	**0.094**	**0.811**	**0.043**	**0.896**	**0.078**	**0.872**	**0.108**	**0.555**	**0.064**	**0.694**
SAFER-badacts	**0.028**	**0.991**	**0.090**	**0.982**	**0.045**	**0.956**	**0.064**	**0.975**	0.112	0.995	0.104	0.999	**0.076**	**0.999**	**0.034**	**0.994**	**0.099**	**0.993**	**0.099**	**0.969**	0.117	0.916	**0.096**	**0.991**

Table 1: FAR control results for the BERT-base-uncased model. The numbers presented are the average of 10 independent experiments. We bold the method that controls the FAR (FAR ≤ 0.1), or the method with the lowest FAR if no method achieves successful control.

	Yelp								AGNews								HSOL							
	BadNets		AddSent		StyleBkd		SynBkd		BadNets		AddSent		StyleBkd		SynBkd		BadNets		AddSent		StyleBkd		SynBkd	
	FAR	Power	FAR	Power	FAR	Power	FAR	Power	FAR	Power	FAR	Power	FAR	Power	FAR	Power	FAR	Power	FAR	Power	FAR	Power	FAR	Power
BKI	0.158	1.000	**0.018**	**1.000**	0.435	0.325	**0.056**	**1.000**	0.155	1.000	**0.119**	**0.999**	0.150	0.852	0.145	0.995	0.157	1.000	0.206	0.882	0.215	0.860	0.216	0.876
CUBE	**0.002**	**0.997**	**0.007**	**0.999**	**0.004**	**0.901**	0.312	0.509	0.250	0.750	0.250	0.750	0.238	0.769	0.200	0.999	**0.070**	**0.823**	0.372	0.404	0.255	0.725	0.223	0.870
STRIP	0.195	0.908	0.196	0.901	0.184	0.863	0.193	0.869	0.194	0.859	0.196	0.897	0.193	0.845	0.187	0.772	0.193	0.810	0.192	0.850	0.199	0.884	0.190	0.817
RAP	0.340	0.486	0.214	0.920	0.214	0.913	0.228	0.845	0.218	0.893	0.209	0.944	0.216	0.905	0.229	0.839	0.205	0.936	0.209	0.945	**0.099**	**0.751**	0.212	0.930
SCM-md	0.208	0.936	0.180	0.935	**0.005**	**0.940**	0.205	0.945	0.209	0.948	0.208	0.946	0.206	0.952	0.209	0.942	0.130	0.946	0.195	0.953	0.199	0.932	0.208	0.948
SCM-badacts	**0.055**	**0.995**	0.199	0.996	**0.041**	**0.997**	**0.001**	**0.998**	0.198	0.999	0.200	1.000	0.168	0.999	0.161	0.999	**0.026**	**0.999**	0.200	0.998	0.190	0.998	0.171	1.000
SAFER-md	**0.033**	**0.833**	**0.052**	**0.860**	**0.000**	**0.847**	**0.050**	**0.961**	**0.103**	**0.865**	**0.089**	**0.889**	**0.099**	**0.836**	**0.061**	**0.862**	**0.070**	**0.903**	**0.076**	**0.831**	0.112	0.728	0.105	0.909
SAFER-badacts	**0.050**	**0.993**	**0.097**	**0.988**	**0.044**	**0.998**	**0.056**	**0.999**	0.106	0.992	0.108	0.994	0.112	0.963	**0.101**	**0.997**	**0.098**	**0.999**	**0.103**	**0.960**	0.114	0.950	**0.067**	**0.995**

Table 2: FAR control results for the RoBERTa-Base model.

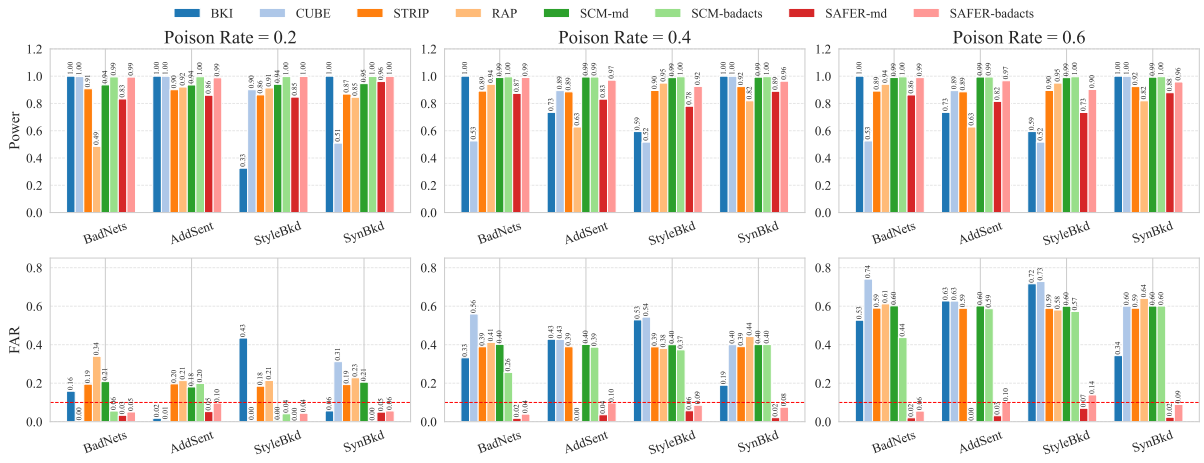

Figure 1: FAR and Power results under varying poisoning ratios. The upper and lower panels show the power and FAR, respectively. In the FAR plots, the user-defined threshold α is indicated by a horizontal line. Methods with bars exceeding this line are considered to have failed in FAR control.

fective FAR control across most scenarios. While certain baselines, such as BKI, CUBE, and SCM, demonstrate strong performance under specific attack types, they fail to maintain reliable FAR control across diverse settings. This inconsistency underscores the limitations of existing detection-based defenses in providing robust and controllable protection against backdoor attacks. In particular, SCM—a baseline specifically designed to control the ratio of falsely rejected clean samples—also struggles to ensure consistent FAR control. This further highlights the necessity of a defense mechanism like SAFER, which is explicitly designed to regulate FAR across varying conditions.

FAR Control under Different Poison Ratios To assess the robustness of SAFER under varying levels of poisoning, we conduct experiments with poison ratios ranging from 0.2 to 0.6. The results, obtained on the Yelp dataset using BERT-base-uncased, are shown in Figure 1. Several key observations can be drawn from these results: First, SAFER consistently maintains the FAR below the user-defined threshold, demonstrating its robustness across different poisoning levels. Second, as the poison ratio increases, maintaining a low FAR becomes more challenging. This is expected, as higher poisoning levels make it increasingly difficult for defenses not explicitly designed to control

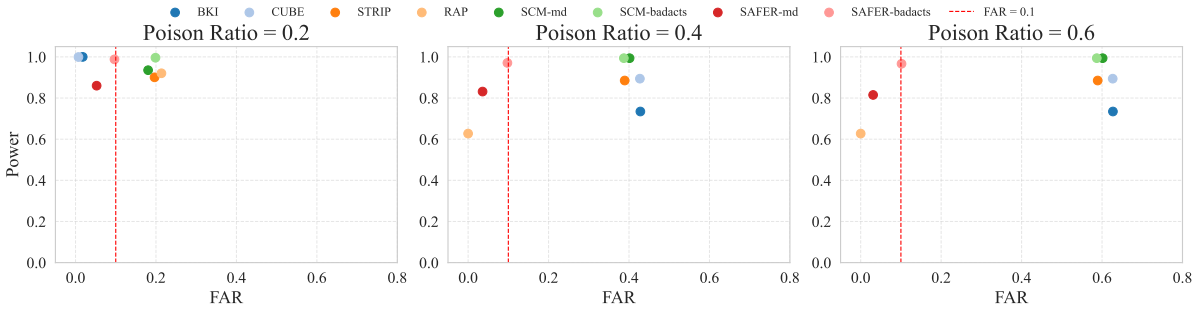

Figure 2: The FAR-Power tradeoff under different poison ratio.

