

CRICO: A Certified Robustness Framework via Input-Dependent Covariance Optimization

Shudi Guo¹, Bin Liu¹, Shihai Wang^{1*}, Tengfei Shi¹, Yu Liu², Haoran Li¹

¹School of Reliability and Systems Engineering, Beihang University, Beijing, China

²Institute of Satellite Information Engineering, Beijing, China

{shudiguo, liubin, wangshihai, shitfing, liuyu415, li_haoran}@buaa.edu.cn

Abstract

Randomized smoothing provides certified robustness guarantees, but its direct application to text remains challenging due to the discrete, sparse, and semantically structured nature of textual inputs. We propose CRICO, an input-dependent covariance optimization framework for certified robustness in the text embedding space. CRICO estimates token-level decision influence in a black-box manner, making it readily extensible to LLMs. By adapting the certified ellipsoid to local decision structure, CRICO significantly enhances certified robustness. Experiments on multiple tasks demonstrate its effectiveness and broad applicability.

1 Introduction

With the rapid development of Large Language Models (LLMs), natural language processing (NLP) has been expanding to broader application scenarios. Despite their remarkable progress in language understanding and generation, current models remain sensitive to subtle perturbations in input texts (Wang et al., 2019; Dong et al., 2025; Abdali et al., 2024). For conventional NLP models, such perturbations are typically studied as adversarial attacks (Szegedy et al., 2013; Miyato et al., 2016), where attackers induce incorrect predictions through synonym substitution (Iyyer et al., 2018; Alzantot et al., 2018; Zang et al., 2019), character-level modification (Gao et al., 2018; Belinkov and Bisk, 2017; Eger et al., 2019), or semantics-preserving transformations (Jia and Liang, 2017). For LLMs, similar vulnerabilities are further manifested in prompt injection attacks (Liu et al., 2024), jailbreak attacks (Yi et al., 2024), and hallucinations triggered (Huang et al., 2025) by unstable inputs.

To mitigate model sensitivity to input perturbations, various defense mechanisms have been proposed, including adversarial training (Madry et al.,

2017; Zhu et al., 2019) and certified robustness methods (Wong and Kolter, 2018; Salman et al., 2019; Gowal et al., 2019). Among them, randomized smoothing (Lecuyer et al., 2019; Cohen et al., 2019) has been widely studied in the image domain as a theoretically guaranteed certification framework. Its core idea is to inject random noise around the input and construct a smoothed classifier that provides a provable robustness radius for model predictions. However, directly applying randomized smoothing based on high-dimensional noise injection to text tasks remains challenging. Unlike continuous and dense image data, in the word embedding space, different tokens and semantic directions may have different impacts on model decisions (Wang et al., 2020).

To address these challenges, we propose CRICO, an input-dependent covariance optimization method for randomized smoothing in the text embedding space.

The main contributions of this paper are:

1. We propose CRICO, an input-dependent covariance optimization framework that constructs token-wise anisotropic Gaussian perturbation distributions in the text embedding space, thereby improving the certified robustness of randomized smoothing for text tasks.
2. We formulate a constrained covariance optimization objective that enlarges the certified region while guiding the region to form an effective geometric shape adapted to the local decision boundary, avoiding the inefficient allocation of isotropic noise in the text embedding space.
3. We systematically evaluate CRICO on traditional text classification tasks and security-related task scenarios, covering both traditional NLP models and LLMs. Experimental results show that the proposed method can provide stable certified robustness guarantees across different model types and task settings.

*Corresponding author.

2 Related Work and Motivation

Randomized smoothing has become an important framework for certified robustness due to its theoretical guarantees and weak dependence on specific model architectures (Lecuyer et al., 2019; Cohen et al., 2019). For text tasks, randomized smoothing has been introduced into discrete perturbation settings, such as synonym substitution (Ye et al., 2020; Zeng et al., 2023). These methods provide certified robustness against discrete word-level perturbations, but do not explore the embedding space and fully capture the non-uniform influence of different tokens and semantic directions on model decisions (Zhang et al., 2024).

In fact, regarding the enhancement of empirical adversarial robustness, substantial work has already been conducted on injecting perturbations within the text embedding space (Liang et al., 2017; Samanta and Mehta, 2017; Jia et al., 2019; Li et al., 2018), yielding promising results. These results indicate that embedding-space methods offer a promising way to study model robustness in text tasks. Since embedding space can cover a richer set of local perturbation directions, including virtual variations that cannot be explicitly represented by a single synonym substitution.

However, directly injecting isotropic Gaussian noise into the embedding space struggles to capture the anisotropic characteristics of the model’s decision boundaries (Miyato et al., 2016). Therefore, the isotropic Gaussian smoothing proposed by Cohen et al. (2019) is not suitable for direct application to the text domain. In fact, to overcome the limitations of isotropic smoothing, several studies have explored certified robustness under anisotropic Gaussian noise. Zhang et al. (2024) added anisotropic noise to the model training process. Hong and Hong (2022) directly extend the theory of Cohen et al. (2019), generalizing their theory to anisotropic Gaussians. However, Hong and Hong (2022) use anisotropic Gaussian noise with a non-zero mean, which is often considered to cause a shift in the model’s decision boundary. In contrast, Eiras et al. (2022) proved the randomized smoothing theory using the Lipschitz constant, and they employed anisotropic Gaussian noise with a zero mean. This paper builds upon the certifiable robustness theory proposed by Eiras et al. (2022), however, the problems addressed, the optimization granularities, and the objective functions differ.

Randomized smoothing has also been explored

for improving LLM robustness. For example, Lou et al. (2024) apply randomized ablation to LLM input texts to improve robustness; Sun et al. (2024) extend SAFER (Ye et al., 2020) to defend against backdoor attacks in LLMs; and Robey et al. (2023) use input randomization to mitigate jailbreak attacks. These methods mainly rely on discrete input perturbations, randomized ablation, or input randomization, while less attention has been paid to constructing input-dependent continuous perturbation distributions in the text embedding space.

3 Randomized Smoothing Certification

3.1 Preliminaries

In text tasks, we consider perturbations introduced by randomized smoothing mechanism in the embedding space rather than directly in the discrete input space. Let $v(\mathbf{x}) \in \mathbb{R}^{n \times d}$ denote the embedding representation of an input text, where n is the sequence length and d is the embedding dimension. Given a noise distribution $\epsilon \sim \mathcal{D}$, the smoothed classifier predicts the class with the largest probability under noisy inputs:

$$g(\mathbf{x}) = \arg \max_{y \in \mathcal{Y}} \mathbb{P}_{\epsilon \sim \mathcal{D}}(f_{\theta}(v(\mathbf{x}) + \epsilon) = y), \quad (1)$$

For a predicted class y_A , randomized smoothing estimates a lower confidence bound $\underline{p}_A$ on the probability of class y_A , and an upper confidence bound $\overline{p}_B$ on the probability of any competing class $y \neq y_A$. If $\underline{p}_A > \overline{p}_B$, the prediction of the smoothed classifier can be certified within a certain perturbation region around $\mathbf{x}$.

3.2 Ellipsoidal Certification under Gaussian Smoothing

Following the ellipsoidal certification framework of Eiras et al. (2022), we consider Gaussian smoothing with a general covariance matrix: $\epsilon \sim \mathcal{N}(\mathbf{0}, \Sigma)$, where $\Sigma \succ 0$ defines the anisotropic perturbation geometry. Under this distribution, the certified region is no longer restricted to an ℓ_2 ball, but becomes an ellipsoid characterized by the Mahalanobis norm: $\|\delta\|_{\Sigma} = \sqrt{\delta^{\top} \Sigma^{-1} \delta}$.

Theorem 1 (Eiras et al., 2022). *Let f be a base classifier and let g be the smoothed classifier constructed with Gaussian noise $\epsilon \sim \mathcal{N}(\mathbf{0}, \Sigma)$, where $\Sigma \succ 0$. For class y_A , suppose we have a lower bound $\underline{p}_A$ and an upper bound $\overline{p}_B$. If the*

following inequality holds:

$$\begin{aligned} \mathbb{P}(f_\theta(v(\mathbf{x}) + \epsilon) = y_A) &\geq \underline{p}_A > \overline{p}_B \\ &\geq \max_{y \neq y_A} \mathbb{P}(f_\theta(v(\mathbf{x}) + \epsilon) = y) \end{aligned} \quad (2)$$

Then, for any adversarial perturbation δ , with Mahalanobis distance $\|\delta\|_\Sigma = \sqrt{\delta^T \Sigma^{-1} \delta}$, the certified robustness holds in the region of an ellipsoid constructed as:

$$\mathcal{E}_\mathbf{x} = \{\delta : \|\delta\|_\Sigma \leq R\}, \quad (3)$$

in which the Mahalanobis radius R is:

$$R = \frac{1}{2}(\Phi^{-1}(\underline{p}_A) - \Phi^{-1}(\overline{p}_B)), \quad (4)$$

and Φ^{-1} denotes the inverse CDF of the standard Gaussian distribution.

When $\Sigma = \sigma^2 I$, i.e., when all elements have the same standard deviation, $\|\delta\|_\Sigma = \frac{1}{\sigma} \|\delta\|_2$, Eq. (4) then reduces to the theory of Cohen et al. (2019).

