CluCERT: Certifying LLM Robustness via Clustering-Guided Denoising Smoothing

Zixia Wang, Gaojie Jin, Jia Hu, Ronghui Mu*

University of Exeter
{zw483, g.jin, j.hu, r.mu2}@exeter.ac.uk

Abstract

Recent advancements in Large Language Models (LLMs) have led to their widespread adoption in daily applications. Despite their impressive capabilities, they remain vulnerable to adversarial attacks, as even minor meaning-preserving changes such as synonym substitutions can lead to incorrect predictions. As a result, certifying the robustness of LLMs against such adversarial prompts is of vital importance. Existing approaches focused on word deletion or simple denoising strategies to achieve robustness certification. However, these methods face two critical limitations: (1) they yield loose robustness bounds due to the lack of semantic validation for perturbed outputs and (2) they suffer from high computational costs due to repeated sampling. To address these limitations, we propose CluCERT, a novel framework for certifying LLM robustness via clustering-guided denoising smoothing. Specifically, to achieve tighter certified bounds, we introduce a semantic clustering filter that reduces noisy samples and retains meaningful perturbations, supported by theoretical analysis. Furthermore, we enhance computational efficiency through two mechanisms: a refine module that extracts core semantics, and a fast synonym substitution strategy that accelerates the denoising process. Finally, we conduct extensive experiments on various downstream tasks and jailbreak defense scenarios. Experimental results demonstrate that our method outperforms existing certified approaches in both robustness bounds and computational efficiency.

1 Introduction

Large Language Models (LLMs) have seen rapid development in recent years. Their strong performance has enabled a wide range of real-world applications (Zhang et al. 2025; Lin et al. 2025; Chen et al. 2024), bringing significant benefits in terms of efficiency and automation. However, LLMs are vulnerable to adversarial inputs (Shayegani et al. 2023; Yi et al. 2024; Huang et al. 2024; Sun, Sen, and Ruan 2024; Wang, Hu, and Mu 2025), which can mislead predictions (Kumar et al. 2023; Mu et al. 2024). In particular, in textual tasks, even minor semantic-preserving perturbations such as synonym substitutions (Wang et al. 2021; Li et al. 2024) or paraphrasing (Fang et al. 2025) can easily mislead LLMs into producing incorrect or harmful outputs. This vulnerability becomes particularly pronounced when dealing with texts that contain subtle semantic distinctions (Tao et al. 2024). For example (as shown in Figure 1), the sentence *"The airline's response to the delayed flights was thoughtful..."* is likely to be interpreted as positive by an LLM. However, replacing *"thoughtful"* with a near-synonym such as *"polite"* can lead the model to a semantically distinct and potentially incorrect judgment.

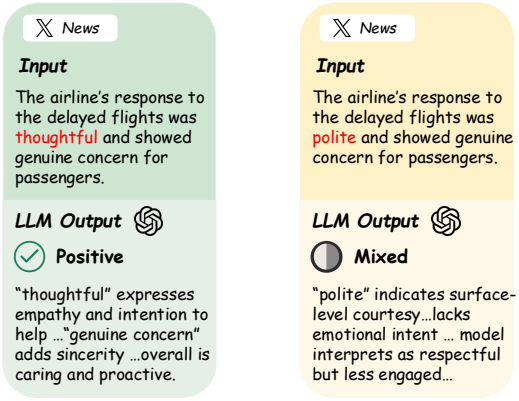

Figure 1: An example showing how a minimal synonym substitution can flip an LLM's sentiment prediction.

To counter this threat, researchers have proposed various empirical defenses, such as adversarial data augmentation (Cheng et al. 2020; Jin et al. 2025) and adversarial fine-tuning (Chen et al. 2020; Zhang et al. 2024b), which have shown some effectiveness in mitigating specific attacks. However, these approaches often lead to a reactive "arms race" (Jin et al. 2020), where evolving attack strategies continually require updated defense mechanisms. As a result, models must be frequently retrained or adapted, leading to an ongoing cycle of patching and evasion. To break this cycle, certified robustness (Raghunathan, Steinhardt, and Liang 2018; Weng et al. 2018; Jin, Yi, and Huang 2025; Sun and Ruan 2023; Zhang, Kouvaros, and Lomuscio 2025; Zhang et al. 2024a; Rocamora, Chrysos, and Cevher 2025) has emerged as a promising alternative. It offers mathematically provable guarantees that model predictions re-

*Corresponding Author

main consistent within a defined perturbation range.

Nevertheless, the large number of parameters, high computational cost, and limited access to internal components make many traditional certification methods hard to apply to LLMs. In contrast, randomized smoothing (Cohen, Rosenfeld, and Kolter 2019) provides a model-independent solution that uses probabilistic techniques to give robustness guarantees with high confidence. It does not require access to model parameters, making it suitable for black-box LLMs. This concept has been successfully established in the vision domain through randomized smoothing (Cohen, Rosenfeld, and Kolter 2019; Levine and Feizi 2020; Jia et al. 2022). Inspired by this, some studies (Jia et al. 2019; Ye, Gong, and Liu 2020; Chao et al. 2025) have applied randomized smoothing to natural language processing.