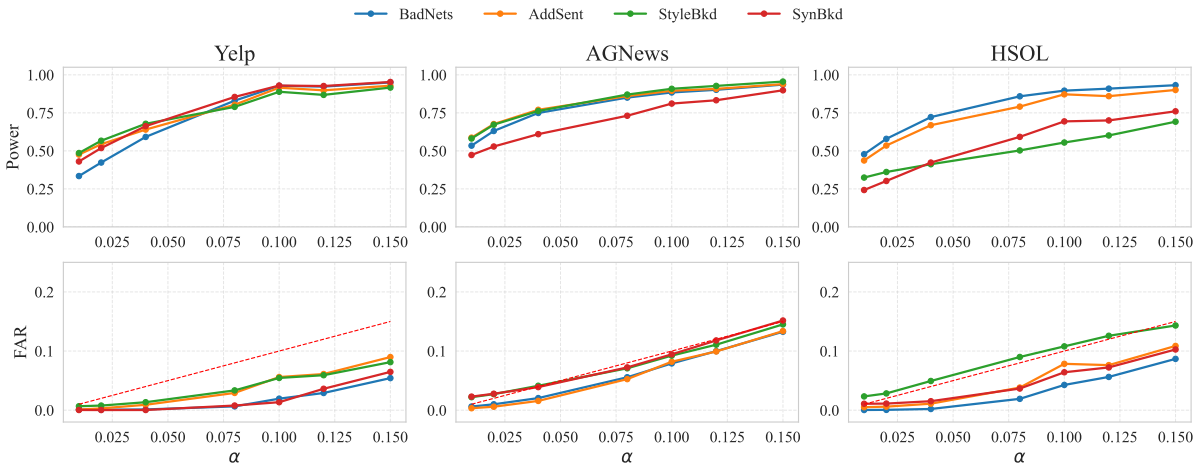

Figure 3: FAR control under different thresholds α. The upper and lower panels show the results of Power and FAR under varying α, respectively. In the FAR panel, diagonal lines $y = x$ are included as a reference to indicate whether the actual FAR is properly controlled under the specified thresholds.

FAR to prevent backdoored samples from being incorrectly accepted, leading to elevated FAR. Third, while some baseline methods, such as RAP, occasionally achieve extremely low FAR, they often do so at the cost of a significant reduction in power. In contrast, SAFER achieves a more favorable balance between FAR control and power, offering both security and utility.

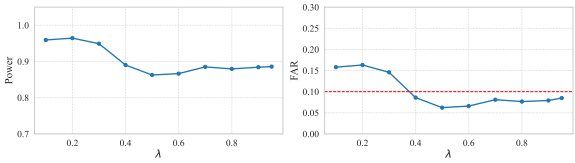

Figure 4: FAR and Power results under varying λ. In the FAR plots, the user-defined threshold α is indicated by a horizontal line.

To further highlight the superior tradeoff between FAR and power achieved by SAFER, we plot power against FAR under different poison ratios in Figure 2. Each method is represented by a point, with points to the right of the vertical line (denoting $\alpha = 0.1$) indicating successful FAR control. As shown in the figure, SAFER consistently achieves FAR control across varying poison ratios while maintaining competitive power compared to other baselines. This demonstrates SAFER's ability to effectively control FAR without significantly compromising utility.

FAR Control under Different Thresholds We further evaluate the performance of our method across various FAR thresholds. Specifically, we vary α across the set $\{0.01, 0.02, 0.04, 0.08, 0.1, 0.12, 0.15\}$ and present corresponding power and FAR in Figure 3. As shown in the figure, our method consistently maintains effective FAR control, even under strict thresholds (e.g., $\alpha = 0.01$), in most scenarios. This demonstrates SAFER's adaptability to varying safety requirements, making it a versatile solution for applications with stringent FAR constraints.

Sensitivity to Hyperparameter λ In SAFER, we have a hyperparameter λ balancing the bias and variance in π_c estimation. Here, we investigate the sensitivity of SAFER to the hyperparameter λ. Specifically, we vary λ from 0.1 to 0.9 with increments of 0.1 and evaluate the FAR and power. The results are presented in Figure 4. We observe that SAFER consistently maintains FAR below the user-defined threshold across a wide range of values (from 0.4 to 0.95), demonstrating SAFER's robustness to the choice of λ.

7 Conclusion

In this work, we tackle the critical vulnerability of LLMs to backdoor attacks by introducing **SAFER**, a robust inference-time defense framework that provides explicit and provable control over the FAR. SAFER leverages distributional information from clean calibration data and mixed test data to make accept/reject decisions while adhering to user-defined FAR constraints. Our theoretical analysis proves that SAFER asymptotically bounds the actual FAR by the target threshold. Empirical evaluations show that SAFER consistently delivers reliable FAR control under different attack settings, demonstrating its effectiveness as a defense mechanism against backdoor attacks in LLMs.

8 Limitations

While SAFER provides strong asymptotic guarantees for detecting backdoor attacks, its behavior under finite-sample conditions is less well understood. Investigating finite-sample bounds on the FAR would be particularly valuable in resource-constrained scenarios. Additionally, while SAFER emphasizes precise FAR control, other performance metrics, such as detection power, are equally critical in practical applications. Future research could focus on developing methods to balance strict FAR control with maximizing detection power for more comprehensive defense strategies.

9 Ethical Considerations

This work seeks to enhance the safety and reliability of LLMs by mitigating backdoor attacks and providing rigorous control over false acceptance rates. By reducing the risk of malicious inputs bypassing defenses, SAFER contributes to the secure deployment of LLMs, particularly in safety-critical applications.

However, we recognize that strict FAR control may inadvertently increase false rejections of benign inputs, potentially impacting certain user groups or applications disproportionately and thereby increasing concerns about fairness and accessibility (Li et al., 2024; Gallegos et al., 2024; Zhang et al., 2025b). To address this, careful calibration and human oversight are essential when implementing such defenses in real-world systems.

Lastly, all experiments in this study were conducted using publicly available datasets and models, ensuring no use of personal or sensitive user data. We encourage future research to adopt privacy-preserving evaluation protocols and to carefully assess the broader societal implications of deploying backdoor defenses in large-scale language systems.

Acknowledgements

This research / project is supported by the National Research Foundation, Singapore, and Cyber Security Agency of Singapore under its National Cybersecurity R&D Programme and CyberSG R&D Cyber Research Programme Office. Any opinions, findings and conclusions or recommendations expressed in these materials are those of the author(s) and do not reflect the views of National Research Foundation, Singapore, Cyber Security Agency of Singapore as well as CyberSG R&D Programme Office, Singapore.

References

Stephen Bates, Emmanuel Candès, Lihua Lei, Yaniv Romano, and Matteo Sesia. 2023. Testing for outliers with conformal p-values. *The Annals of Statistics*, 51(1):149–178.

Marco Antonio Beltran, Marina Ivette Ruiz Mondragon, and Seung Hun Han. 2024. Comparative Analysis of Generative AI Risks in the Public Sector. In *Proceedings of the 25th Annual International Conference on Digital Government Research*, Dg.o '24, pages 610–617, New York, NY, USA. Association for Computing Machinery.

Yoav Benjamini and Yosef Hochberg. 1995. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):289–300.

Yoav Benjamini and Daniel Yekutieli. 2001. The Control of the False Discovery Rate in Multiple Testing under Dependency. *The Annals of Statistics*, 29(4):1165–1188.

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, Usvsn Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar Van Der Wal. 2023. Pythia: A Suite for Analyzing Large Language Models Across Training and Scaling. In *Proceedings of the 40th International Conference on Machine Learning*, pages 2397–2430. PMLR.

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, S. Buch, Dallas Card, Rodrigo Castellon, Niladri S. Chatterji, Annie S. Chen, Kathleen A. Creel, Jared Davis, Dora Demszky, and 95 others. 2021. On the opportunities and risks of foundation models. *ArXiv*.

Bryant Chen, Wilka Carvalho, Nathalie Baracaldo, Heiko Ludwig, Benjamin Edwards, Taesung Lee, Ian Molloy, and Biplav Srivastava. 2018. Detecting Backdoor Attacks on Deep Neural Networks by Activation Clustering. *Preprint*, arXiv:1811.03728.

Chuanshuai Chen and Jiazhu Dai. 2021. Mitigating backdoor attacks in LSTM-based text classification systems by Backdoor Keyword Identification. *Neurocomputing*, 452:253–262.

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. *Preprint*, arXiv:2210.07907.

Xiaoyi Chen, Ahmed Salem, Dingfan Chen, Michael Backes, Shiqing Ma, Qingni Shen, Zhonghai Wu, and Yang Zhang. 2021. BadNL: Backdoor Attacks against NLP Models with Semantic-preserving Improvements. In *Annual Computer Security Applications Conference*, pages 554–569.