In this paper, we build upon this ellipsoidal certification framework and optimize the covariance matrix Σ in an input-dependent manner.

4 CRICO: Certified Robustness Framework via Input-Dependent Covariance Optimization

4.1 CRICO Framework Overview

Motivated by the theoretical result above, we propose the framework CRICO (Certified Robustness Framework via Input-Dependent Covariance Optimization). Our framework constructs a smoothed classifier based on anisotropic Gaussian noise and generates Mahalanobis ellipsoidal regions induced by the covariance matrix Σ . The core goal of our framework is to simultaneously (i) maximize the coverage of the certified region and (ii) at the same time adapt its geometry to the local decision structure of each input, thereby improving certification coverage through more effective perturbation-budget allocation, rather than spending excessive budget on directions where expansion rapidly approaches the local decision boundary.

Concretely, by injecting Gaussian noise $\epsilon \sim \mathcal{N}(0, \Sigma)$, the covariance matrix Σ determines the geometry of the certified region: a Mahalanobis ellipsoid characterized by a Mahalanobis-distance constraint. Constructing a Σ is crucial: an appropriate Σ should not only enlarge the ellipsoid volume to improve certification coverage, but also ensure

that the ellipsoid expands along directions requiring the minimum perturbation budget.

As illustrated by the two subfigures between Step 4 and Step 5 in Fig.1, where the curve denotes the decision boundary, we prefer the certified region to be a broader ellipsoid rather than an ℓ_2 -norm ball. Moreover, for two ellipsoids with the same volume, the certified region should avoid excessive expansion along the decision-sensitive direction, i.e., the normal direction of the decision boundary, since such expansion increases the risk that perturbed samples cross the local decision boundary, thereby reducing prediction stability.

To achieve these goals, our framework follows the design principles below:

Coverage. Under the certification constraints, the certified ellipsoidal region is enlarged as much as possible by maximizing the volume metric, thereby increasing the proportion of certifiable samples and the overall robust coverage.

Alignment. Decision-critical directions near the decision boundary are characterized by the sensitivity of each token in the input text to the model’s decision. The anisotropic structure of Σ is then adjusted accordingly, so that the geometry of the certified region is consistent with decision boundary and adversarial threat directions, avoiding ineffective expansions where the region is "large but threat-irrelevant" or where "unconstrained volume expansion approaches the decision boundary".

Validity. The construction of Σ is kept compatible with Theorem 1, i.e., the way Σ is obtained does not affect the certifiability of the randomized smoothing model.

Simplicity and Efficiency. Since multiple Monte Carlo samples are required for each input text in randomized smoothing, the framework should be easy to implement and scalable.

Although CRICO is inspired by ANCER (Eiras et al., 2022), the two approaches differ substantially in both optimization granularity and objective. ANCER adopts an element-wise covariance optimization strategy in continuous representation spaces to enlarge the certified ellipsoid primarily. CRICO instead introduces token-wise block-diagonal covariance optimization and uses token importance to guide the perturbation budget across different token subspaces. Through the *Coverage* and *Alignment* principles, CRICO not only enlarges the certified region, but also adapts its geometry to the token-level decision-sensitive structure of each input.

An overview of the proposed CRICO framework

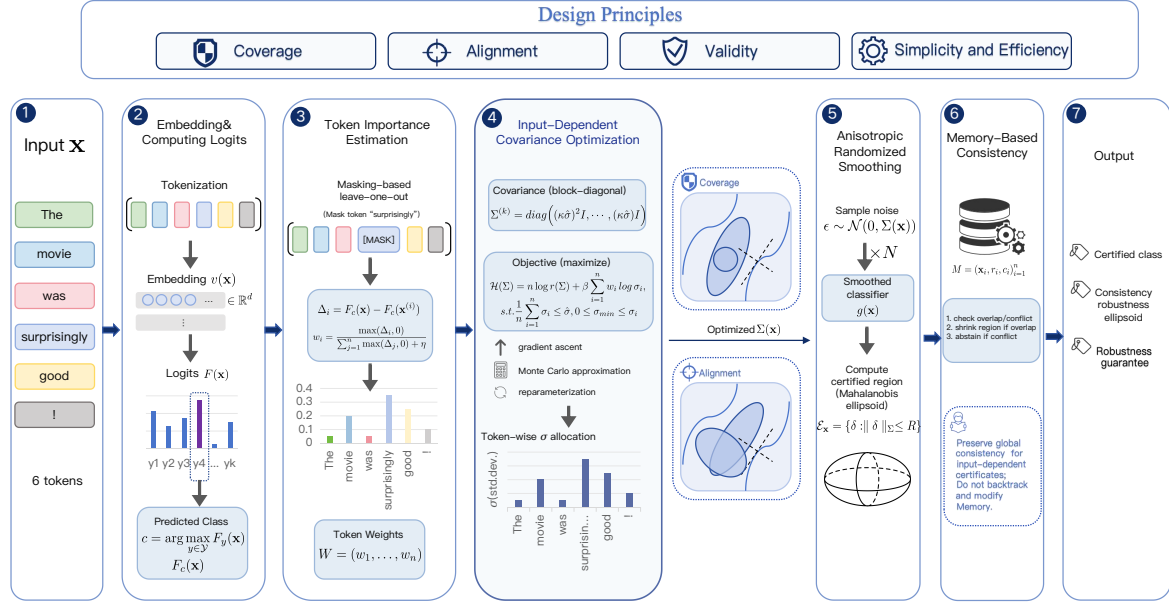


Figure 1: Overall pipeline of CRICO

is illustrated in Fig. 1.

4.2 Surrogate Objective for Certified Ellipsoid Volume

Unlike isotropic smoothing with a fixed standard deviation σ , we aim to learn a data-dependent covariance matrix Σ to align with the local decision boundary. Since Σ induces an ellipsoidal certified region in the embedding space, we first align the notation for the ellipsoid and its volume.

Definition 1 (Certified Ellipsoid). Given an input $\mathbf{x}$ and a perturbation vector $\delta \in \mathbb{R}^{nd}$, the certified region induced by Σ can be written as:

$$\mathcal{E}_{\mathbf{x}} = \left\{ \delta \in \mathbb{R}^{nd} : \|\delta\|_{\Sigma} \leq r \right\}, \quad (5)$$

where $\|\delta\|_{\Sigma} = \sqrt{\delta^T \Sigma^{-1} \delta}$ and $\Sigma = \text{diag}(\sigma_1^2 I_d, \sigma_2^2 I_d, \dots, \sigma_n^2 I_d)$. The volume of $\mathcal{E}_{\mathbf{x}}$ is:

$$\mathcal{V}(\mathcal{E}_{\mathbf{x}}) = \frac{\pi^{nd/2}}{\Gamma(nd/2 + 1)} r^{nd} \prod_{i=1}^n \sigma_i^d. \quad (6)$$

Surrogate objective. Eq. (6) contains a constant factor, then the expression can be regarded as:

$$\mathcal{V}(\mathcal{E}_{\mathbf{x}}) \propto r^{nd} \cdot \prod_{i=1}^n \sigma_i^d. \quad (7)$$

r here is a function of Σ , i.e. $r(\Sigma)$ (Eq. (3,4)). Our goal is to find the parameter Σ that maximizes the right-hand side of Eq. (7), equivalently:

$\max \left(d(n \log r(\Sigma) + \sum_{i=1}^n \log \sigma_i) \right)$. Since d is fixed, we have the proxy formula below to maximize:

$$n \log r(\Sigma) + \sum_{i=1}^n \log \sigma_i. \quad (8)$$

4.3 Token Importance for Decision-Sensitive Alignment

To explicitly model decision-relevant tokens in covariance optimization and thereby effectively constrain the expansion direction of the certified ellipsoid, we adopt a masking-based leave-one-out importance estimator to assign tokens with different levels of influence distinct weights: we mask each token and measure the logit change of the target class to quantify the token’s contribution to the current decision. As a simple, efficient, and black-box importance estimator directly related to the covariance optimization objective, the leave-one-out method aligns with the design principles of the CRICO framework: *Simplicity and Efficiency*.

Specifically, we define a tuple $W = (w_1, \dots, w_n)$ to store token importances and initialize $w_i = 0$. Let the base classifier output a logit vector $F(\mathbf{x}) \in \mathbb{R}^{|\mathcal{Y}|}$. We define the original predicted class as $c = \arg \max_{y \in \mathcal{Y}} F_y(\mathbf{x})$.

We then apply a masking operation to the i -th token to obtain $\mathbf{x}^{(i)}$, and compute the logit drop of

the target class c as a proxy of token importance:

$$\Delta_i = F_c(\mathbf{x}) - F_c(\mathbf{x}^{(i)}). \quad (9)$$

Finally, we normalize the importances into non-negative weights:

$$w_i = \frac{\max(\Delta_i, 0)}{\sum_{j=1}^n \max(\Delta_j, 0) + \eta}. \quad (10)$$

Here $\eta > 0$ is a numerical stabilizer. The whole process is shown in Algorithm 1 in Appendix A.1. And the pipeline is illustrated in Steps 2 and 3 of Fig.1.

4.4 Constrained Covariance Optimization

Based on the above analysis, we further propose an optimization objective for Σ . Our goal is to expand the certifiable region (*Coverage*) while ensuring it better conforms to the local decision boundaries of the model (*Alignment*).