While randomized smoothing is widely explored in the vision domain, its application in text is limited. In the previous work, Zeng et al. (2021) adopt randomized ablation by replacing words with `[MASK]` tokens. This strategy disrupts semantic and grammatical structure, resulting in certificates that are ineffective against semantics-preserving attacks. Building on this, Ji et al. (2024) use LLMs to fill in masked tokens in an attempt to preserve meaning. However, their approach relies on unverifiable LLM outputs, which may introduce contextually inappropriate substitutions. Moreover, both methods suffer from low efficiency, as large-scale certification and repeated use of LLMs are computationally expensive.

To address these challenges, we propose CluCERT, an efficient certified robustness framework for LLMs, built upon a clustering-guided denoising smoothing strategy (as shown in Figure 2). CluCERT aims to enhance certified robustness bounds while maintaining computational efficiency. To this end, we highlight two key components that contribute to **efficiency**. First, we introduce the semantic refinement, where irrelevant tokens are removed based on their contribution to the input's core meaning. This results in a fixed-length, refined input that avoids repeated sampling on semantically unimportant words. Importantly, the certified radius is computed over this filtered input, ensuring that the robustness guarantee aligns with the core semantic content. Second, instead of relying on LLM-generated token replacements, we introduce a lightweight synonym substitution strategy based on WordNet and embedding similarity, which avoids model queries while preserving semantic consistency. To further **denoise**, we apply semantic clustering over the perturbed samples, filtering out samples that are semantically inconsistent with the original input. This process increases the probability of the most frequent response while reducing the impact of less frequent ones.

This paper's main **contributions** are summarized as follows: (i) We develop a fast synonym substitution strategy combined with a semantic refinement module that removes less informative tokens. This design significantly improves computational efficiency and reduces the overall cost of certification. (ii) We propose a novel framework, CluCERT, for certifying the robustness of LLMs. It incorporates a clustering-guided denoising module that preserves only meaning-consistent perturbations, resulting in tighter certified robustness bounds. We provide both theoretical analysis and empirical results to validate the effectiveness of our approach. (iii) We conduct extensive experiments across multiple tasks to demonstrate the effectiveness and efficiency of our approach. To the best of our knowledge, we are the first to apply certified robustness techniques to math word problem solving, a domain that requires precise semantics and sensitive to input perturbations.

2 Related Works

Randomized smoothing(Cohen, Rosenfeld, and Kolter 2019) has become a mainstream approach in certified robustness research due to its ability to provide probabilistic guarantees without requiring access to model internals. The core idea is to ensure that a model's prediction remains unchanged under perturbations within a predefined set (e.g., modifying up to n pixels), thereby establishing provable robustness bounds. To enhance the certified robustness bounds of randomized smoothing, Salman et al. (2020) first introduced the concept of *"denoising"smoothing* in the context of computer vision. Carlini et al. (2022) further proposed the integration of diffusion models to remove the added Gaussian noise during the smoothing process, thereby strengthening robustness guarantees for standard models. However, prior work has mainly focused on the image domain. In this work, we extend the denoising concept to text by introducing a semantic consistency mechanism that selects perturbations aligned with the original meaning. In the vision domain, Levine and Feizi (2020) proposed a random ablation method for defending against sparse adversarial attacks. The approach involves randomly removing pixels from an image before classification to assess their influence on the output. Jia et al. (2022) extended this idea to the top-k prediction setting. Inspired by these methods in image-based tasks, Zeng et al. (2021) transferred the pixel deletion strategy to text classification by randomly replacing words with `[MASK]` tokens. However, their approach was developed for small-scale models and lacks scalability to LLMs. Ji et al. (2024) introduced LLMs to fill in masked tokens for semantic denoising. Nonetheless, their method suffers from two major issues: first, the substitutions are not verified, which may introduce semantic drift; second, invoking LLMs for large-scale denoising leads to considerable computational cost.

Our work aims to strike a balance between the loose robustness bounds typically associated with standard models and the high computational overhead of the denoising process required for smoothed models. By incorporating a semantic clustering filter and an efficient perturbation generation strategy, we improve robustness certification while reducing the overall computational burden.

3 Background

3.1 Notation

We consider the standard setting of text classification, where the input is a sentence represented as a sequence of discrete tokens $w = (w_1, w_2, \ldots, w_n) \in \mathcal{W}$. Each token w_i is drawn from a finite vocabulary $\mathcal{S}$, and n denotes the sentence length. The goal is to assign an input sentence w to one