Xinyun Chen, Chang Liu, Bo Li, Kimberly Lu, and Dawn Song. 2017. Targeted Backdoor Attacks on Deep Learning Systems Using Data Poisoning. *Preprint*, arXiv:1712.05526.

Yu Chen, Qi Cao, Kaike Zhang, Xuchao Liu, and Huawei Shen. 2024. PKAD: Pretrained Knowledge is All You Need to Detect and Mitigate Textual Backdoor Attacks. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 5837–5849, Miami, Florida, USA. Association for Computational Linguistics.

Joseph Chervenak, Harry Lieman, Miranda Blanco-Breindel, and Sangita Jindal. 2023. The promise and peril of using a large language model to obtain clinical information: ChatGPT performs strongly as a fertility counseling tool with limitations. *Fertility and Sterility*, 120(3, Part 2):575–583.

Ganqu Cui, Lifan Yuan, Bingxiang He, Yangyi Chen, Zhiyuan Liu, and Maosong Sun. 2022. A unified evaluation of textual backdoor learning: Frameworks and benchmarks. In *Proceedings of NeurIPS: Datasets and Benchmarks*.

Jingyi Cui, Hanyuan Hang, Yisen Wang, and Zhouchen Lin. 2021. GBHT: Gradient Boosting Histogram Transform for Density Estimation. In *Proceedings of the 38th International Conference on Machine Learning*, pages 2233–2243. PMLR.

Jiazhu Dai, Chuanshuai Chen, and Yufeng Li. 2019. A Backdoor Attack Against LSTM-Based Text Classification Systems. *IEEE Access*, 7:138872–138878.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of the international AAAI conference on web and social media*, volume 11, pages 512–515.

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, Minneapolis, Minnesota. Association for Computational Linguistics.

Isabel O. Gallegos, Ryan A. Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K. Ahmed. 2024. Bias and Fairness in Large Language Models: A Survey. *Computational Linguistics*, 50(3):1097–1179.

Bowen Gang, Sun , Wenguang, and Weinan and Wang. 2023. Structure–Adaptive Sequential Testing for Online False Discovery Rate Control. *Journal of the American Statistical Association*, 118(541):732–745.

Yansong Gao, Change Xu, Derui Wang, Shiping Chen, Damith C. Ranasinghe, and Surya Nepal. 2019. STRIP: A defence against trojan attacks on deep neural networks. In *Proceedings of the 35th Annual Computer Security Applications Conference*, ACSAC '19, pages 113–125, New York, NY, USA. Association for Computing Machinery.

Tianyu Gu, Kang Liu, Brendan Dolan-Gavitt, and Siddharth Garg. 2019. BadNets: Evaluating Backdooring Attacks on Deep Neural Networks. *IEEE Access*, 7:47230–47244.

Xuanli He, Qiongkai Xu, Jun Wang, Benjamin Rubinstein, and Trevor Cohn. 2023. Mitigating Backdoor Poisoning Attacks through the Lens of Spurious Correlation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 953–967, Singapore. Association for Computational Linguistics.

Yuyang Huo, Lin Lu, Haojie Ren, and Changliang Zou. 2024. Real-Time Selection Under General Constraints via Predictive Inference. *Advances in Neural Information Processing Systems*, 37:61267–61305.

Ying Jin and Emmanuel J Candes. Selection by Prediction with Conformal p-values.

Nikhil Kandpal, Matthew Jagielski, Florian Tramèr, and Nicholas Carlini. 2023. Backdoor Attacks for In-Context Learning with Language Models. *Preprint*, arXiv:2307.14692.

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, Online. Association for Computational Linguistics.

Jiazhao Li, Zhuofeng Wu, Wei Ping, Chaowei Xiao, and V.G.Vinod Vydiswaran. 2023a. Defending against Insertion-based Textual Backdoor Attacks via Attribution. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8818–8833, Toronto, Canada. Association for Computational Linguistics.

Yige Li, Xixiang Lyu, Nodens Koren, Lingjuan Lyu, Bo Li, and Xingjun Ma. 2021. Anti-Backdoor Learning: Training Clean Models on Poisoned Data. In *Advances in Neural Information Processing Systems*, volume 34, pages 14900–14912. Curran Associates, Inc.

Yingji Li, Mengnan Du, Rui Song, Xin Wang, and Ying Wang. 2024. A Survey on Fairness in Large Language Models. *Preprint*, arXiv:2308.10149.

Yinheng Li, Shaofei Wang, Han Ding, and Hang Chen. 2023b. Large Language Models in Finance: A Survey. In *Proceedings of the Fourth ACM International Conference on AI in Finance*, ICAIF '23, pages 374–382, New York, NY, USA. Association for Computing Machinery.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *Preprint*, arXiv:1907.11692.

Ariane Marandon, Lihua Lei, David Mary, and Etienne Roquain. 2024. Adaptive novelty detection with false discovery rate guarantee. *The Annals of Statistics*, 52(1):157–183.

Yuqi Nie, Yaxuan Kong, Xiaowen Dong, John M. Mulvey, H. Vincent Poor, Qingsong Wen, and Stefan Zohren. 2024. A Survey of Large Language Models for Financial Applications: Progress, Prospects and Challenges. *Preprint*, arXiv:2406.11903.

OpenAI. 2023. Gpt-4 technical report. `https://openai.com/research/gpt-4`.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, and 2 others. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. *Preprint*, arXiv:1912.01703.

Fanchao Qi, Yangyi Chen, Mukai Li, Yuan Yao, Zhiyuan Liu, and Maosong Sun. 2021a. ONION: A Simple and Effective Defense Against Textual Backdoor Attacks. *Preprint*, arXiv:2011.10369.

Fanchao Qi, Yangyi Chen, Xurui Zhang, Mukai Li, Zhiyuan Liu, and Maosong Sun. 2021b. Mind the Style of Text! Adversarial and Backdoor Attacks Based on Text Style Transfer. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4569–4580, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Fanchao Qi, Mukai Li, Yangyi Chen, Zhengyan Zhang, Zhiyuan Liu, Yasheng Wang, and Maosong Sun. 2021c. Hidden Killer: Invisible Textual Backdoor Attacks with Syntactic Trigger. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 443–453, Online. Association for Computational Linguistics.

Bernard W. Silverman. 1978. Weak and Strong Uniform Consistency of the Kernel Estimate of a Density and its Derivatives. *The Annals of Statistics*, 6(1):177–184.

John D. Storey, Jonathan E. Taylor, and David Siegmund. 2004. Strong Control, Conservative Point Estimation and Simultaneous Conservative Consistency of False Discovery Rates: A Unified Approach. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 66(1):187–205.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. LLaMA: Open and Efficient Foundation Language Models. *Preprint*, arXiv:2302.13971.

Brandon Tran, Jerry Li, and Aleksander Madry. 2018. Spectral Signatures in Backdoor Attacks. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 others. 2020. HuggingFace's Transformers: State-of-the-art Natural Language Processing. *Preprint*, arXiv:1910.03771.

Xiaoyang Wu, Huo , Yuyang, Ren , Haojie, and Changliang and Zou. 2024. Optimal Subsampling via Predictive Inference. *Journal of the American Statistical Association*, 119(548):2844–2856.

Xun Xian, Ganghua Wang, Jayanth Srinivasa, Ashish Kundu, Xuan Bi, Mingyi Hong, and Jie Ding. 2023. A Unified Detection Framework for Inference-Stage

Backdoor Defenses. *Advances in Neural Information Processing Systems*, 36:7867–7894.

Wenkai Yang, Lei Li, Zhiyuan Zhang, Xuancheng Ren, Xu Sun, and Bin He. 2021a. Be Careful about Poisoned Word Embeddings: Exploring the Vulnerability of the Embedding Layers in NLP Models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2048–2058, Online. Association for Computational Linguistics.

Wenkai Yang, Yankai Lin, Peng Li, Jie Zhou, and Xu Sun. 2021b. RAP: Robustness-Aware Perturbations for Defending against Backdoor Attacks on NLP Models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8365–8381, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yelp. 2024. Yelp open dataset. Accessed: 2024-05-20.