For the certified ellipsoid induced by anisotropic Gaussian perturbations, its volume is jointly determined by the certified radius and the scaling factors σ_i over different token sub-embedding spaces. Ignoring constants irrelevant to optimization, the proxy formula of the certified ellipsoid is given in Eq. (8). Since the second term in the formula reflects the joint expansion contribution across token subspaces, this motivates us to optimize σ_i at the token level, so that the certified region can preferentially expand along directions consistent with the local decision structure. We further introduce token-wise importance weights w_i (Eq. (10)) and transform the second term as an importance-weighted anisotropic allocation term for token-wise expansion scale allocation: $\sum_{i=1}^n w_i \log \sigma_i$. In this way, the certified region is encouraged to maintain a large overall scale while better fitting the local decision boundary.

Based on this, we define the final optimization objective as follows:

$$\begin{aligned} \mathcal{H}(\Sigma) &= n \log r(\Sigma) + \beta \sum_{i=1}^n w_i \log \sigma_i, \\ \text{s.t. } \frac{1}{n} \sum_{i=1}^n \sigma_i &\leq \hat{\sigma}, 0 < \sigma_{\min} \leq \sigma_i \end{aligned} \quad (11)$$

We constrain the average noise level by $\frac{1}{n} \sum_{i=1}^n \sigma_i \leq \hat{\sigma}$, where $\hat{\sigma}$ is a global noise budget, which keeps the overall perturbation budget under control and prevents trivially enlarging the covariance scales at the expense of

clean accuracy. $0 < \sigma_{\min} \leq \sigma_i$ for numerical stability.

The objective in Eq. (11) is derived from a log-volume proxy of the certified ellipsoid. By optimizing Σ with gradient ascent, CRICO enlarges the anisotropic certified region while maintaining local prediction stability. The first term, $n \log r(\Sigma)$, is certification-driven: if the noise scale grows along decision-sensitive directions, noisy samples are more likely to cross the local decision boundary, reducing $\underline{p}_A$ and hence $r(\Sigma)$. Thus, this term discourages unsafe covariance expansion.

The second term is an importance-weighted regularizer that shapes the ellipsoid. It prevents the optimization from relying only on the radius term and becoming overly conservative. Here, β balances certified-region enlargement and sensitivity-aware covariance adaptation.

Specifically, to obtain the final value of Σ , we optimize a Monte Carlo approximation $\hat{r}(\Sigma)$ of $r(\Sigma)$ (Cohen et al., 2019). For the input $\mathbf{x}$, we sample num times, then calculate Eq. (11) and iterate. See Algorithm 2 in Appendix A.2 for the detailed algorithm.

4.5 Memory-Based Certification Consistency

For a single input $\mathbf{x}$, once the covariance matrix Σ of the anisotropic Gaussian noise is fixed, the certification procedure for randomized smoothing coincides with the isotropic case in Cohen et al. (2019). The key obstacle to certifying anisotropic Gaussian smoothing arises from the input-dependent nature of Σ , i.e., the data-dependent covariance that can be written as $\Sigma(\mathbf{x})$.

When certifying data-dependent classifiers, ensuring the global consistency of certification is a widely recognized challenge. Prior works Alfarrar et al. (2022); Eiras et al. (2022) addresses this issue by introducing a memory-based consistency checking procedure. We adopt the memory-based cstrategy of ANCER (Eiras et al., 2022) rather than that of Alfarrar et al. (2022), since the latter involves an additional reassignment of the current prediction according to previously stored certificates, which introduces a relatively subjective post-hoc modification to the certification result. In contrast, ANCER is more consistent with our view that an issued certification represents a robustness commitment of the framework that should not be modified proactively. Since ANCER is designed for image inputs whereas CRICO operates on textual representations, the covariance construction and imple-

mentation details differ substantially, we therefore reimplemented this component for CRICO, with the complete implementation provided in our code repository.

Concretely, we maintain a memory set (a triple repository): $M = (\mathbf{x}_i, r_i, c_i)_{i=1}^n$, where $\mathbf{x}_i$ denotes a processed sample, r_i denotes its certified region, and c_i is the certified class. For a newly arrived sample $(\mathbf{x}_{n+1}, r_{n+1}, c_{n+1})$, the core objective of the memory-based algorithm is as follows: when we attempt to issue a certification statement for $\mathbf{x}_{n+1}$ that "within a certain region, the model prediction is constantly c_{n+1} ," we first verify whether this new statement would introduce a logical contradiction with any existing certified statements stored in M . For example, there may be intersecting certification regions, but the predicted labels are different. If a potential inconsistency is detected, we resolve it by shrinking r_{n+1} (thereby tightening the certified region) to eliminate the conflict, or, when necessary, by abstaining from issuing this certification statement for the sample.

We emphasize that the guiding principle of memory-based strategy is that every triple stored in M corresponds to a statement that has already been certified and accepted. To ensure the stability and traceability of CRICO framework’s outputs, we do not allow retroactive modification or deletion of previously stored certification results. This design is intended to provide reliable, persistent robustness guarantees that do not flip midway, rather than allowing certification conclusions to vary with the order in which data are processed.

However, this design leads to order dependency issues in memory-based mechanisms. The ANCER (Eiras et al., 2022) and Alfarrá et al. (2022) also reports this problem. Our experimental report in the Appendix B.1 states that conflicting overlaps between certified regions of different samples occur only rarely in practice, and thus the practical impact of order dependence is generally limited.

5 Experiments

5.1 Setup

The experiments in this paper are conducted from two perspectives. First, we evaluate the main certified performance of the CRICO framework on standard text classification tasks with both conventional models and LLMs to see whether the principle *Validity* holds. Second, we examine the value of CRICO in security-related attack scenarios. We

first introduce the main experimental setup. More experimental settings and results can be found in the Appendix.

Models and Datasets We use three conventional text classification models. The first is a 1-layer bidirectional LSTM with 150 hidden units (Graves, 2012). The second is TextCNN (Kim, 2014), where the number of filters is set to 100 for IMDB, 200 for AG News and 500 for Amazon to account for dataset size differences. The third is the bert-base-uncased BERT model (Devlin et al., 2019). To evaluate CRICO in LLM settings, we also use two instruction-tuned LLMs: Qwen/Qwen2.5-7B-Instruct¹ (Qwen Team, 2024) and meta-llama/Llama-3.2-3B-Instruct² (Llama Team, 2024).

For datasets, we use three standard text classification benchmarks, IMDB (Maas et al., 2011), AG News (Zhang et al., 2015), and Amazon (McAuley and Leskovec, 2013), to evaluate the main performance of CRICO on classification tasks.

Embedding Layers For BiLSTM and TextCNN, we use GloVe embeddings (Pennington et al., 2014). For BERT, Qwen, and Llama, we feed the input ids generated by the tokenizer into the model and perform the conversion from input ids to model embeddings inside the model. This allows CRICO to operate consistently in the embedding space across different architectures.

Baselines We compare CRICO with three representative NLP certified robustness baselines. SAFER (Ye et al., 2020) is the first work to apply randomized smoothing to NLP and is a classical smoothing-based method. Unlike SAFER, which uses random word substitution, RanMask (Zeng et al., 2023) adopts random ablation and represents a mask-based smoothing method. Text-CRS (Zhang et al., 2024) is also based on word-substitution smoothing, but it searches for synonyms in the word embedding space using a staircase distribution.

Evaluation Metrics We use the following evaluation metrics:

¹The Huggingface page of Qwen2.5-7B-Instruct is: <https://huggingface.co/Qwen/Qwen2.5-7B-Instruct>

²The Huggingface page of Llama-3.2-3B-Instruct is: <https://huggingface.co/meta-llama/Llama-3.2-3B-Instruct>

Clean Accuracy. Acc_{clean} measures the classification accuracy on clean test samples, i.e., the accuracy of clean samples under the base classifier: $\frac{1}{N} \sum_{i=1}^N \mathbf{1}[f(\mathbf{x}_i) = y_i]$.

Certified Accuracy at Radius r : $Acc_{cert}(r) = \frac{1}{N} \sum_{i=1}^N \mathbf{1}[g(\mathbf{x}_i) = y_i \wedge R_i \geq r]$.

Average Certified Radius. The Average Certified Radius serves to reflect the certification capability itself; therefore, sampling is restricted solely to those inputs where certification was successful. Thus, letting $\mathcal{S} = \{i : g(\mathbf{x}_i) = y_i\}$, the Average Certified Radius is given by: $\bar{R} = \frac{1}{|\mathcal{S}|} \sum_{i=1}^{|\mathcal{S}|} R_i$.

Median Certified Radius. The same as Average Certified Radius, we report Median Certified Radius (R_M) over correctly classified samples.

It should be noted that Acc_{clean} is evaluated on the base classifier $f(\mathbf{x})$, whereas Acc_{cert} is evaluated on the randomized smoothed classifier $g(\mathbf{x})$. Since they are not exactly the same predictive model, Acc_{cert} is usually expected to be no higher than Acc_{clean} , but the opposite may occasionally occur.

Code for our experiments can be found in: <https://github.com/ShudiGuo/crico-certified-robustness.git>

5.2 Results

Certified Robustness Performance of CRICO Framework Table 1 presents the main results of CRICO with $\beta = 1$. We use this value as the default setting.