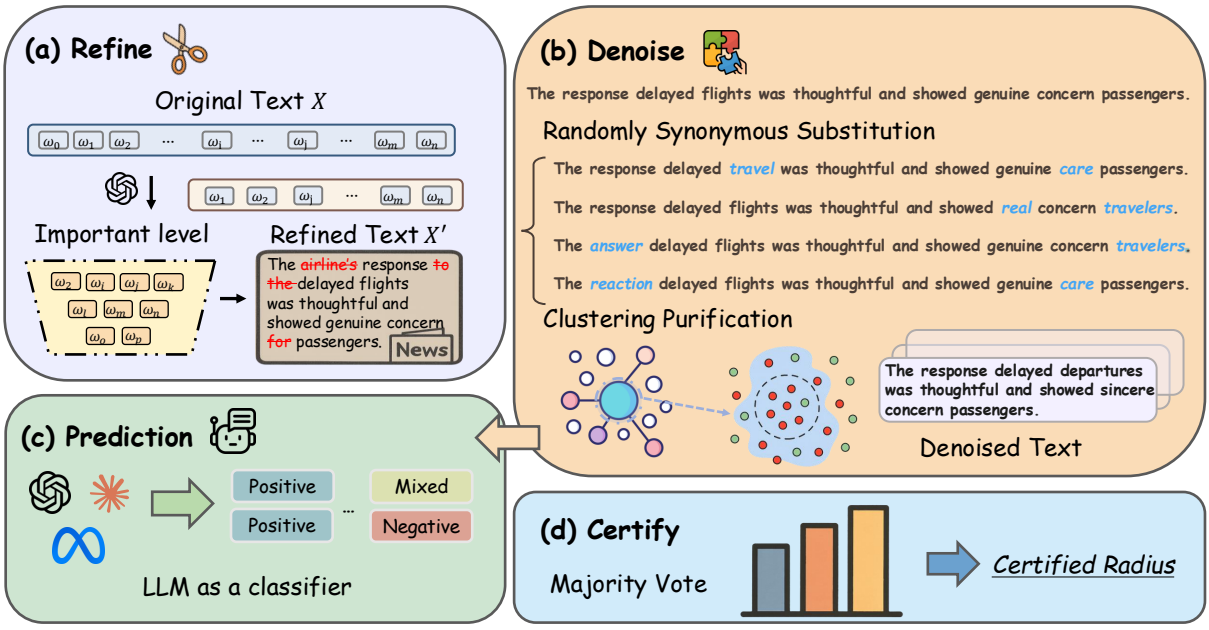

Figure 2: Overview of our certified robustness framework **CluCERT**. (a) **Refine** removes irrelevant tokens using LLM generation to improve efficiency. (b) **Denoise** generates adversarial variants via synonym substitution and applies semantic clustering for purification. (c) **Predict** uses different LLMs for classification and aggregates outputs via majority vote. (d) **Certify** computes the certified radius based on the voting outcome.

of the target classes in a finite label set $\mathcal{Y}$. A *soft classifier* is defined as a mapping $F : \mathcal{W} \to \Delta^{|\mathcal{Y}|}$, which outputs a probability distribution over labels, where $\Delta^{|\mathcal{Y}|}$ is the $|\mathcal{Y}|$-dimensional probability simplex. The corresponding *hard classifier* is given by $f(w) := \arg\max_{c \in \mathcal{Y}} F(w)_c$, returning the most likely label.

3.2 Robustness Certification

Certified robustness provides formal guarantees on a model's prediction stability under bounded perturbations, independent of specific attack strategies. Formally, let $f : \mathcal{W} \to \mathcal{Y}$ be a text classifier, and let $w \in \mathcal{W}$ be a clean input sentence. A certification algorithm computes a certified radius $d \in \mathbb{N}$ such that for any perturbed sentence $w' \in \mathcal{W}$ satisfying $\|w - w'\|_0 \leq d$, the prediction remains unchanged, i.e., $f(w') = f(w)$. Here, $\|w - w'\|_0 = \sum_{i=1}^{n} \mathbb{I}[w_i \neq w'_i]$ denotes the Hamming distance, corresponding to the number of modified tokens.

This form of certification guarantees that the classifier remains robust against all adversarial inputs within an ℓ_0 ball of radius r, regardless of how the perturbations are constructed. The certified radius serves as a conservative estimate of the model's robustness; empirical studies (Zhang et al. 2023) have shown that robustness under the same perturbation budget is typically no worse and often higher.

3.3 Randomized Smoothing on Text

Let $f : \mathcal{W} \to \mathcal{Y}$ be a base classifier that maps an input sentence $w \in \mathcal{W}$ to a class label in $\mathcal{Y}$. To improve its robustness, we construct a *smoothed classifier* g by averaging the predictions of f over randomly perturbed input.

Each perturbation is generated by randomly masking some words in w and replacing them with semantically similar alternatives. Let $T(w)$ denote a randomly perturbed version of w. The smoothed classifier is defined as:$g(w) := \arg\max_{c \in \mathcal{Y}} \mathbb{P}[f(T(w)) = c]$, where $\mathbb{P}[f(T(w)) = c]$ represents the probability that the perturbed input is classified as label c. For any class $c \in \mathcal{Y}$, we define its smoothed probability as: $p_c(w) := \mathbb{P}_{\tilde{w} \sim \mathcal{D}(w)}[f(\tilde{w}) = c]$, where $\mathcal{D}(w)$ reflects the randomness in the perturbation process.

By ensembling predictions over a distribution of semantically plausible inputs, randomized smoothing effectively transforms the base model into a robust classifier that can tolerate sparse word-level perturbations, making it suitable for formal certification.