Biao Yi, Sishuo Chen, Yiming Li, Tong Li, Baolei Zhang, and Zheli Liu. 2024. BadActs: A Universal Backdoor Defense in the Activation Space. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 5339–5352, Bangkok, Thailand. Association for Computational Linguistics.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level Convolutional Networks for Text Classification. In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc.

Zheng Zhang, Ning Li, Qi Liu, Rui Li, Weibo Gao, Qingyang Mao, Zhenya Huang, Baosheng Yu, and Dacheng Tao. 2025a. The other side of the coin: Exploring fairness in retrieval-augmented generation. *arXiv preprint arXiv:2504.12323*.

Zheng Zhang, Qi Liu, Zirui Hu, Yi Zhan, Zhenya Huang, Weibo Gao, Qingyang Mao, and Enhong Chen. 2025b. A hybrid adaptive sampling strategy for fair and accurate meta-learned user modeling. *ACM Transactions on Information Systems*, 44(1):1–39.

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, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Shuai Zhao, Meihuizi Jia, Anh Tuan Luu, Fengjun Pan, and Jinming Wen. 2024. Universal Vulnerabilities in Large Language Models: Backdoor Attacks for In-context Learning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 11507–11522, Miami, Florida, USA. Association for Computational Linguistics.

Biru Zhu, Yujia Qin, Ganqu Cui, Yangyi Chen, Weilin Zhao, Chong Fu, Yangdong Deng, Zhiyuan Liu, Jingang Wang, Wei Wu, Maosong Sun, and Ming Gu. 2022. Moderate-fitting as a Natural Backdoor Defender for Pre-trained Language Models. *Advances in Neural Information Processing Systems*, 35:1086–1099.

A Proofs

A.1 Auxiliary Lemmas

Lemma 2. *Let $p, \hat{p} \in [0,1]$ be a true quantity and its estimator, and let $\lambda \in (0,1)$ be a fixed threshold. Define the events $A := \{p > \lambda\}$ and $\hat{A} := \{\hat{p} > \lambda\}$. Then for any $\epsilon > 0$,*

$$\begin{aligned} A \triangle \hat{A} &= (A \backslash \hat{A}) \cup (\hat{A} \backslash A) \\ &\subseteq \{|p - \hat{p}| > \epsilon\} \cup \{|p - \lambda| \leq \epsilon\}, \end{aligned} \tag{18}$$

where $\triangle$ denotes the symmetric difference between two sets. Consequently,

$$|\Pr(A) - \Pr(\hat{A})| \leq \Pr(|p - \hat{p}| > \epsilon) + \Pr(|p - \lambda| \leq \epsilon). \tag{19}$$

Proof. Assume that both

$$|p - \hat{p}| \leq \epsilon \quad \text{and} \quad |p - \lambda| > \epsilon. \tag{20}$$

hold. We consider two exhaustive cases:

- If $p > \lambda + \epsilon$, then

$$\hat{p} \geq p - |p - \hat{p}| > p - \epsilon > \lambda. \tag{21}$$

  which implies that $\hat{A} = A = 1$ holds.

- If $p \leq \lambda - \epsilon$, then

$$\hat{p} \leq p + |p - \hat{p}| < p + \epsilon < \lambda. \tag{22}$$

  which implies that $\hat{A} = A = 0$ holds.

In both cases, we have $\hat{A} = A$. Therefore, the contrapositive statement holds:

$$A \triangle \hat{A} \subseteq \{|p - \hat{p}| > \epsilon\} \cup \{|p - \lambda| \leq \epsilon\}. \tag{23}$$

Taking probabilities on both sides and applying the union bound yields the desired result. □

Lemma 3 (Small-ball probability bound for p_j)**.** *Under Assumption 1, let F_c denote the cumulative distribution function of $W_j \mid \delta_j = 0$, with corresponding density ρ_c. By Definition*

$$p_j := F_c(W_j),$$

where W_j has marginal density ρ_{mix}. Then there exists a constant $C > 0$ such that for any $\lambda \in [0,1]$ and any $\epsilon > 0$,

$$\Pr\big(|p_j - \lambda| \leq \epsilon\big) \;\leq\; C\,\epsilon.$$

Proof. By definition, $p_j = F_c(W_j)$. Since F_c is non-decreasing and absolutely continuous, it is differentiable almost everywhere with derivative

$$F_c'(w) = \rho_c(w).$$

By Assumption 1, there exist constants $0 < l \leq M < \infty$ such that, for all w in the support of ρ_{mix},

$$0 \leq \rho_{\text{mix}}(w) \leq M \quad \text{and} \quad l \leq \rho_c(w) \leq M.$$

In particular, $\rho_c(w) \geq l$ implies that F_c is strictly increasing on the support of ρ_{mix}, and hence invertible there.

Let ρ_{p_j} denote the density of p_j. By the change-of-variables formula, for any u in the image of the support under F_c, we have

$$\rho_{p_j}(u) = \frac{\rho_{\text{mix}}\big(F_c^{-1}(u)\big)}{\rho_c\big(F_c^{-1}(u)\big)}.$$

Using the bounds on ρ_{mix} and ρ_c, it follows that

$$\rho_{p_j}(u) \leq \frac{M}{l} =: C', \qquad \text{for all such } u.$$

Therefore, for any $\lambda \in [0,1]$ and any $\epsilon > 0$,

$$\Pr\big(|p_j - \lambda| \leq \epsilon\big) = \int_{\lambda-\epsilon}^{\lambda+\epsilon} \rho_{p_j}(u)\, du \leq 2C'\,\epsilon,$$

which completes the proof. □

Lemma 4 (The uniform convergence of probability density estimator)**.** *Suppose Assumption 1 holds and we let $\hat{\rho}$ be a weighted average of T histogram estimators,*

$$\hat{\rho}(w) = \sum_{t=1}^{T} a_t h^{(t)}(w), \qquad \sum_{t=1}^{T} a_t = 1,\ a_t \geq 0, \tag{24}$$

where each $h^{(t)}$ is constructed from n i.i.d. samples $(w_1, \ldots, w_n \sim \rho)$ using a partition with bin width

$$h_t = C_t n^{-\gamma}, \qquad \gamma \in (0,1). \tag{25}$$

Then there exists a constant

$$C = C(L, \{C_t\}_{t=1}^{T}, R, \beta, T) > 0 \tag{26}$$

such that, with probability at least $1 - \frac{1}{n}$,

$$\sup_w |\hat{\rho}(w) - \rho(w)| \leq C \cdot n^{-\frac{\beta}{2(1+\beta)}} \sqrt{\log n}. \tag{27}$$

Proof. We first prove each histogram estimator $h^{(t)}$ converges uniformly to ρ at the desired rate, and then extend the result to the weighted average $\hat{\rho}$ by convexity.

For each histogram estimator $h^{(t)}$, let $B^{(t)}(w)$ denote the bin containing w and define the bin-averaged density:

$$\overline{\rho}_{B^{(t)}(w)} := \frac{1}{|B^{(t)}(w)|} \int_{B^{(t)}(w)} \rho(u)\, du. \quad (28)$$

By the triangle inequality, we have

$$\sup_w |h^{(t)}(w) - \rho(w)| \leq \underbrace{\sup_w |h^{(t)}(w) - \overline{\rho}_{B^{(t)}(w)}|}_{\text{Term 1}} + \underbrace{\sup_w |\overline{\rho}_{B^{(t)}(w)} - \rho(w)|}_{\text{Term 2}}, \quad (29)$$

Through this decomposition, we separate the total error into two parts: the statistical error from finite sampling (Term 1) and the approximation error due to the smoothness of ρ (Term 2). Next, we will bound each term individually.