CRICO provides non-trivial certified robustness across most settings, with performance varying across model architectures. BERT achieves the most stable results and maintains high certified accuracy at large radii. BiLSTM and TextCNN perform well on IMDB and AG News. They perform moderately on Amazon, which is partly limited by the base model, whose Acc_{clean} is also not very high.

BERT exhibits identical median radii across the three datasets. This is because BERT’s predictions satisfy $p_A > 99\%$, a condition that causes Monte Carlo sampling and the confidence bound configuration to jointly determine a maximum finite certified radius.

For LLMs, CRICO still offers certified robustness guarantees, though with lower certified accuracy. One possible reason for the lower certified accuracy is that perturbations injected into the embedding space are further processed by Transformer

layers, leading to less stable class-probability estimates under Monte Carlo sampling.

Moreover, to preserve the black-box nature of our setting, we directly evaluate LLMs without extensive task-specific fine-tuning, which also explains why their clean accuracy is lower than the performance typically reported for fully adapted LLM classifiers. Especially, the near-random low accuracies of Qwen on Amazon and Llama on AG News may primarily result from limited zero-shot compatibility between the specific models and datasets, rather than performance degradation caused by CRICO.

Overall, these results show that CRICO can enlarge certified regions effectively, while its practical performance remains closely tied to the stability and task adaptation of the underlying classifier.

Comparison of Certified Robustness Performance Against Baselines Table 2 compares CRICO with existing methods on IMDB. Since these baselines are mainly designed for discrete text space, they do not directly provide certified radius. Thus we report the certified accuracy of CRICO at $r = 0$ and $\beta = 0.7$, and focus on the certified coverage rather than the radius magnitude.

CRICO achieves competitive performance on BiLSTM and TextCNN. In particular, CRICO reaches 89.40% on TextCNN, substantially outperforming baselines. On BiLSTM, CRICO obtains 77.40%, which is comparable to RanMask, indicating that the covariance optimization can effectively improve certified robustness for conventional classifiers in the embedding space.

However, the result of CRICO on BERT is less competitive, with a certified accuracy of 59.20%. One possible reason is that we did not extensively tune the Monte Carlo sampling size, which may lead to less stable class-probability estimates. This result does not necessarily indicate that CRICO is unsuitable for BERT, but suggests that more extensive hyperparameter tuning may be required for complex model architectures.

Ablation: w/o Token-Importance Weighting

We conduct an ablation by removing the token-importance weighting from the objective while retaining the same β -controlled regularization strength:

$$\mathcal{H}_{w/o-imp}(\Sigma) = n \log r(\Sigma) + \beta \sum_{i=1}^n \log \sigma_i. \quad (12)$$

Models	Datasets	Acc_{clean} (%)	$Acc_{cert}(r)$ (%)						$\bar{R}$	R_M
			@0	@0.5	@1.0	@1.5	@2.0	@2.5		
BiLSTM	IMDB	78.03	77.80	77.80	77.20	75.30	70.90	61.40	2.9814	3.3053
	AG News	91.26	84.80	83.20	79.30	71.30	65.50	59.70	2.8008	3.4320
	Amazon	62.52	53.00	46.00	35.10	25.40	16.80	8.70	1.5289	1.4508
TextCNN	IMDB	88.44	87.40	79.50	39.90	13.40	5.90	3.00	1.0688	0.9453
	AG News	92.09	73.50	69.70	59.20	45.90	32.10	20.70	1.9103	1.8086
	Amazon	58.87	11.70	5.90	1.90	0.50	0.10	0.00	0.5658	0.5072
BERT	IMDB	83.43	59.80	59.80	59.80	59.60	59.40	59.40	2.7405	2.7487
	AG News	90.04	87.60	87.60	87.20	87.20	87.00	85.60	2.7285	2.7487
	Amazon	57.87	49.60	49.60	49.00	48.40	47.00	40.20	2.6044	2.7487
Llama	IMDB	57.04	49.00	34.00	28.00	24.50	22.00	20.00	1.5391	1.4691
	AG News	29.67	27.00	8.50	3.00	3.00	3.00	3.00	0.6779	0.4319
	Amazon	41.38	13.00	12.50	12.50	12.00	12.00	12.00	2.6053	2.7487
Qwen	IMDB	77.05	53.50	47.50	36.50	23.50	15.00	12.00	1.5023	1.3809
	AG News	42.97	31.50	24.00	17.50	7.00	2.50	2.00	1.0908	1.1848
	Amazon	20.06	20.00	18.50	15.50	12.00	8.00	8.00	1.7788	1.7144

Table 1: Comparison on multiple models and datasets

Methods	BiLSTM	TextCNN	BERT
SAFER	72.60	88.80	73.60
RanMask	78.80	77.20	85.60
Text-CRS	76.00	86.40	88.80
CRICO($r=0$)	77.40	89.40	59.20

Table 2: Comparison with Baselines on IMDB (%)

These ablation experiments aim to compare the differences between "uniform volume preference" and "semantic-aware volume preference".

Table 3 reports the ablation results where "w/o" denotes the variant that removes the token-importance weights. Since this ablation mainly focuses on the change in certified coverage, we directly compare the certified accuracy at $r = 0$.

Overall, removing token-importance weighting leads to a clear drop in certified accuracy in most settings, especially for BiLSTM and TextCNN. Across different values of β , the w/o variant is generally inferior to the full CRICO framework. This demonstrates that token-importance-guided anisotropic allocation effectively guides the noise budget toward decision-relevant directions, making it a key mechanism for improving the certified robustness of CRICO. Without the constraints of this regularization mechanism, the ellipsoidal region is highly prone to crossing the decision boundary, resulting lower accuracy.

One exception is observed on BERT with IMDB, where the w/o variant slightly outperforms the full model. A possible explanation is that the average text length in the IMDB significantly exceeds that

of AG News and Amazon, causing IMDB text to exhibit instability within multi-layer Transformers.

Certified Robustness of CRICO against Adversarial Attacks

To further evaluate the certified robustness of CRICO in security-related attack scenarios, we construct adversarial examples using black-box text attacks and examine whether CRICO can recover and certify correct predictions on these successfully attacked samples. In security-related attack scenarios, we use the Textattack³ framework (Morris et al., 2020) to generate Textbugger (Li et al., 2018) and Textfooler (Jin et al., 2020) black-box adversarial examples. The performance of CRICO against adversarial attacks on LLMs is presented in the Appendix B.5.

It is important to note that we use base models to generate adversarial examples and retain only the samples for which the attack succeeds. Therefore, in this security-related setting, all retained samples are misclassified by the base model, i.e., $Acc_{clean} = 0\%$. We then evaluate whether CRICO can recover correct predictions and provide certified radii for these recovered samples.

The results show that, even when $Acc_{clean} = 0\%$, CRICO can still recover a non-trivial portion of correct predictions and assign certified radius ranges to them. As r increases from 0 to 2.5, the certified accuracy generally decreases smoothly, indicating that CRICO provides not only prediction recovery on adversarial samples but also continuous and radius-controlled guarantees. These results demonstrate the practical potential of CRICO as a

³The TextAttack project is available at: <https://github.com/QData/TextAttack/tree/master>

Models	Datasets	0.0	0.1	0.3	0.5	0.7	0.9	1.0	1.5	2.0	2.5
BiLSTM	IMDB	77.30	76.50	79.50	74.80	77.40	75.90	77.80	75.70	75.10	70.30
	IMDB(w/o)		61.70	55.10	52.10	54.40	57.30	55.00	62.10	60.20	63.10
	AG News	89.60	91.50	89.60	90.90	88.60	86.90	84.80	84.20	79.20	75.00
	AG News(w/o)		74.60	62.80	60.90	58.70	62.40	62.70	63.70	64.50	66.70
	Amazon	56.30	56.60	57.60	55.00	53.50	49.70	53.00	46.90	43.90	42.00
	Amazon(w/o)		40.80	24.60	26.70	27.00	25.20	26.80	26.70	28.20	29.90
TextCNN	IMDB	85.50	88.20	85.80	87.60	89.40	86.70	87.40	86.70	87.30	86.40
	IMDB(w/o)		73.70	58.00	63.30	60.40	60.00	59.30	55.40	52.40	52.70
	AG News	83.80	83.10	82.60	81.20	76.90	73.40	73.50	60.30	51.30	49.60
	AG News(w/o)		41.20	33.90	32.60	32.00	30.60	31.90	27.70	27.30	29.80
	Amazon	14.00	14.10	11.80	13.00	11.10	12.20	11.70	12.60	11.50	10.10
	Amazon(w/o)		12.30	3.90	0.60	0.20	0.00	0.20	0.20	0.30	0.20
BERT	IMDB	58.80	59.20	60.20	58.20	59.20	57.80	59.80	59.60	59.20	63.20
	IMDB(w/o)		67.00	61.00	63.00	62.80	62.00	62.20	62.20	61.40	60.80
	AG News	87.20	92.40	91.80	89.00	87.20	91.40	87.60	88.40	90.00	87.40
	AG News(w/o)		88.40	82.80	85.60	85.60	83.80	80.20	80.80	77.60	72.40
	Amazon	52.40	55.00	49.20	51.40	47.40	49.80	49.60	44.80	45.40	43.20
	Amazon(w/o)		42.80	32.60	33.00	33.20	32.20	29.80	31.40	32.40	26.80

Table 3: Ablation $Acc_{cert}(r = 0)$ (%) results under different values of β

Models	Datasets	r	TextFooler	TextBugger
BiLSTM	IMDB	0	16.07	26.81
		0.5	16.07	26.81
		1.5	14.67	24.74
		2.5	9.60	15.09
	AG News	0	39.33	34.01
		0.5	38.87	33.36
		1.5	25.53	18.39
		2.5	8.33	4.62
	Amazon	0	33.90	32.59
		0.5	28.67	26.11
		1.5	11.00	9.93
		2.5	0.72	0.66
TextCNN	IMDB	0	51.47	55.07
		0.5	31.20	34.33
		1.5	0.53	0.93
		2.5	0.00	0.00
	AG News	0	30.00	29.53
		0.5	27.00	26.27
		1.5	8.40	9.60
		2.5	0.13	0.07
	Amazon	0	7.46	6.97
		0.5	1.39	1.22
		1.5	0.00	0.00
		2.5	0.00	0.00

Table 4: Certified robustness performance(%) in security-related attack scenarios

certified defense in security-related scenarios.