4 Methodology

In this section, we present our theoretical contributions for certified robustness in text classification. Specifically, we begin by formally defining the two key operations in our framework, which correspond to the core processes of noise injection and semantic denoising.

Given a perturbation ratio parameter $m \in (0, 1)$, we randomly select $s = \lfloor (1 - m) \cdot n \rfloor$ positions from a sentence of length n as the retention set $\mathcal{T} \subseteq [n]$, and replace the remaining positions with the special symbol `[MASK]`. The mask operation is defined as $\mathcal{M}(w, \mathcal{T}) = (\tilde{w}_1, \ldots, \tilde{w}_n)$, where $\tilde{w}_i = w_i$ if $i \in \mathcal{T}$ and $\tilde{w}_i =$ `[MASK]` if $i \notin \mathcal{T}$. For example, let the input sentence be $w =$ (`A`, `B`, `C`, `D`, `E`), and let $m = 0.4$, then $s = \lfloor 0.6 \cdot 5 \rfloor = 3$. Randomly selecting the retention set $\mathcal{T} = \{1, 3, 5\}$, the masked sentence becomes $\mathcal{M}(w, \mathcal{T}) =$ (`A`, `[MASK]`, `C`, `[MASK]`, `E`).

Second, we replace [MASK] positions with semantically similar words, such as synonyms, context-predicted words, or embedding-similar words. Let the semantic recovery mapping be $\mathcal{R} : \mathcal{W}_{\texttt{mask}} \to \mathcal{W}$, then the final perturbed sample is $T(w, \mathcal{T}) := \mathcal{R}(\mathcal{M}(w, \mathcal{T}))$. Continuing from the previous case, suppose the [MASK] positions are replaced with semantically similar B′ and D′ respectively, then $T(w, \mathcal{T}) =$ (A, B′, C, D′, E), where B′ and D′ are substitutes for B and D, and the overall semantics remain consistent.

Thus, let the base classifier be $f : \mathcal{W} \to \mathcal{Y}$. We construct a smoothed classifier by aggregating predictions over randomly perturbed versions of the input. Formally, the smoothed classifier is defined as:

$$g(w) := \arg\max_{y \in \mathcal{Y}} \mathbb{P}_{\mathcal{T} \sim \mathcal{U}_{n,s}}[f(T(w, \mathcal{T})) = y] \tag{1}$$

where $\mathcal{U}_{n,s}$ denotes the uniform distribution over all $\binom{n}{s}$ possible selections of s positions from n positions.

For any $c \in \mathcal{Y}$, the smoothed probability is defined as:

$$p_c(w) := \mathbb{P}_{\mathcal{T} \sim \mathcal{U}_{n,s}}[f(T(w, \mathcal{T})) = c] \tag{2}$$

4.1 Building a Smoothed LLM

To build a smoothed classifier for textual tasks, and inspired by the insights from (Jia et al. 2022) and (Zeng et al. 2021), we propose a novel certification bound that differs from existing methods by incorporating both the sampling shift term Δ_t and a semantic recovery stability factor γ, enabling robustness certification under textual perturbations.

Theorem 1 ((Levine and Feizi 2020))**.** *For any $w' \in \mathcal{W}$ satisfying $\|w - w'\|_0 \leq d$, the smoothed probability for any class $c \in \mathcal{Y}$ satisfies*

$$|p_c(w) - p_c(w')| \leq \gamma \cdot \Delta_t, \tag{3}$$

where

$$\Delta_t := 1 - \frac{\binom{n-d}{s}}{\binom{n}{s}}, \tag{4}$$

and

$$\gamma \in [0, 1]. \tag{5}$$

Here, d denotes the maximum number of perturbed words, i.e., the ℓ_0-distance between w and w'. $\Delta_t = 1 - \binom{n-d}{s}/\binom{n}{s}$ quantifies the sampling distribution shift induced by ℓ_0 perturbations, while $\gamma \in [0, 1]$ captures the semantic recovery stability. A complete proof is provided in **Appendix A**.

In practice, we approximate the exact smoothed probability $p_c(w)$ by averaging the model's predictions over a finite number of randomly sampled mask patterns, since enumerating all possible masks is computationally infeasible. Specifically, we sample N mask sets $\mathcal{T}_1, \ldots, \mathcal{T}_N \sim \mathcal{U}_{n,s}$ and estimate: $\hat{p}_c(w) = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}[f(T(w, \mathcal{T}_i)) = c]$.

To provide rigorous guarantees despite finite sampling, we construct confidence intervals using the Clopper-Pearson method (Clopper and Pearson 1934). Let $\underline{p}_c(w)$ and $\overline{p}_c(w)$ denote the lower and upper bounds of the $(1-\alpha)$ confidence interval for $p_c(w)$. Our certification procedure (as shown in Algorithm 1) leverages these conservative estimates to ensure probabilistic soundness.

Corollary 1. *Let $c = \arg\max_y \underline{p}_y(w)$ be the class with highest lower confidence bound. If for all $j \neq c$:*

$$\underline{p}_c(w) - \overline{p}_j(w) > 2\gamma \cdot \Delta_t \tag{6}$$

then for all $w' \in \mathcal{B}_t(w) := \{w' : \|w - w'\|_0 \leq t\}$, we have $g(w') = c$ with probability at least $1 - \alpha$.