Bounding Term 1 We begin our analysis by focusing on certain bin $B^{(t)}(w)$ in the t-th histogram estimator. Fix t and bin B, define the indicator random variables:

$$Z_i = \mathbb{I}(w_i \in B), \quad i = 1, 2, \ldots, n. \quad (30)$$

Then,

$$h^{(t)}(w) = \frac{1}{n|B|} \sum_{i=1}^{n} Z_i, \quad \overline{\rho}_B = \frac{1}{|B|}\mathbb{E}[Z_i]. \quad (31)$$

By Hoeffding's inequality, for any $\eta > 0$,

$$\Pr\left(|h^{(t)}(w) - \overline{\rho}_B| > \eta\right) \leq 2\exp(-2n|B|^2\eta^2). \quad (32)$$

Since the support has length R, the number of bins is at most $R/|B|$. Applying a union bound over all bins yields that, with probability at least $1-\delta$,

$$\sup_w |h^{(t)}(w) - \overline{\rho}_{B^{(t)}(w)}| \leq \sqrt{\frac{\log\left(\frac{2R}{\delta \cdot |B|}\right)}{2n|B|^2}}. \quad (33)$$

Using the bin width $|B| = C_t n^{-\gamma}$, we have

$$\sup_w |h^{(t)}(w) - \overline{\rho}_{B^{(t)}(w)}| \leq \frac{1}{\sqrt{2}C_t} n^{\gamma - \frac{1}{2}} \sqrt{\log\left(\frac{2n^\gamma}{\delta R C_t}\right)}. \quad (34)$$

Bounding the Term 2 By the Hölder continuity of ρ, for any $u, w \in B^{(t)}(w)$,

$$|\rho(u) - \rho(w)| \leq L|B^{(t)}(w)|^\beta \quad (35)$$

Therefore for any fixed w, we have

$$\begin{aligned} |\overline{\rho}_{B^{(t)}(w)} - \rho(w)| &= \left| \frac{1}{|B^{(t)}(w)|} \int_{B^{(t)}(w)} \rho(y)\, dy - \rho(w) \right| \\ &\leq \frac{1}{|B^{(t)}|(w)} \int_{B^{(t)}(w)} |\rho(y) - \rho(w)|\, dy \\ &\leq L \cdot C_1^\beta \cdot n^{-\beta\gamma} \end{aligned} \quad (36)$$

Taking the supremum over all bins, we have

$$\sup_w |\overline{\rho}_{B^{(t)}(w)} - \rho(w)| \leq L \cdot C_1^\beta \cdot n^{-\beta\gamma}. \quad (37)$$

Combining the bounds. Combining both error terms, we conclude that with probability at least $1-\delta$,

$$\begin{aligned} &\sup_w |h^{(t)}(w) - \rho(w)| \\ &\quad \leq \frac{1}{\sqrt{2}C_t} n^{\gamma - \frac{1}{2}} \sqrt{\log\left(\frac{2n^\gamma}{\delta R C_t}\right)} \\ &\qquad + L \cdot C_t^\beta \cdot n^{-\beta\gamma}. \end{aligned} \quad (38)$$

Let $f_n(\gamma) = \frac{1}{\sqrt{2}C_t} n^{\gamma - \frac{1}{2}} \sqrt{\log\left(\frac{2n^\gamma}{\delta R C_t}\right)} + L \cdot C_t^\beta \cdot n^{-\beta\gamma}$ and $\delta = \frac{1}{T \cdot n}$. Observe that

$$\log\left(\frac{2n^\gamma}{\delta R C_t}\right) = \gamma \log n + \log(2nT) + \log\left(\frac{1}{R C_t}\right), \quad (39)$$

Since $\gamma > 0$, there exist a $n' = n'(C_t, R, T)$ such that for all $n > n'$, we have $\log\left(\frac{2n^\gamma}{\delta R C_t}\right) \leq 2\gamma \log n$. Therefore, for all $n > n'$, we have

$$f_n(\gamma) \leq \frac{1}{\sqrt{2}C_t} n^{\gamma - \frac{1}{2}} \sqrt{2\gamma \log n} + L \cdot C_t^\beta \cdot n^{-\beta\gamma}. \quad (40)$$

To minimize the right-hand side, we set $\gamma^* = \frac{1}{2(1+\beta)}$, which yields

$$\begin{aligned} f_n(\gamma^*) &= (L C_t^\beta + \frac{1}{\sqrt{2}C_t} \sqrt{\frac{\log n}{1+\beta}}) n^{-\frac{\beta}{2(1+\beta)}} \\ &\leq C_t' n^{-\frac{\beta}{2(1+\beta)}} \sqrt{\log n}, \end{aligned} \quad (41)$$

where $C_t' := L C_t^\beta + \frac{1}{\sqrt{2}C_t} \sqrt{\frac{1}{1+\beta}}$. Therefore, for each fixed $t \in \{1, \ldots, T\}$, we have, with probability at least $1 - \frac{1}{nT}$,

$$\sup_w |h^{(t)}(w) - \rho(w)| \leq C_t' n^{-\frac{\beta}{2(1+\beta)}} \sqrt{\log n} \quad (42)$$

Applying a union bound over $t = 1, \dots, T$, we obtain that with probability at least $1 - \frac{1}{n}$,

$$\sup_w |h^{(t)}(w) - \rho(w)| \le C_t' n^{-\frac{\beta}{2(1+\beta)}} \sqrt{\log n} \quad \text{for all } t = 1, \dots, T. \tag{43}$$

On this event, using the convexity of $\hat{\rho}(w) = \sum_{t=1}^T a_t h^{(t)}(w)$, we have

$$\begin{aligned}\sup_w |\hat{\rho}(w) - \rho(w)| &\le \sum_{t=1}^T a_t \sup_w |h^{(t)}(w) - \rho(w)| \\ &\le \left(\max_{t\in[T]} C_t'\right) n^{-\frac{\beta}{2(1+\beta)}} \sqrt{\log n}.\end{aligned}$$

This completes the proof.

□

Lemma 5 (Consistency of the null proportion estimator)**.** *Under Assumption 2, there exist positive constants $C_1, C_2 > 0$ such that, with probability at least $1 - \frac{1}{n_1}$,*

$$|\pi_0 - \hat{\pi}_0| \le C_1 \sqrt{\frac{\log n_0}{n_0}} + C_2 \sqrt{\frac{\log n_1}{n_1}}, \tag{44}$$

Proof. Recall that the empirical p-value estimator is defined as

$$\hat{p}_j = \frac{1 + \sum_{i\in[n_0]} \mathbb{I}(G(w_j) > G(w_i))}{1 + n_0}, \tag{45}$$

where G is a monotone function. Define $A_j := \{p_j > \lambda\}$ and $\hat{A}_j := \{\hat{p}_j > \lambda\}$. By the law of total probability:

$$\begin{aligned}\Pr(A_j) = \Pr(p_j > \lambda \mid y_j^{\mathrm{bd}} = 0) \cdot \pi_0 \\ + \Pr(p_j > \lambda \mid y_j^{\mathrm{bd}} = 1) \cdot (1 - \pi_0).\end{aligned} \tag{46}$$

For clean samples, $p_j \sim \mathrm{Uniform}[0, 1]$, which implies

$$\Pr(p_j > \lambda \mid y_j^{\mathrm{bd}} = 0) = 1 - \lambda. \tag{47}$$

By the assumption of the lemma, we have $\Pr(p_j > \lambda \mid y_j^{\mathrm{bd}} = 1) = 0$. Therefore,

$$\begin{aligned}\Pr(A_j) &= \pi_0(1 - \lambda) \\ &\quad + (1 - \pi_0) \cdot \Pr(p_j > \lambda \mid y_j^{\mathrm{bd}} = 1) \\ &= \pi_c \cdot (1 - \lambda).\end{aligned} \tag{48}$$

Then the error of the null proportion estimator can thus be written as

$$\begin{aligned}|\hat{\pi}_0 - \pi_0| &= \frac{1}{1-\lambda} \left| \Pr(A_j) - \frac{1}{n_1} \sum_{j\in[n_1]} \mathbb{I}(\hat{A}_j) \right| \\ &\le \frac{1}{1-\lambda} \Bigg(\underbrace{\left| \Pr(A_j) - \Pr(\hat{A}_j) \right|}_{(1)} \\ &\quad + \underbrace{\left| \Pr(\hat{A}_j) - \frac{1}{n_1} \sum_{j\in[n_1]} \mathbb{I}(\hat{A}_j) \right|}_{(2)} \Bigg)\end{aligned} \tag{49}$$

We now bound each term individually.