6 Conclusion

We propose CRICO, an input-dependent covariance optimization framework for certified robustness in the text embedding space. Compared with isotropic smoothing, CRICO produces certified regions that better adapt to local decision behavior. Experiments show that CRICO provides effective and stable certified robustness across different mod-

els and tasks.

Limitations

CRICO still has several limitations.

First, CRICO is affected by the architecture of the base model and the input length. Complex Transformer models may introduce instability in Monte Carlo estimation, while leave-one-out token-importance estimation may exhibit attention decay on long texts. We adopt leave-one-out because it provides a simple and effective black-box solution under the already substantial computational cost of randomized smoothing. Alternative methods, including interpretability-based approaches such as SHAP, may improve long-text stability but would introduce additional overhead when combined with covariance optimization. Future work will explore more efficient importance estimation strategies for complex models and long-text scenarios.

Second, in LLM experiments, we adopt a unified and strict black-box setting for all models, without any task-specific fine-tuning. Consequently, the LLMs are not specifically adapted to the downstream classification tasks, which may result in lower certification performance than fully trained conventional classifiers. Future work may further investigate lightweight task adaptation while preserving the black-box certification setting.

Third, CRICO is currently designed and evaluated primarily for classification tasks, including the LLM-based experiments in this work. Its applicability to generative tasks and other more complex NLP scenarios remains to be further explored.

References

- Sara Abdali, Richard Anarfi, Carlos J Barberan, Jia He, and Erfan Shayegani. 2024. Securing large language models: Threats, vulnerabilities and responsible practices. *arXiv preprint arXiv:2403.12503*.
- Motasem Alfarra, Adel Bibi, Philip HS Torr, and Bernard Ghanem. 2022. Data dependent randomized smoothing. In *Uncertainty in Artificial Intelligence*, pages 64–74. PMLR.
- Moustafa Alzantot, Yash Sharma, Ahmed Elgohary, Bo-Jhang Ho, Mani Srivastava, and Kai-Wei Chang. 2018. Generating natural language adversarial examples. *arXiv preprint arXiv:1804.07998*.
- Yonatan Belinkov and Yonatan Bisk. 2017. Synthetic and natural noise both break neural machine translation. *arXiv preprint arXiv:1711.02173*.
- Jeremy Cohen, Elan Rosenfeld, and Zico Kolter. 2019. Certified adversarial robustness via randomized smoothing. In *International Conference on Machine Learning*, pages 1310–1320. PMLR.
- 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.
- Yi Dong, Ronghui Mu, Yanghao Zhang, Siqi Sun, Tianle Zhang, Changshun Wu, Gaojie Jin, Yi Qi, Jinwei Hu, Jie Meng, Saddek Bensalem, and Xiaowei Huang. 2025. Safeguarding large language models: A survey. *Artificial Intelligence Review*, 58(12):382.
- Steffen Eger, Gözde Gül Şahin, Andreas Rücklé, Ji-Ung Lee, Claudia Schulz, Mohsen Mesgar, Krishnkant Swarnkar, Edwin Simpson, and Iryna Gurevych. 2019. Text processing like humans do: Visually attacking and shielding nlp systems. *arXiv preprint arXiv:1903.11508*.
- Francisco Eiras, Motasem Alfarra, Philip Torr, M. Pawan Kumar, Puneet K. Dokania, Bernard Ghanem, and Adel Bibi. 2022. [ANCER: Anisotropic certification via sample-wise volume maximization](#). *Transactions on Machine Learning Research*.
- Ji Gao, Jack Lanchantin, Mary Lou Soffa, and Yanjun Qi. 2018. Black-box generation of adversarial text sequences to evade deep learning classifiers. In *2018 IEEE Security and Privacy Workshops (SPW)*, pages 50–56. IEEE.
- Sven Gowal, Krishnamurthy Dj Dvijotham, Robert Stanforth, Rudy Bunel, Chongli Qin, Jonathan Uesato, Relja Arandjelovic, Timothy Mann, and Pushmeet Kohli. 2019. Scalable verified training for provably robust image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4842–4851.
- Alex Graves. 2012. Long short-term memory. *Supervised sequence labelling with recurrent neural networks*, pages 37–45.
- Hanbin Hong and Yuan Hong. 2022. Certified adversarial robustness via anisotropic randomized smoothing. *arXiv preprint arXiv:2207.05327*.
- Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55.
- Mohit Iyyer, John Wieting, Kevin Gimpel, and Luke Zettlemoyer. 2018. Adversarial example generation with syntactically controlled paraphrase networks. *arXiv preprint arXiv:1804.06059*.
- Robin Jia and Percy Liang. 2017. Adversarial examples for evaluating reading comprehension systems. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2021–2031.
- Robin Jia, Aditi Raghunathan, Kerem Göksel, and Percy Liang. 2019. Certified robustness to adversarial word substitutions. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4129–4142.
- Di Jin, Zhijing Jin, Joey Tianyi Zhou, and Peter Szolovits. 2020. Is bert really robust? a strong baseline for natural language attack on text classification and entailment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8018–8025.
- Yoon Kim. 2014. [Convolutional neural networks for sentence classification](#). In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751, Doha, Qatar. Association for Computational Linguistics.
- Mathias Lecuyer, Vaggelis Atlidakis, Roxana Geambasu, Daniel Hsu, and Suman Jana. 2019. Certified robustness to adversarial examples with differential privacy. In *2019 IEEE Symposium on Security and Privacy (SP)*, pages 656–672. IEEE.
- Jinfeng Li, Shouling Ji, Tianyu Du, Bo Li, and Ting Wang. 2018. Textbugger: Generating adversarial text against real-world applications. *arXiv preprint arXiv:1812.05271*.
- Bin Liang, Hongcheng Li, Miaoqiang Su, Pan Bian, Xirong Li, and Wenchang Shi. 2017. Deep text classification can be fooled. *arXiv preprint arXiv:1704.08006*.