Proof. For any $w' \in \mathcal{B}_t(w)$ and $j \neq c$, the probability drift bound gives $p_c(w') \geq p_c(w) - \gamma\Delta_t$ and $p_j(w') \leq p_j(w) + \gamma\Delta_t$. Thus,

$$\begin{aligned} p_c(w') - p_j(w') &\geq (p_c(w) - \gamma\Delta_t) - (p_j(w) + \gamma\Delta_t) \\ &= p_c(w) - p_j(w) - 2\gamma\Delta_t > 0 \end{aligned}$$

which implies $c = \arg\max_y p_y(w')$, i.e., $g(w') = c$. □

This bound extends the classical randomized smoothing certification by explicitly incorporating both the sampling shift and the uncertainty in semantic recovery stability. In binary classification tasks, the condition simplifies to $p_c(w) - \gamma\Delta_t > 0.5$, allowing robustness certification for text classification under a confidence level of $(1 - \alpha)$.

Furthermore, to enhance the efficiency of our smoothing framework and ensure consistent processing across all modules, we introduce a *Refine* operation that selects the top-L most informative tokens from a sentence based on LLM-derived importance scores.

Definition 1 (Refine Operation)**.** *Given an input sentence $w = (w_1, w_2, \ldots, w_n)$, where w_i denotes the i-th token, we define the refine function $R : \mathcal{W} \to \mathcal{W}_L$ as*

$$R(w) := \{w_i \in w : \textit{rank}_\sigma(w_i) \leq L\}.$$

Here, $\sigma : \mathcal{W} \to \mathbb{R}^n$ is an importance scoring function computed by a language model that assigns each token w_i a real-valued score $\sigma(w_i)$. The operator $\textit{rank}_\sigma(w_i)$ denotes the position of w_i when all tokens in w are sorted in descending order of importance. The parameter L is the target output length, ensuring $|R(w)| = L$ for all inputs with $|w| \geq L$. We denote $\mathcal{W}_L := \{w \in \mathcal{W} : |w| = L\}$ as the space of refined sequences of fixed length.

This operation extracts the core semantic content of the input while ensuring a uniform output length across all texts. All downstream modules in our framework operate on these fixed-length refined inputs. In practice, we achieve this by injecting a carefully designed prompt into the LLM to rank input words by importance and prune less informative ones accordingly (see **Appendix B** for details).

4.2 Cluster-guided Denoising

In this section, we formally define the concept of clustering and provide a theoretical proof for how clustering-guided denoising smoothing improves certified robustness.

Let $\mathcal{W}_t(w) = \{w'_1, \ldots, w'_N\}$ denote the perturbation set obtained through mask and semantic filling. We introduce an embedding function $\phi : \mathcal{W} \to \mathbb{R}^m$ that maps each perturbed sample to its semantic representation $z_i = \phi(w'_i)$. By applying a clustering algorithm to partition $\{z_i\}_{i=1}^{N}$

Algorithm 1: Clustering-Guided Denoising Smoothing

procedure Classifier(w, f, s, N)
  Sample $\mathcal{B} = \{w'_1, \ldots, w'_N\}$ where $w'_i = T(w, \mathcal{T}_i)$
  $\tilde{\mathcal{B}} \leftarrow \{w' \in \mathcal{B} : \phi(w') \in \mathcal{C}_{\max}\}$ ▷ *Clustering*
  counts$[c] \leftarrow |\{w' \in \tilde{\mathcal{B}} : f(w') = c\}|$ for all $c \in \mathcal{Y}$
  return counts

procedure Predict(w, f, s, N)
  counts $\leftarrow$ Classifier(w, f, s, N)
  $\hat{c} \leftarrow \arg\max_c$ counts$[c]$ ▷ *Majority vote*
  $\tilde{p}_{\hat{c}} \leftarrow$ counts$[\hat{c}] / \sum_c$ counts$[c]$
  return $\hat{c}, \tilde{p}_{\hat{c}}$

procedure Certify($w, y, f, s, N, N', \gamma, \alpha$)
  $\hat{c}, \tilde{p}_{\hat{c}} \leftarrow$ Predict(w, f, s, N)
  Let n_A, n_B be counts of the top-2 classes
  if BinomPValue$(n_A, n_A + n_B, 0.5) > \alpha$
    return ABSTAIN
  counts $\leftarrow$ Classifier(w, f, s, N')
  Compute bounds $\{\underline{p}_c, \overline{p}_c\}_{c \in \mathcal{Y}}$ from counts
  $r^* \leftarrow \max\{t : \underline{p}_y - \overline{p}_j > 2\gamma\Delta_t, \forall j \neq y\}$
  return r^*

into K clusters $\mathcal{C}_1, \ldots, \mathcal{C}_K$, we identify the largest cluster index $k_{\max} := \arg\max_k |\mathcal{C}_k|$ and denote its corresponding cluster as $\mathcal{C}_{k_{\max}}$. This induces a refined perturbation subspace $\tilde{\mathcal{W}}_t(w) := \{w'_i \mid \phi(w'_i) \in \mathcal{C}_{k_{\max}}\}$, over which we re-estimate the smoothed probabilities as $\tilde{p}_c := \frac{|\{w' \in \tilde{\mathcal{W}}_t(w) : f(w') = c\}|}{|\tilde{\mathcal{W}}_t(w)|}$. We select the largest cluster as the primary semantic group, as semantically consistent perturbations tend to concentrate within it, whereas adversarial or noisy samples are typically dispersed among smaller clusters.