Term (1): By Lemma 2, for any $\epsilon > 0$,

$$\begin{aligned}&|\Pr(A_j) - \Pr(\hat{A}_j)| \\ &\quad \le \Pr(|\hat{p}_j - p_j| > \epsilon) + \Pr(|p_j - \lambda| \le \epsilon).\end{aligned} \tag{50}$$

For the first term on the right-hand side, conditioned on w_j, the $V_i = \mathbb{I}(G(w_j) > G(w_i))$ are i.i.d. Bernoulli with mean p_j. By Hoeffding's inequality,

$$\Pr(|\hat{p}_j - p_j| > \epsilon \mid w_j) \le 2\exp(-2n_0\epsilon^2).$$

Since the bound is uniform over w_j, it holds unconditionally:

$$\Pr(|\hat{p}_j - p_j| > \epsilon) \le 2\exp(-2n_0\epsilon^2).$$

For the second term, by Lemma 3, there exists a constant $C > 0$ such that

$$\Pr(|p_j - \lambda| \le \epsilon) \le C \cdot \epsilon.$$

Combining the two bounds, we have

$$|\Pr(A_j) - \Pr(\hat{A}_j)| \le 2\exp(-2n_0\epsilon^2) + C \cdot \epsilon.$$

Term (2): Define $U_j = \mathbb{I}(\hat{p}_j > \lambda)$. Conditioned on $\mathcal{D}^{\mathrm{cal}}$, the U_j are i.i.d. random variables. By applying Hoeffding's inequality:

$$\Pr\left(\left| \mathbb{E}[U_j] - \frac{1}{n_1} \sum_{j\in[n_1]} U_j \right| < \epsilon \mid \mathcal{D}^{\mathrm{cal}} \right) \ge 1 - 2\exp(-2n_1\epsilon^2). \tag{51}$$

Again, since the bound does not depend on $\mathcal{D}^{\mathrm{cal}}$, it holds unconditionally. Thus, with probability at least $1 - \delta$, the second term is bounded as

$$\textbf{(2)} \le \sqrt{\frac{\log\left(\frac{2}{\delta}\right)}{2n_1}} \tag{52}$$

Final Bound: Combining the two terms , we obtain, with probability at least $1-\delta$,

$$|\hat{\pi}_0-\pi_0| \le \frac{1}{1-\lambda}\left(2\exp(-2n_0\epsilon^2) + C\cdot\epsilon + \sqrt{\frac{\log\left(\frac{2}{\delta}\right)}{2n_1}}\right). \tag{53}$$

Choosing

$$\epsilon = \sqrt{\frac{\log n_0}{n_0}}, \quad \delta = \frac{1}{n_1} \tag{54}$$

yields

$$|\hat{\pi}_0-\pi_0| \le \frac{1}{1-\lambda}\left(2\cdot\frac{1}{{n_0}^2} + C\cdot\sqrt{\frac{\log n_0}{n_0}} + \sqrt{\frac{\log(2n_1)}{2n_1}}\right) \tag{55}$$

with probability at least $1-\frac{1}{n_1}$.

For sufficiently larger n_0 and n_1, the first term will be absorbed by the second term and the factor of $\sqrt{\log(2n_1)}$ can be bounded by a constant multiple of $\sqrt{\log n_1}$. Therefore, there exist constants $C', C'' > 0$ such that, with probability at least $1-\frac{1}{n_1}$,

$$\begin{aligned}|\hat{\pi}_0 - \pi_0| &\le \frac{1}{1-\lambda}\left(C'\cdot\sqrt{\frac{\log n_0}{n_0}} + C''\sqrt{\frac{\log n_1}{n_1}}\right)\\ &= C_1'\cdot\sqrt{\frac{\log n_0}{n_0}} + C_1''\cdot\sqrt{\frac{\log n_1}{n_1}},\end{aligned} \tag{56}$$

where $C_1' = \frac{C'}{1-\lambda}$ and $C_1'' = \frac{C''}{1-\lambda}$. □

Lemma 6 (Uniform convergence of $\hat{L}$)**.** *Suppose Assumption 1-2 hold, there exist constants $C_1, C_2 > 0$ such that, with probability at least $1-\frac{1}{n_0}-\frac{2}{n_1}$,*

$$\begin{aligned}&\sup_w \left|L(w) - \hat{L}(w)\right|\\ &\quad\le C_1 {n_0}^{-\frac{\beta_1}{2(1+\beta_1)}}\sqrt{\log n_0}\\ &\qquad + C_2\, {n_1}^{-\frac{\beta_2}{2(1+\beta_2)}}\sqrt{\log n_1}.\end{aligned} \tag{57}$$

Proof. Recall that

$$L(w) = \frac{\rho_c(w)\,\pi_0}{\rho_{\text{mix}}(w)}, \qquad \hat{L}(w) = \frac{\hat{\rho}_c(w)\,\hat{\pi}_0}{\hat{\rho}_{\text{mix}}(w)}.$$

Then

$$\begin{aligned}&\sup_w \left|L(w) - \hat{L}(w)\right|\\ &= \sup_w \left|\frac{\hat{\rho}_c(w)\,\hat{\pi}_0}{\hat{\rho}_{\text{mix}}(w)} - \frac{\rho_c(w)\,\pi_0}{\rho_{\text{mix}}(w)}\right|\\ &= \sup_w \frac{1}{\rho_{\text{mix}}(w)\hat{\rho}_{\text{mix}}(w)}\Big(\hat{\pi}_0\,\rho_{\text{mix}}(w)\left|\hat{\rho}_c(w) - \rho_c(w)\right|\\ &\qquad + \rho_{\text{mix}}(w)\rho_c(w)\left|\hat{\pi}_0 - \pi_0\right|\\ &\qquad + \pi_0\,\rho_c(w)\left|\rho_{\text{mix}}(w) - \hat{\rho}_{\text{mix}}(w)\right|\Big).\end{aligned} \tag{58}$$

By Assumption 1, there exists $l > 0$ such that $\inf_w \rho_{\text{mix}}(w) \ge l$. Moreover, Lemma 4 implies that, with probability at least $1-\frac{1}{n_1}$,

$$\sup_w \left|\hat{\rho}_{\text{mix}}(w) - \rho_{\text{mix}}(w)\right| \le C_{\rho_{\text{mix}}}\, {n_1}^{-\frac{\beta_2}{2(1+\beta_2)}}\sqrt{\log n_1}.$$

Hence, for sufficiently large n_1,

$$\inf_w \hat{\rho}_{\text{mix}}(w) \ge l - C_{\rho_{\text{mix}}}\, {n_1}^{-\frac{\beta_2}{2(1+\beta_2)}}\sqrt{\log n_1} \ge \frac{l}{2}.$$

Also, $\sup_w \rho_c(w) \le M$ for some $M > 0$. Combining these bounds yields

$$\begin{aligned}&\sup_w \left|L(w) - \hat{L}(w)\right|\\ &\quad\le \frac{2}{l}\Big(\sup_w \left|\hat{\rho}_c(w) - \rho_c(w)\right|\\ &\qquad + M\left|\hat{\pi}_0 - \pi_0\right| + \sup_w \left|\hat{\rho}_{\text{mix}}(w) - \rho_{\text{mix}}(w)\right|\Big).\end{aligned} \tag{59}$$

Applying Lemma 4 to both $\hat{\rho}_c$ and $\hat{\rho}_{\text{mix}}$, and Lemma 5 to $\hat{\pi}_0$, we obtain that, with probability at least $1-\frac{1}{n_0}-\frac{2}{n_1}$,

$$\begin{aligned}&\sup_w \left|L(w) - \hat{L}(w)\right|\\ &\le \frac{2}{l}\Bigg(C_{\rho_c}\, {n_0}^{-\frac{\beta_1}{2(1+\beta_1)}}\sqrt{\log n_0}\\ &\qquad + C_{\rho_{\text{mix}}}\, {n_1}^{-\frac{\beta_2}{2(1+\beta_2)}}\sqrt{\log n_1}\\ &\qquad + C_{\pi_0}\sqrt{\frac{\log n_0}{n_0}} + C'_{\pi_0}\sqrt{\frac{\log n_1}{n_1}}\Bigg).\end{aligned} \tag{60}$$

For sufficiently large n_0 and n_1, the terms $\sqrt{\frac{\log n_0}{n_0}}$ and $\sqrt{\frac{\log n_1}{n_1}}$ are dominated by ${n_0}^{-\frac{\beta_1}{2(1+\beta_1)}}\sqrt{\log n_0}$ and ${n_1}^{-\frac{\beta_2}{2(1+\beta_2)}}\sqrt{\log n_1}$, respectively. Absorbing constants into C_1 and C_2 completes the proof. □

A.2 Proof of Lemma 1

Proof. By definition, the FAR is given by:

$$\text{FAR} = \mathbb{E}\left[\frac{\sum_{j\in[n_1]}(1-\hat{\delta}_j)\delta_j}{1 \vee \sum_{j\in[n_1]}(1-\hat{\delta}_j)}\right].$$