- Yupe Liu, Yuqi Jia, Runpeng Geng, Jinyuan Jia, and Neil Zhenqiang Gong. 2024. Formalizing and benchmarking prompt injection attacks and defenses. In *33rd USENIX Security Symposium (USENIX Security 24)*, pages 1831–1847.
- Llama Team. 2024. [The llama 3 herd of models](#). *arXiv preprint arXiv:2407.21783*.
- Qian Lou, Xin Liang, Jiaqi Xue, Yancheng Zhang, Rui Xie, and Mengxin Zheng. 2024. Cr-utp: Certified robustness against universal text perturbations on large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 9863–9875.
- Andrew Maas, Raymond E Daly, Peter T Pham, Dan Huang, Andrew Y Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*, pages 142–150.
- Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. 2017. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*.
- Julian McAuley and Jure Leskovec. 2013. Hidden factors and hidden topics: understanding rating dimensions with review text. In *Proceedings of the 7th ACM Conference on Recommender Systems*, pages 165–172.
- Takeru Miyato, Andrew M Dai, and Ian Goodfellow. 2016. Adversarial training methods for semi-supervised text classification. *arXiv preprint arXiv:1605.07725*.
- John X. Morris, Eli Lifland, Jin Yong Yoo, Jake Grigsby, Di Jin, and Yanjun Qi. 2020. [Textattack: A framework for adversarial attacks, data augmentation, and adversarial training in nlp](#). *Preprint*, arXiv:2005.05909.
- NeurAlchemy. 2026. [Prompt injection and jailbreak detection dataset](#).
- Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.
- Qwen Team. 2024. [Qwen2.5: A party of foundation models](#).
- Alexander Robey, Eric Wong, Hamed Hassani, and George J Pappas. 2023. Smoothllm: Defending large language models against jailbreaking attacks. *arXiv preprint arXiv:2310.03684*.
- Hadi Salman, Jerry Li, Ilya Razenshteyn, Pengchuan Zhang, Huan Zhang, Sebastien Bubeck, and Greg Yang. 2019. Provably robust deep learning via adversarially trained smoothed classifiers. *Advances in Neural Information Processing Systems*, 32.
- Suranjana Samanta and Sameep Mehta. 2017. Towards crafting text adversarial samples. *arXiv preprint arXiv:1707.02812*.
- Siqi Sun, Procheta Sen, and Wenjie Ruan. 2024. Crowd: Certified robustness via weight distribution for smoothed classifiers against backdoor attack. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 17056–17070.
- Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. 2013. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*.
- Lingxiao Wang, Jing Huang, Kevin Huang, Ziniu Hu, Guangtao Wang, and Quanquan Gu. 2020. Improving neural language generation with spectrum control. In *International Conference on Learning Representations*.
- Wenqi Wang, Run Wang, Lina Wang, Zhibo Wang, and Aoshuang Ye. 2019. Towards a robust deep neural network in texts: A survey. *arXiv preprint arXiv:1902.07285*.
- Eric Wong and Zico Kolter. 2018. Provable defenses against adversarial examples via the convex outer adversarial polytope. In *International Conference on Machine Learning*, pages 5286–5295. PMLR.
- Mao Ye, Chengyue Gong, and Qiang Liu. 2020. Safer: A structure-free approach for certified robustness to adversarial word substitutions. *arXiv preprint arXiv:2005.14424*.
- Sibo Yi, Yule Liu, Zhen Sun, Tianshuo Cong, Xinlei He, Jiaying Song, Ke Xu, and Qi Li. 2024. Jailbreak attacks and defenses against large language models: A survey. *arXiv preprint arXiv:2407.04295*.
- Yuan Zang, Fanchao Qi, Chenghao Yang, Zhiyuan Liu, Meng Zhang, Qun Liu, and Maosong Sun. 2019. Word-level textual adversarial attacking as combinatorial optimization. *arXiv preprint arXiv:1910.12196*.
- Jiehang Zeng, Jianhan Xu, Xiaoqing Zheng, and Xuanjing Huang. 2023. Certified robustness to text adversarial attacks by randomized [mask]. *Computational Linguistics*, 49(2):395–427.
- Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. *Advances in Neural Information Processing Systems*, 28.
- Xinyu Zhang, Hanbin Hong, Yuan Hong, Peng Huang, Binghui Wang, Zhongjie Ba, and Kui Ren. 2024. Text-crs: A generalized certified robustness framework against textual adversarial attacks. In *2024 IEEE Symposium on Security and Privacy (SP)*, pages 2920–2938. IEEE.
- Chen Zhu, Yu Cheng, Zhe Gan, Siqi Sun, Tom Goldstein, and Jingjing Liu. 2019. Freelib: Enhanced adversarial training for natural language understanding. *arXiv preprint arXiv:1909.11764*.

A Detailed Algorithms of CRICO

In this section, we provide the detailed algorithm procedures used in CRICO.

A.1 Obtaining the Importance of Each Token

Algorithm 1 describes the procedure for estimating token-wise importance scores. Given an input sequence $\mathbf{x} = (x_1, \dots, x_n)$, we first obtain the prediction of the base classifier and identify the original predicted class. Then, each token is masked once, and the change in the target-class logit is used as a leave-one-out estimate of the token’s contribution to the current decision. Tokens that induce larger logit drops are assigned higher importance scores. Finally, the scores are clipped to be non-negative and normalized into $W = (w_1, \dots, w_n)$, which is later used to guide the anisotropic allocation of token-wise noise scales.

A.2 Input-Dependent Covariance Optimization

Algorithm 2 presents the optimization procedure for the input-dependent covariance matrix Σ . Starting from an initialized diagonal covariance matrix, the algorithm repeatedly samples Gaussian perturbations from the current covariance, estimates the smoothed class probabilities via Monte Carlo sampling, and computes the corresponding certified radius estimate. The covariance is then updated by maximizing the objective in Eq. (11), which jointly considers the certified radius and the importance-weighted anisotropic allocation term. After each update, we project the token-wise noise scales back to the feasible region by enforcing the global noise budget $\hat{\sigma}$ and the lower bound $\sigma_{\min}$. The final covariance $\Sigma^{(K)}$ is used to construct the certified ellipsoidal region for the input.

B More Experimental Results and Analysis

B.1 Order-Dependence Analysis of the Memory-Based Mechanism

The memory-based consistency mechanism is theoretically order-dependent. Prior work has reported that conflicting intersections between certified regions occur only rarely in practice, making the practical impact of order dependence generally limited (Alfarra et al., 2022; Eiras et al., 2022). To further investigate this issue in CRICO, we conducted an additional experiment on 1,000 samples

and recorded the numbers of memory-induced **ABSTAIN** and **SHRINK** operations.

ABSTAIN is triggered when two regions intersect and have different predicted classes while $\mathbf{x}_{n+1}$ lies inside an existing certified region. **SHRINK** is triggered when the regions intersect and have different predicted classes but $\mathbf{x}_{n+1}$ lies outside the existing region, in which case the newly generated region is shrunk to eliminate the conflicting intersection before being added to memory. **Final ABSTAIN** denotes samples rejected because of the Clopper–Pearson confidence bound.

As shown in Table 5, only five **SHRINK** operations are observed across all nine experimental settings (9,000 processed samples), with at most two occurring in any single setting, while no memory-induced **ABSTAIN** is observed. These results indicate that conflicting overlaps are rare in CRICO and therefore provide relatively few opportunities for input order to affect the final certification results.

B.2 Runtime per Example

We measured the average certification time on the IMDB dataset using a fixed setting of $\beta = 1$, 10000 Monte Carlo sampling for certification and 50 covariance-optimization iterations on an NVIDIA GeForce RTX 4090 GPU (24 GB). Since CRICO certification consists of covariance optimization, Monte Carlo sampling and other implementation overhead, we decompose the total runtime as: $T_{\text{cert}} = T_{\text{opt}} + T_{\text{MC}} + T_{\text{other}}$. We report the important parts in Table 6. Values are reported as mean $\pm$ sample standard deviation over the timed examples. The results show that covariance optimization constitutes the dominant computational cost of CRICO. This observation further motivates our framework design principle: **Simplicity and Efficiency**, so as to avoid introducing additional computational overhead on top of the already expensive covariance optimization procedure.

B.3 LLM Experimental Setup

For the LLM experiments, we evaluate CRICO on Qwen2.5-7B-Instruct and Llama3.2-3B-Instruct. Since Qwen and Llama are causal language models, while our evaluation requires discrete class predictions, we perform lightweight task adaptation to align them with text classification setting. It is important to note that this adaptation does not change the black-box nature of CRICO, only convert generative LLMs into a discriminative classification interface.

Algorithm 1: Obtaining the importance of each token

Function GetImportance($\mathbf{x} = (x_1, \dots, x_n)$):
 $logits \leftarrow F(\mathbf{x})$ // logits for input $\mathbf{x}$
 $c \leftarrow \arg \max_{y \in \mathcal{Y}} logits$ // original prediction class
 $s \leftarrow \max logits$ // the logit value corresponding to c
 Initialize $W \leftarrow (0, \dots, 0) \in \mathbb{R}^n$
 for $i \leftarrow 1$ **to** n **do**
 $\mathbf{x}^{(i)} \leftarrow \mathbf{x}$
 $x_i^{(i)} \leftarrow [\text{MASK}]$
 $s^{(i)} \leftarrow \max F(\mathbf{x}^{(i)})$ // the logit value corresponding to c
 $\Delta_i \leftarrow s - s^{(i)}$
 $w_i \leftarrow \max(\Delta_i, 0)$
 end
 $S \leftarrow \text{sum}(W)$
 $W \leftarrow W / (S + \eta)$
 return W

Algorithm 2: Σ Optimization Algorithm

Function OptimizeSigma($\mathbf{x} = (x_1, \dots, x_n), \hat{\sigma}, \sigma_{\min}, \kappa, K, num, \beta, \alpha, W$):
 Initialize $\Sigma^{(0)} = \text{diag}((\kappa \hat{\sigma})^2 I_d, \dots, (\kappa \hat{\sigma})^2 I_d)$
 for $k \leftarrow 1$ **to** K **do**
 sample $\epsilon \times num \in \mathcal{N}(0, \Sigma^{(k)})$
 $s_\theta \leftarrow \text{softmax}(F_\theta(v(\mathbf{x}) + \epsilon_i))$
 $\tau(\Sigma) \leftarrow \frac{1}{num} \sum_i^{num} s_\theta$
 $y_A \leftarrow \arg \max_y \tau_y(\Sigma), \hat{p}_A \leftarrow \max_{y_A} \tau_{y_A}(\Sigma), \hat{p}_B \leftarrow \max_{y \neq y_A} \tau_y(\Sigma)$
 $\hat{r}(\Sigma) \leftarrow \frac{1}{2}(\Phi^{-1}(\hat{p}_A) - \Phi^{-1}(\hat{p}_B))$ // Eq. (4)
 $\mathcal{H}(\Sigma^{(k)}) \leftarrow n \log \hat{r}(\Sigma) + \beta \sum_{i=1}^n w_i \log \sigma_i$ // Eq. (11)
 $\Sigma^{(k+1)} \leftarrow \Sigma^{(k)} + \alpha \nabla_{\Sigma^{(k)}} \mathcal{H}(\Sigma^{(k)})$
 if $\frac{1}{n} \sum_{i=1}^n \sigma_i^{(k+1)} > \hat{\sigma}$ **then**
 for $i \leftarrow 1$ **to** n **do**
 $\sigma_i^{(k+1)} \leftarrow \frac{\hat{\sigma}}{\bar{\sigma}} \cdot \sigma_i^{(k+1)}$
 end
 end
 for $i \leftarrow 1$ **to** n **do**
 $\sigma_i^{(k+1)} \leftarrow \max(\sigma_i^{(k+1)}, \sigma_{\min})$
 end
 end
 return $\Sigma^{(K)}$

Specifically, we implement a unified classification interface for LLMs. First, we provide a `_compute_embeddings()` function, allowing the external certification procedure to obtain or inject token embeddings directly. Second, the model supports both `input_ids` and `inputs_embeds` as inputs, which enables CRICO to perform embedding-space perturbation. Third, instead of using the full causal LM output with shape $\mathbb{R}^{B \times L \times V}$, we extract

the hidden state of the last valid token and project it onto a predefined label-word space, obtaining classification logits in $\mathbb{R}^{B \times C}$, where C is the number of classes. This design makes the LLM output compatible with the CRICO framework used in our main experiments. Fourth, For all datasets, the tokenizer’s `max_length` is uniformly set to 256.