Theorem 2. *Let*

$$r^* = \max\left\{t \in \mathbb{N} : p_c(w) - \max_{j \neq c} p_j(w) > 2\gamma\Delta_t\right\} \quad (7)$$

be the original certified radius. Suppose semantic clustering induces probability shifts such that there exists some $\epsilon > 0$ satisfying:

$$\tilde{p}_c(w) \geq p_c(w) + \epsilon,$$
$$\tilde{p}_j(w) \leq p_j(w) - \epsilon \quad \textit{for all } j \neq c.$$

Then the certified radius with clustering $\tilde{r}^*$ satisfies $\tilde{r}^* > r^*$

The complete proof is provided in **Appendix C**. We also provide the detailed clustering algorithm implementation in **Appendix D**. This theorem shows that denoising can provably improve the certification bound. By leveraging clustering to focus on semantically coherent perturbations, we reduce the influence of low-consistency perturbations and shift the prediction distribution towards the correct class, thus enabling tighter robustness guarantees.

Lemma 1. *If the embedding mapping ϕ and classifier f satisfy the Lipschitz condition:*

$$\|\phi(w'_i) - \phi(w'_j)\| \leq \rho \Rightarrow \mathbb{P}[f(w'_i) = f(w'_j)] \geq 1 - L\rho$$

and the clustering diameter satisfies diam$(\phi(\tilde{\mathcal{W}}_t(w))) \leq r$, then:

$$\textit{Var}_{w' \in \tilde{\mathcal{W}}_t(w)}[\mathbb{I}[f(w') = c]] \leq Lr \quad (8)$$

Proof. Let $\mu_c = \mathbb{P}_{w' \in \tilde{\mathcal{W}}_t(w)}[f(w') = c]$ denote the class probability within the refined cluster. Since the semantic diameter of the cluster is at most r, for any two samples $w'_i, w'_j \in \tilde{\mathcal{W}}_t(w)$, we have $\|\phi(w'_i) - \phi(w'_j)\| \leq r$, and thus by the Lipschitz condition,

$$\mathbb{P}[f(w'_i) = f(w'_j)] \geq 1 - Lr.$$

Now fix an arbitrary reference point $w_0 \in \tilde{\mathcal{W}}_t(w)$. If $f(w_0) = c$, then for all $w' \in \tilde{\mathcal{W}}_t(w)$ we have $\mathbb{P}[f(w') = c] \geq 1 - Lr$, implying $\mu_c \geq 1 - Lr$. Conversely, if $f(w_0) \neq c$, then $\mathbb{P}[f(w') \neq c] \geq 1 - Lr$, so $\mu_c \leq Lr$.

In either case, the variance of the binary variable $\mathbb{I}[f(w') = c]$ satisfies

$$\text{Var}[\mathbb{I}[f(w') = c]] = \mu_c(1 - \mu_c) \leq Lr.$$

□

Building on Lemma 1, we observe that semantic clustering not only increases the mean prediction probability for the majority class but also reduces its variance. When perturbed samples are tightly grouped in the embedding space, the Lipschitz continuity of the classifier implies more stable predictions. This stability leads to lower variance and, consequently, enables tighter and more reliable certified robustness bounds.

In conclusion, we interpret cluster-guided denoising as projecting the perturbation set $\mathcal{W}_t(w)$ onto a semantically consistent subspace $\tilde{\mathcal{W}}_t(w) = \Pi_{\mathcal{S}_{\text{sem}}}(\mathcal{W}_t(w))$. The target subspace $\mathcal{S}_{\text{sem}}$ is characterized by: (i) tight clustering with $\max_{w', w'' \in \mathcal{S}_{\text{sem}}} |\phi(w') - \phi(w'')| \leq r$, and (ii) high prediction consistency where $\mathbb{P}[f(w') = f(w'')] \geq 1 - Lr$ for any $w', w'' \in \mathcal{S}_{\text{sem}}$. This projection acts as semantic denoising, improving model stability and enlarging certified robustness regions without modifying the base model.

5 Experiments

5.1 Experimental Setup

Overview We consider two major task settings to evaluate our proposed framework CluCERT: ***(a)*** certified robustness of a given LLM under word substitutions, and ***(b)*** empirical robustness under attacks.