Using the law of total expectation and the independence of δ_j, we can write:

$$\begin{aligned}\text{FAR} &= \mathbb{E}\left[\mathbb{E}\left[\frac{\sum_{j\in[n_1]}(1-\hat{\delta}_j)\delta_j}{1 \vee \sum_{j\in[n_1]}(1-\hat{\delta}_j)}\,\middle|\,\{w_j\}_{j\in[n_1]}\right]\right]\\ &= \mathbb{E}\left[\frac{\sum_{j\in[n_1]}(1-\hat{\delta}_j)\mathbb{E}[\delta_j \mid w_j]}{1 \vee \sum_{j\in[n_1]}(1-\hat{\delta}_j)}\right].\end{aligned} \tag{61}$$

Since $L_j = \Pr(\delta_j = 0|w_j) = 1 - \mathbb{E}[\delta_j \mid w_j]$. Then,

$$\text{FAR} = \mathbb{E}\left[\frac{\sum_{j\in[n_1]}(1-\hat{\delta}_j)(1-L_j)}{1 \vee \sum_{j\in[n_1]}(1-\hat{\delta}_j)}\right].$$

□

A.3 Proof of Theorem 1

Proof. Let $\hat{L}_j$ denote the estimated value of L_j used in the decision rule. We decompose the FAR as:

$$\begin{aligned}\text{FAR} &= \mathbb{E}\left[\frac{\sum_{j\in[n_1]}(1-\hat{\delta}_j)(1-\hat{L}_j)}{1 \vee \sum_{j\in[n_1]}(1-\hat{\delta}_j)}\right] \\ &\quad + \mathbb{E}\left[\frac{\sum_{j\in[n_1]}(1-\hat{\delta}_j)(\hat{L}_j - L_j)}{1 \vee \sum_{j\in[n_1]}(1-\hat{\delta}_j)}\right] \\ &= \alpha + \mathbb{E}\left[\frac{\sum_{j\in[n_1]}(1-\hat{\delta}_j)(\hat{L}_j - L_j)}{1 \vee \sum_{j\in[n_1]}(1-\hat{\delta}_j)}\right],\end{aligned} \tag{62}$$

where the first term equals α by construction of the Algorithm 1.

To bound the second term, note that according to lemma 6 for any j,

$$\begin{aligned}|L_j - \hat{L}_j| \le C_1 {n_0}^{-\frac{\beta_1}{2(1+\beta_1)}}\sqrt{\log n_0} \\ + C_2 {n_1}^{-\frac{\beta_2}{2(1+\beta_2)}}\sqrt{\log n_1}\end{aligned} \tag{63}$$

with probability at least $1 - \frac{1}{n_0} - \frac{2}{n_0}$. Let event $\mathcal{E}$ denote the event that this bound holds. Then, we can write:

$$\begin{aligned}&\mathbb{E}\left[\left|\frac{\sum_j(1-\hat{\delta}_j)(\hat{L}_j - L_j)}{1 \vee \sum_j(1-\hat{\delta}_j)}\right|\right] \\ &\le \mathbb{E}\left[\frac{\sum_j(1-\hat{\delta}_j)|L_j - \hat{L}_j|}{1 \vee \sum_j(1-\hat{\delta}_j)}\right] \\ &\le \mathbb{E}\left[\frac{\sum_j(1-\hat{\delta}_j)|L_j - \hat{L}_j|}{1 \vee \sum_j(1-\hat{\delta}_j)}\middle|\mathcal{E}\right]\cdot\Pr(\mathcal{E}) \\ &\quad + \mathbb{E}\left[\frac{\sum_j(1-\hat{\delta}_j)|L_j - \hat{L}_j|}{1 \vee \sum_j(1-\hat{\delta}_j)}\middle|\mathcal{E}^c\right]\cdot\Pr(\mathcal{E}^c) \\ &\le \mathbb{E}\left[\frac{\sum_j(1-\hat{\delta}_j)|L_j - \hat{L}_j|}{1 \vee \sum_j(1-\hat{\delta}_j)}\middle|\mathcal{E}\right] + \Pr(\mathcal{E}^c) \\ &\le C_1 {n_0}^{-\frac{\beta_1}{2(1+\beta_1)}}\sqrt{\log n_0} + C_2 {n_1}^{-\frac{\beta_2}{2(1+\beta_2)}}\sqrt{\log n_1} \\ &\quad + \frac{1}{n_0} + \frac{2}{n_1}\end{aligned} \tag{64}$$

Dataset	Classes	Train Size	Val Size	Test Size
Yelp	2	14,000	3,000	3,000
HSOL	2	5823	2,485	2,485
AG News	4	108,000	7,600	7,600

Table 3: Summary of datasets used in experiments.

For sufficiently large n_0 and n_1, the terms $\frac{1}{n_0}$ and $\frac{2}{n_1}$ can be absorbed into the first two terms respectively.

□

B Details of Scoring Function

In this section, we provide detailed formulations for the two scoring functions used in our experiments: Mahalanobis Distance (MDS) and BadActs Score.

Both scoring functions rely on an additional clean data set for their calculations. Specifically, we divide the original calibration set into two equal parts. The first half is used to estimate the mean and covariance matrix for MDS and to compute the clean activation range for BadActs. The second half serves as the true calibration set, as detailed in Section 4.

- **Mahalanobis Distance:** We estimate the mean (μ) and covariance (Σ) of the feature representations of clean samples using the first half of the calibration set. The Mahalanobis distance for each test sample is then computed as:

$$W_{\text{Mahalanobis}}(x_j) = \sqrt{(f(x_j)-\mu)^\top \Sigma^{-1}(f(x_j)-\mu)}, \tag{65}$$

  where $f(x_j)$ represents the feature vector of the test sample x_j, and μ and Σ are derived from the clean samples. The Mahalanobis distance quantifies how far a test sample's feature representation deviates from the distribution of clean samples.

  For this score, we set $G(w) = -w$ so that smaller scores indicate a higher likelihood of being a backdoor sample.

- **BadActs Score:** We follow the scoring methodology proposed in BadActs (Yi et al., 2024). For each test sample x_j, the intermediate score for the k-th activation value is defined as:

$$\Phi^k(x_j) = \mathbb{1}_{[\mu^k - a\sigma^k, \mu^k + a\sigma^k]}(r_{j,k}),$$

  where $\mathbb{1}_{[c,d]}(r)$ is an indicator function that evaluates to 1 if r lies within the interval $[c, d]$, and 0 otherwise. Here, μ^k and σ^k are the mean and standard deviation of the k-th activation value

	Yelp								AGNews								HSOL							
	BadNets		AddSent		StyleBkd		SynBkd		BadNets		AddSent		StyleBkd		SynBkd		BadNets		AddSent		StyleBkd		SynBkd	
	FAR	Power	FAR	Power	FAR	Power	FAR	Power	FAR	Power	FAR	Power	FAR	Power	FAR	Power	FAR	Power	FAR	Power	FAR	Power	FAR	Power
BKI	0.158	1.000	0.114	1.000	0.435	0.324	0.083	1.000	0.157	1.000	0.107	0.999	0.154	0.835	0.151	0.989	0.160	1.000	0.206	0.888	0.215	0.869	0.213	0.881
CUBE	**0.001**	**0.991**	0.334	0.498	**0.013**	**0.866**	0.321	0.522	0.200	0.999	0.250	0.749	0.203	0.945	0.200	0.999	**0.000**	**0.934**	**0.000**	**0.890**	0.255	0.618	0.200	0.998
STRIP	0.197	0.925	0.200	0.930	0.194	0.899	0.196	0.897	0.197	0.877	0.198	0.925	0.195	0.882	0.197	0.857	0.196	0.927	0.196	0.899	0.201	0.939	0.199	0.958
RAP	0.209	0.934	0.212	0.927	0.210	0.929	0.211	0.932	0.215	0.907	0.210	0.940	0.240	0.786	0.222	0.873	0.208	0.911	0.235	0.815	0.215	0.867	0.207	0.956
SCM-md	**0.000**	**0.982**	**0.000**	**0.986**	**0.000**	**0.995**	**0.000**	**0.991**	0.201	0.994	0.200	0.996	0.131	0.995	**0.007**	**0.994**	0.200	0.996	0.200	0.996	0.197	0.994	0.119	0.996
SCM-badacts	0.189	0.999	0.185	0.999	0.002	0.999	0.175	0.998	0.188	0.995	0.193	0.995	0.172	0.995	0.043	0.992	0.191	0.994	0.063	0.995	0.172	0.994	0.201	0.995
SAFER-md	**0.002**	**0.979**	**0.000**	**0.928**	**0.001**	**0.948**	**0.000**	**0.842**	**0.077**	**0.950**	**0.095**	**0.983**	0.108	0.988	**0.011**	**0.906**	**0.033**	**0.813**	**0.048**	**0.923**	0.108	0.426	**0.078**	**0.986**
SAFER-badacts	**0.039**	**0.997**	**0.005**	**0.996**	**0.025**	**1.000**	**0.094**	**0.999**	**0.078**	**0.983**	**0.084**	**0.930**	**0.094**	**0.954**	**0.074**	**0.995**	**0.099**	**0.574**	**0.071**	**0.811**	0.114	0.698	**0.081**	**0.972**

Table 4: FAR control results for the Pythia-1B model.