For prompt construction, we follow three principles: the prompt should be fixed, concise, and

Models	Datasets	Memory ABSTAIN	Memory SHRINK	Final ABSTAIN
BiLSTM	IMDB	0	0	0
	AG News	0	0	1
	Amazon	0	0	154
TextCNN	IMDB	0	0	15
	AG News	0	2	35
	Amazon	0	1	181
BERT	IMDB	0	1	1
	AG News	0	0	1
	Amazon	0	1	221

Table 5: Order-dependence analysis of the memory-based consistency mechanism.

Model	Optimization (s/example)	MC Sampling (s/example)	Total Certification (s/example)
BiLSTM	10.5379 ± 0.8972	1.5672 ± 0.0071	12.1079 ± 0.8989
TextCNN	5.3159 ± 0.2763	1.8135 ± 0.0065	7.1322 ± 0.2771
BERT	8.3347 ± 0.6749	3.3096 ± 0.0196	11.6476 ± 0.6730

Table 6: Average certification time per example with $\beta = 1$.

use simple label words. Each prompt consists of three components: prefix, content, and suffix. The content part corresponds to the input text, while the prefix and suffix provide task instructions for classification. Since CRICO perturbs only the input content in the embedding space, especially in prompt injection experiments, we explicitly place the content between the prefix and suffix and construct a content_mask to indicate which tokens belong to the input text. This ensures that noise injection and covariance optimization are applied only to the content tokens, rather than to the task instruction or label description.

We design four prompt templates in total (Table 7), corresponding to IMDB, AG News, Amazon, and prompt injection attacks. All templates follow the same prefix-content-suffix structure, ensuring a consistent LLM classification interface, perturbation region, and certification pipeline.

B.4 Sensitivity Analysis of Hyperparameters

As shown in Eq.(11) and Algorithm 2, CRICO involves several hyperparameters, including $\sigma_{\min}$, κ , K , num , α , β , and the global noise budget $\hat{\sigma}$. Among them, $\sigma_{\min}$ constrains the lower bound of each token-wise standard deviation σ_i , κ controls the initialization scale of Σ , K and num denote the number of optimization iterations and Monte Carlo samples, respectively, and α is the gradient update step size. These parameters mainly

affect numerical stability, convergence behavior, or estimation accuracy, and are standard choices in stochastic optimization and Monte Carlo estimation rather than design-specific components of CRICO. Therefore, we test and fix them throughout the experiments as $\sigma_{\min} = 0.005$, $\kappa = 0.5$, $K = 50$, $num = 500$, and $\alpha = 0.05$. In the certification stage, we use 1000 Monte Carlo samples for predicting the smoothed label, 10000 samples of BiLSTM/TextCNN and 1000 samples of BERT/Qwen/LLaMA for estimating the certified region. We focus our hyperparameter analysis on β , which directly controls the contribution of the importance-weighted anisotropic allocation term in Eq. (11).

Fig. 2 shows the certified radius distributions under different values of β , where each boxplot corresponds to the correctly classified samples under a specific β setting. The results show that the influence of β is model- and dataset-dependent, but appropriate token-importance weighting generally leads to more favorable radius distributions. BiLSTM and BERT exhibit relatively stable certified radii, while TextCNN is more sensitive to the choice of β .

Fig. 3 reports the certified accuracy curves under different β values. As expected, $Acc_{cert}(r)$ decreases as the radius threshold r increases. Appropriate β values usually lead to better certified coverage, confirming the benefit of token-importance-

Dataset	Prompt Prefix	Suffix	Label Words
IMDB	Classify the sentiment of the following movie review as positive or negative. Review:	Answer:	negative, positive
Amazon	Classify the sentiment of the following product review as one of the five labels: awful, bad, okay, good, excellent. Review:	Answer:	awful, bad, okay, good, excellent
AG News	Classify the topic of the following news article as world, sports, business, or technology. Article:	Answer:	world, sports, business, technology
Prompt Injection	Classify whether the following input is benign or malicious for prompt injection detection. Input:	Answer:	benign, malicious

Table 7: Prompt templates used for LLM-based classification. Each input text is inserted between the prompt prefix and suffix. The label words are used to map the LLM output to task-specific class logits.

guided anisotropic allocation. Although Qwen and Llama obtain lower certified accuracy than conventional classifiers and BERT, their non-zero certified accuracy still shows that CRICO can provide certified robustness guarantees for LLMs.

We also observe that the median radius in Fig. 2 is fixed in some settings. This is because when the estimated $p_A > 99\%$, the Monte Carlo sampling size and confidence-bound configuration jointly determine a maximum finite certificate value. Once most samples satisfy this condition, the median radius concentrates around this maximum value. Overall, these results show that β is an important hyperparameter for CRICO and validate the effectiveness of token-importance-guided anisotropic covariance optimization.

B.5 Certified Robustness of CRICO against Adversarial Attacks on LLMs

For LLM-based security experiments, we use the `neuralchemy_prompt_injection`⁴ dataset (NeurAlchemy, 2026) for prompt injection detection. Unlike adversarial example generation frameworks such as TextAttack, NeurAlchemy already contains prompt injection attack samples. Therefore, we directly use the attack samples from NeurAlchemy for evaluation. With our prompt template, prompt injection detection is formulated as a binary classification task, where Llama and Qwen are required to predict whether an input is benign or malicious. Consistent with the previous attack setting, we only retain samples that successfully attack the

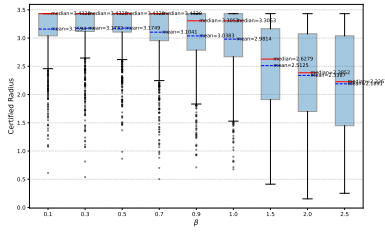
Models	r	NeurAlchemy
Llama	0	4.50
	0.5	2.50
	1.5	1.00
	2.5	0.00
Qwen	0	0.50
	0.5	0.00
	1.5	0.00
	2.5	0.00

Table 8: Certified robustness performance(%) on LLMs in security-related attack scenarios

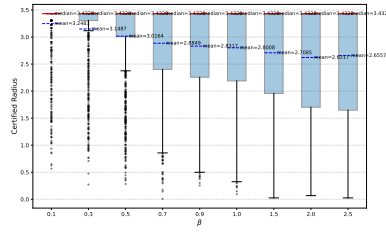
original Llama and Qwen models. Thus, in this LLM-based security evaluation, the base-model accuracy is $Acc_{clean} = 0\%$.

Table 8 reports the $Acc_{cert}(r)$ of CRICO on NeurAlchemy attack samples. Even though all retained samples are misclassified by the base models, CRICO can still recover a small portion of correct predictions and provide corresponding certified radius ranges, especially for Llama. Specifically, Llama achieves 4.50% certified accuracy at $r = 0$. In contrast, Qwen shows weaker recovery, achieving only 0.50% at $r = 0$ and dropping to 0 afterward. These results show that CRICO can still provide certified robustness guarantees in LLM prompt injection scenarios. The relatively low certified accuracy is mainly due to our black-box evaluation protocol. As a result, the base LLMs are not fully adapted to the prompt injection detection task, which limits the certified performance observed in this setting.