Datasets and models To demonstrate both the certified radius enhancement our framework (settings ***a***), we follow prior work (Ji et al. 2024) and conduct experiments on two widely used sentiment classification datasets: SST-2 (Socher et al. 2013) and AGNews (Zhang, Zhao, and LeCun 2015). We additionally include the GSM8K dataset (Cobbe et al. 2021) to assess robustness in mathematical problem-solving. For these evaluations, we use ChatGPT-3.5 [1] as base models. To assess empirical robustness under adversarial attacks

[1] https://platform.openai.com

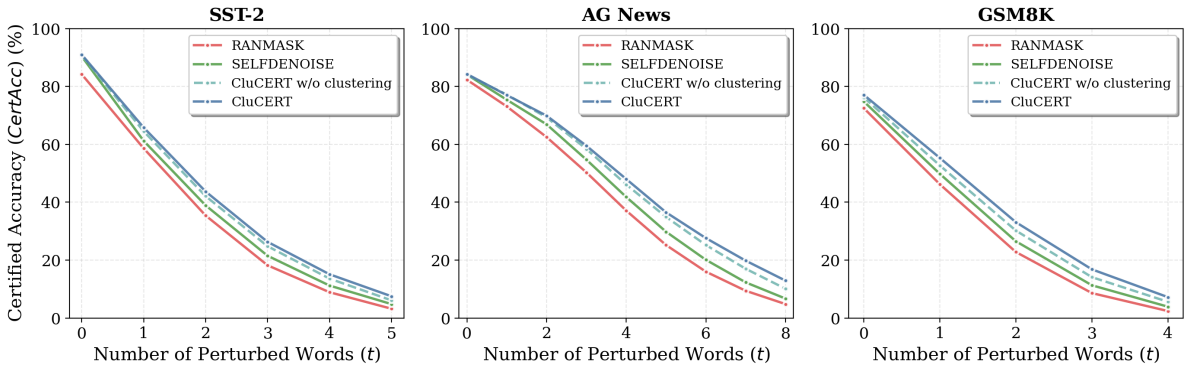

Figure 3: Certified accuracy on SST-2, AG News, and GSM8K under different numbers of perturbed words.

(setting ***b***), we adopt Textbugger (Li et al. 2019) and DeepWordBug (Gao et al. 2018), using the open-source LLM Vicuna [2] for evaluation.

Implementation details and metrics Following prior work (Zeng et al. 2021), we randomly selected 2,000 examples from each dataset for both certified and empirical robustness experiments. For each input instance, we generated 1,000 perturbed samples using word substitutions. All experiments are conducted under a consistent prompt interface to ensure a fair comparison across defense strategies.

To enable efficient local synonym substitution, we construct a hybrid candidate pool centered on WordNet [3], augmented with embedding-based nearest neighbors and domain-specific lexicons. After generating perturbed sentences, we compute sentence-level semantic similarity between each perturbed sentence and the original input using a pre-trained BERT model. We employ a semantic similarity threshold τ to filter perturbations. Here, τ is task-specific and determined through validation. The confidence level for certification is set to $\alpha = 0.05$.

For certified robustness evaluation, our primary metrics include the average certified radius r_{avg} and the certified accuracy under varying perturbation levels. Certified accuracy at level δ is defined as $\text{CertAcc}(\delta) = \frac{1}{N}\sum_{j=1}^{N}\mathbb{I}(r_j \geq \delta)$, where r_j is the certified radius of the j-th input. For empirical robustness evaluation, we use the attack success rate (ASR) against multiple adversarial attack strategies. Additionally, to assess the efficiency, we use the execution time t and the token cost c as metrics, which reflect the computational and monetary cost when querying LLMs.

Baselines We compare with two baselines: RanMASK (Zeng et al. 2021), a randomized smoothing method without denoising, and SelfDenoise (Ji et al. 2024), which applies denoising but lacks verification. These highlight the effects of our denoising and clustering modules, respectively. For consistency, both baselines are evaluated with the same *Refine* module as our method (denoted with an asterisk *).

[2]https://huggingface.co/lmsys/vicuna-7b-v1.5

[3]https://wordnet.princeton.edu/

5.2 Certified Robustness Evaluation

While prior methods rely on white-box access and often require fine-tuning or retraining, CluCERT instead operates in a black-box setting, interacting with the model solely through API queries. This design makes it applicable to closed-source large language models and more reflective of real-world deployment. Specifically, we use ChatGPT-3.5 as the base model and evaluate its certifiable robustness without accessing model weights. In contrast to vision tasks (Cohen, Rosenfeld, and Kolter 2019; Carlini et al. 2022) that typically assume continuous perturbations, textual perturbations are inherently discrete, such as word substitutions. Moreover, certified radii in text are integer-valued, indicating the maximum number of word substitutions that preserve the model's prediction. To capture robustness under this discrete setting, we report the average certified radius (r_{avg}), along with the coefficient of variation (Coe), to characterize both the level and stability of model robustness across different perturbation budgets.

Method	SST-2		AGNews		GSM8K	
	r_{avg}	Coe	r_{avg}	Coe	r_{avg}	Coe
RanMASK	1.24	1.12	2.78	0.88	0.80	1.35
SelfDenoise	1.38	1.07	3.08	0.82	0.92	1.25
CluCERT(-Clu)	1.51	1.02	3.38	0.79	1.05	1.15
CluCERT	**1.58**	**1.00**	**3.51**	**0.79**	**1.16**	**1.07**

Table 1: Robustness certificates across datasets. CluCERT(-Clu) refers to our framework without clustering. r_{avg} denotes the average certified radius (larger indicates stronger robustness), and Coe is the coefficient of variation (smaller indicates higher stability).