		BERT-base-uncased				RoBERTa-Base				Pythia-1B			
		Badnets	Addsent	Stylebkd	Synbkd	Badnets	Addsent	Stylebkd	Synbkd	Badnets	Addsent	Stylebkd	Synbkd
Yelp	md	0.000±0.000	0.000±0.000	0.000±0.000	0.000±0.000	0.000±0.000	0.000±0.000	0.000±0.000	0.000±0.000	0.000±0.000	0.000±0.000	0.000±0.000	0.000±0.000
	badacts	0.000±0.000	0.000±0.000	0.004±0.002	0.000±0.000	0.000±0.000	0.000±0.000	0.000±0.000	0.000±0.000	0.000±0.000	0.000±0.000	0.000±0.000	0.000±0.000
AGNews	md	0.000±0.000	0.000±0.000	0.001±0.000	0.000±0.000	0.000±0.000	0.000±0.000	0.002±0.001	0.000±0.000	0.000±0.000	0.000±0.000	0.000±0.000	0.000±0.000
	badacts	0.000±0.000	0.000±0.000	0.002±0.001	0.000±0.000	0.000±0.000	0.000±0.000	0.006±0.002	0.001±0.002	0.000±0.000	0.000±0.000	0.007±0.002	0.000±0.000
HSOL	md	0.000±0.000	0.000±0.000	0.002±0.002	0.000±0.000	0.000±0.000	0.000±0.000	0.000±0.000	0.000±0.000	0.000±0.000	0.000±0.000	0.001±0.001	0.000±0.000
	badacts	0.000±0.000	0.000±0.000	0.007±0.003	0.000±0.000	0.000±0.000	0.000±0.000	0.007±0.004	0.000±0.000	0.000±0.000	0.000±0.000	0.064±0.009	0.000±0.000

Table 5: Proportion of backdoor samples whose scores exceed the λ-quantile of clean samples under different poisoning ratios.

estimated from clean samples, a is a hyperparameter (set to 3 in our experiments), and $r_{j,k}$ is the k-th activation value of x_j.

The final BadActs score is then computed as:

$$W_{\text{Badacts}}(x_j) = \frac{1}{L \cdot d} \sum_{k=1}^{L \cdot d} \Phi^k(x_j),$$

where $L \cdot d$ is the total number of activation values across all layers. This score reflects the proportion of activation values in x_j that fall within the expected range based on clean samples. For this score, we set $G(w) = w$.

C Dataset Details

We provide a summary of the datasets used in our experiments in Table 3. We follow the same data partition steps as in (Cui et al., 2022).

For the SynBkd attack on the AG News dataset, due to the large dataset size, poisoning the entire dataset is computationally expensive (over 100 hours). Therefore, we randomly sample 10% of the dataset (approximately 10,800 samples) for backdoor training.

Use of Potentially Offensive Data Some of the datasets used in our experiments may contain offensive, toxic, or otherwise sensitive content (e.g., HSOL). These datasets are publicly available and have been widely used in prior work on backdoor attacks and content moderation for language models. We use them solely for research purposes, with the goal of evaluating and improving the robustness and safety of LLMs against malicious behaviors. We do not introduce new offensive content beyond what is already present in the original datasets, nor do we deploy models trained on these data in real-world applications. We acknowledge the potential risks associated with handling such data and follow standard research practices to minimize exposure and misuse.

D Additional Experimental Results

Experimental Results on Larger Models We further conduct experiments on the Pythia-1B model (Biderman et al., 2023) to evaluate the effectiveness of our method on larger-scale models. The results are reported in Table 4. We observe a trend consistent with that in the main manuscript: our method reliably achieves the desired FAR control while maintaining strong detection power across most attack settings.

Validity of Assumption 2 In Assumption 2, we posit that the scores of clean and backdoor samples are separable; specifically, the score of any backdoor sample does not exceed the λ-quantile of the clean sample scores. To empirically validate this assumption, we compute the proportion of backdoor samples whose scores exceed the λ-quantile of clean samples under different poisoning ratios. The results are presented in Table 5. As shown, the violation proportion is extremely low (often zero) across different datasets, attack methods, and scoring functions, supporting the practical validity of this assumption.

	$\mu = 60$			$\mu = 70$			$\mu = 80$			$\mu = 90$		
	FAR	Power	Violation	FAR	Power	Violation	FAR	Power	Violation	FAR	Power	Violation
$\sigma = 10$	0.047±0.039	0.869±0.226	0.000±0.000	0.054±0.044	0.836±0.234	0.000±0.000	0.074±0.048	0.732±0.277	0.001±0.000	0.116±0.050	0.584±0.319	0.011±0.003
$\sigma = 20$	0.074±0.048	0.732±0.277	0.001±0.000	0.090±0.052	0.648±0.295	0.003±0.001	0.116±0.050	0.585±0.319	0.011±0.003	0.166±0.029	0.614±0.341	0.037±0.007

Table 6: Simulation results of violation proportion

	Mahalanobis Distance Score						Badacts Score					
	n_0=500		n_0=1000		n_0=2000		n_0=500		n_0=1000		n_0=2000	
	FAR	Power	FAR	Power	FAR	Power	FAR	Power	FAR	Power	FAR	Power
n_1=500	0.102±0.058	0.814±0.126	0.102±0.043	0.839±0.066	0.095±0.031	0.840±0.047	0.107±0.009	0.991±0.005	0.108±0.006	0.989±0.003	0.108±0.005	0.988±0.002
n_1=1000	0.114±0.066	0.826±0.125	0.118±0.045	0.849±0.082	0.110±0.039	0.845±0.065	0.067±0.034	0.979±0.016	0.065±0.028	0.983±0.006	0.069±0.018	0.985±0.004
n_1=2000	0.069±0.046	0.748±0.133	0.062±0.042	0.758±0.086	0.060±0.030	0.768±0.079	0.083±0.030	0.991±0.004	0.085±0.019	0.989±0.004	0.092±0.014	0.989±0.003

Table 7: Impact of calibration and test set sizes

To further examine the impact of potential violations on FAR control, we simulate clean and backdoor scores using Gaussian distributions with varying means and variances, and evaluate FAR and detection power under different poisoning ratios. Specifically, clean sample scores are drawn from $\mathcal{N}(100, \sigma^2)$, while backdoor sample scores are drawn from $\mathcal{N}(\mu, \sigma^2)$, where μ and σ^2 are chosen to represent different degrees of separation between the two distributions. We then apply our detection algorithm to the simulated data and compute FAR and power for each configuration.

The results are summarized in Table 6. We observe that as long as the distributions are sufficiently separated (e.g., $\mu \leq 70$ when $\sigma = 20$, and $\mu \leq 80$ when $\sigma = 10$), the violation proportion remains low, and our method continues to achieve reliable FAR control.

Varying Calibration and Test Set Sizes In this experiment, we evaluate the impact of different calibration and test set sizes on the FAR control of our method. Specifically, we vary both the calibration and test set sizes from 500 to 2000, and measure the resulting FAR and detection power on the AGNews dataset. The results are reported in Table 7.

From these results, we observe that our method consistently maintains effective FAR control even with relatively small sample sizes (e.g., 500 calibration samples and 500 test samples). This demonstrates the practicality and robustness of our approach in real-world scenarios, where the number of available samples may be limited.