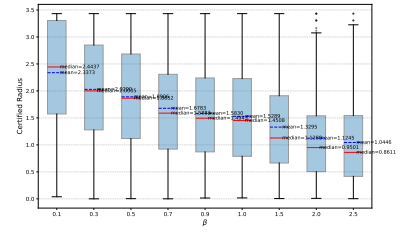
⁴The Huggingface page of `neuralchemy_prompt_injection` is: <https://huggingface.co/datasets/neuralchemy/Prompt-injection-dataset>



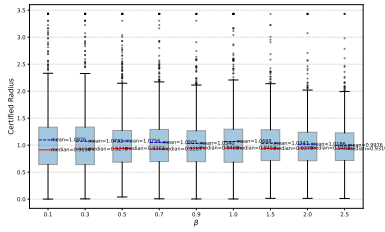
(a) BiLSTM on IMDB



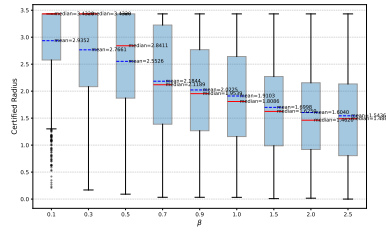
(b) BiLSTM on AG News



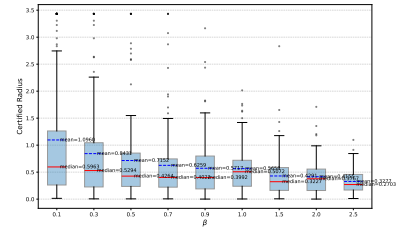
(c) BiLSTM on Amazon



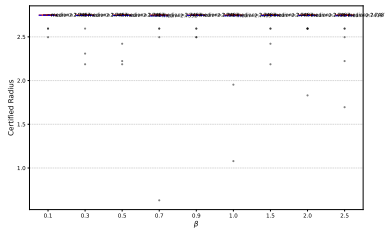
(d) TextCNN on IMDB



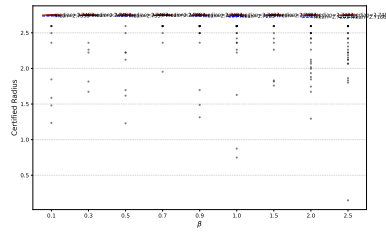
(e) TextCNN on AG News



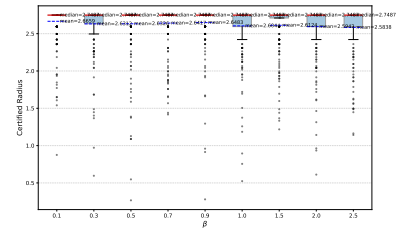
(f) TextCNN on Amazon



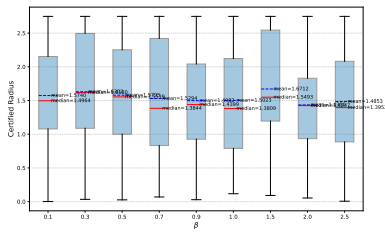
(g) BERT on IMDB



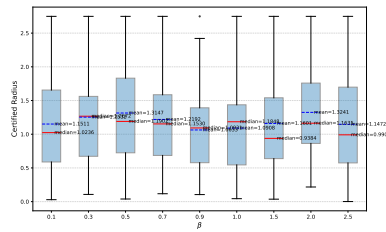
(h) BERT on AG News



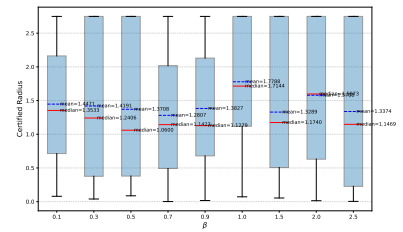
(i) BERT on Amazon



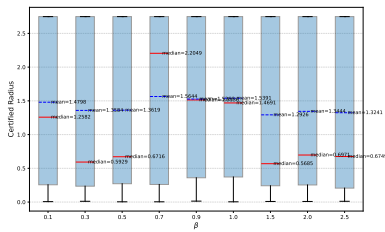
(j) Qwen on IMDB



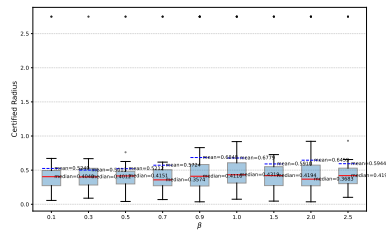
(k) Qwen on AG News



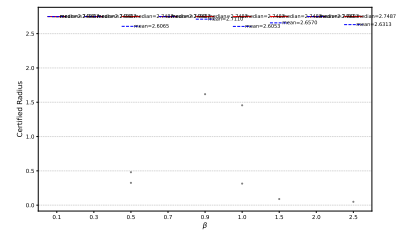
(l) Qwen on Amazon



(m) Llama on IMDB

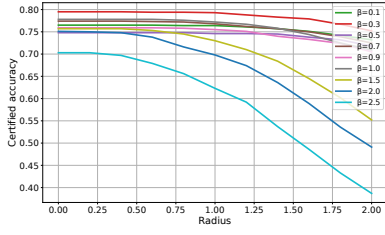


(n) Llama on AG News

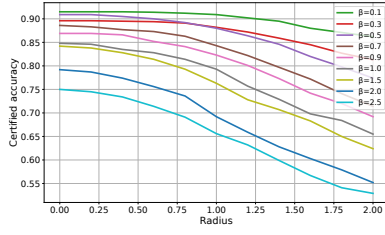


(o) Llama on Amazon

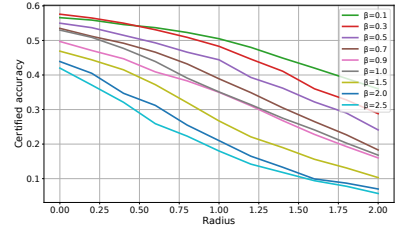
Figure 2: Boxplot of certified radii under different β



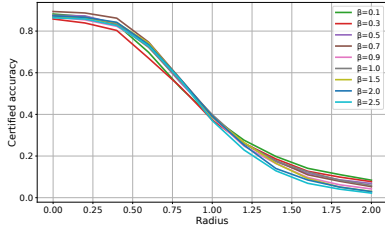
(a) BiLSTM on IMDB



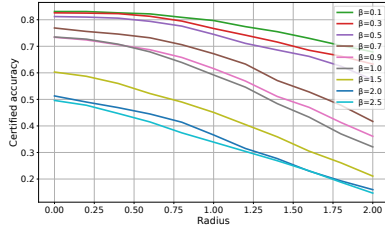
(b) BiLSTM on AG News



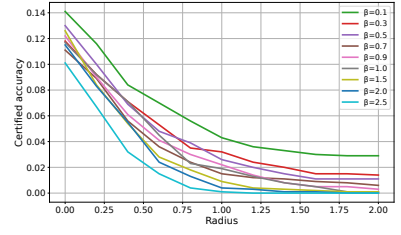
(c) BiLSTM on Amazon



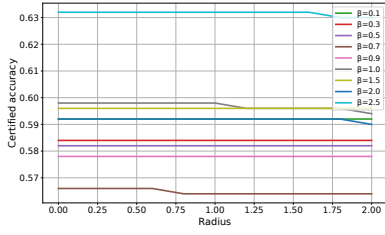
(d) TextCNN on IMDB



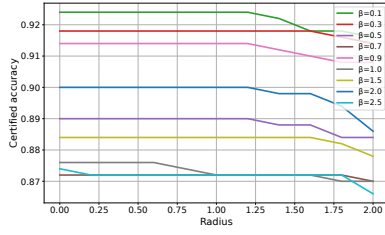
(e) TextCNN on AG News



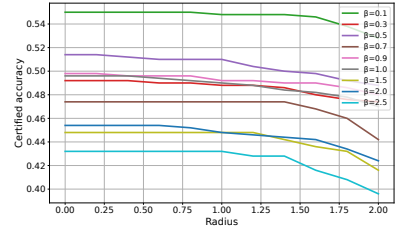
(f) TextCNN on Amazon



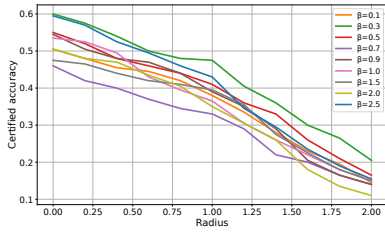
(g) BERT on IMDB



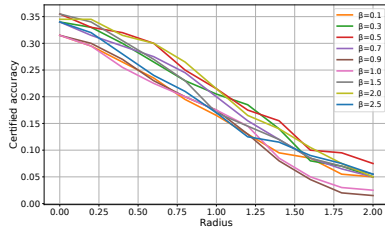
(h) BERT on AG News



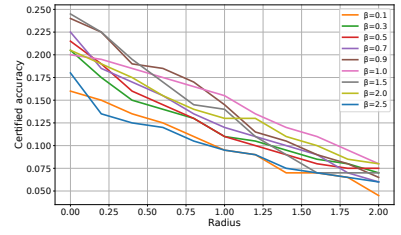
(i) BERT on Amazon



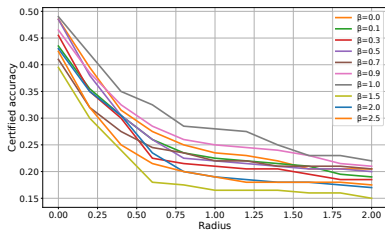
(j) Qwen on IMDB



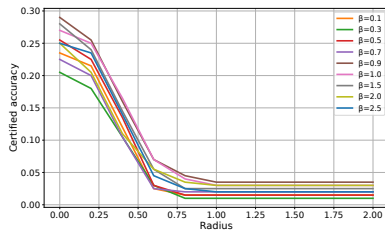
(k) Qwen on AG News



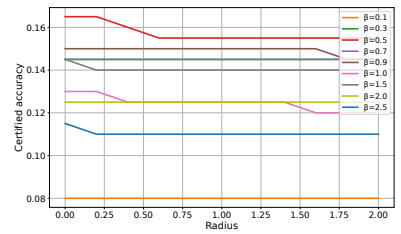
(l) Qwen on Amazon



(m) Llama on IMDB



(n) Llama on AG News



(o) Llama on Amazon

Figure 3: Certified robustness accuracy under different β