Beyond standard text classification tasks, we extend CluCERT to mathematical question answering, a setting that poses greater challenges for robustness due to its deterministic outputs and high sensitivity to input perturbations. As shown in Table 1, CluCERT achieves the highest r_{avg} and the lowest Coe across all evaluated datasets, indicating superior and stable certified robustness. On GSM8K, CluCERT attains an r_{avg} of 1.16, outperforming RanMASK

AGNews	Clean Acc.	ASR (TB)	ASR (DWB)
RanMask	82.3	47.1	42.6
SelfDenoise	84.1	34.4	30.9
CluCERT (w/o Refine)	**85.0**	32.7	28.5
CluCERT	84.3	**29.8**	**26.9**
SST-2	Clean Acc.	ASR (TB)	ASR (DWB)
RanMask	84.2	52.3	47.4
SelfDenoise	90.4	44.6	35.7
CluCERT (w/o Refine)	**91.0**	45.5	33.3
CluCERT	**91.0**	**41.3**	**31.2**

Table 2: Clean accuracy and attack success rate (%) under TextBugger (TB) and DeepWordBug (DWB) attacks on AGNews and SST-2. Lower ASR indicates stronger robustness.

(0.80) and SelfDenoise (0.92), with the smallest Coe of 1.07. This demonstrates its ability to maintain robustness under structurally complex and fragile input conditions. On AGNews, which features longer inputs and greater semantic diversity, CluCERT again outperforms baselines, achieving $r_{\text{avg}} = 3.51$ and $Coe = 0.79$, showing tolerance to stronger perturbations and consistency across examples. SST-2 exhibits lower certified radii, likely due to shorter inputs and the reliance on a few key emotional tokens, yet CluCERT still delivers the best performance, highlighting its adaptability to different task types.

To further understand the effect of the clustering mechanism, we conduct ablation studies focusing on the clustering-based denoising module. As shown in Figure 3, integrating the synonym clustering strategy consistently improves certified accuracy across various perturbation levels, with particularly notable gains on long-text datasets such as AGNews. These results suggest that clustering mitigates semantic drift introduced by low-quality substitutions and enhances prediction stability. Compared to the variant without clustering, CluCERT yields both higher certified accuracy and smoother robustness curves throughout the entire perturbation range. Importantly, these empirical observations align with our theoretical analysis, confirming the effectiveness of clustering in balancing semantic preservation with perturbation diversity.

5.3 Empirical Robustness under Attacks

We evaluate CluCERT under two representative black-box attacks, TextBugger and DeepWordBug, using the default settings of the TextAttack toolkit (Morris et al. 2020). Experiments are conducted on two widely used classification benchmarks, SST-2 and AGNews. RanMASK and SelfDenoise serve as baselines, and we additionally include a variant of CluCERT without the proposed *Refine* module. Table 2 reports clean accuracy and attack success rate (ASR), where lower ASR indicates stronger robustness.

CluCERT consistently achieves the lowest attack success rates (ASR) across both datasets and attack types, demonstrating robust and stable empirical performance. On SST-2, it maintains the highest clean accuracy (91.0%) while achieving the lowest ASR, clearly outperforming all baseline methods. On AGNews, CluCERT also exhibits the strongest robustness under adversarial attacks. Although its clean accuracy slightly decreases compared to the variant without the *Refine* module, this trade-off results in a substantial gain in robustness. We attribute this improvement primarily to the *Refine* module. By compressing the input, filtering peripheral content, and emphasizing core semantics, this module helps the model focus on key information, thereby enhancing robustness and prediction stability. While the process may occasionally remove marginally informative tokens or introduce output variance induced by prompt design, it generally reduces input length and minimizes noisy information. This leads to improved resistance to perturbations and increased inference efficiency.

The integration of clustering-based substitution and the *Refine* module enables CluCERT to strike a balance between clean accuracy and adversarial robustness. The consistent drop in ASR across datasets and attack types confirms the effectiveness and generalizability of our approach.

5.4 Efficiency

In this section, we analyze the efficiency of the proposed method. As shown in Figure 4, we present the estimated time cost of each processing stage when generating 1000 perturbed samples for a single input. Our method achieves approximately 6.8× speedup compared to the baseline method, significantly reducing the computational overhead for deployment. In addition, the LLM-based substitution stage incurs non-negligible token-level inference costs, which may lead to substantial economic burdens in large-scale text processing scenarios.

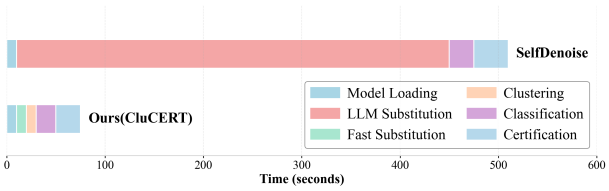

Figure 4: Time cost for each mode from SelfDenoise and CluCERT

The speedup primarily stems from our fast synonym substitution strategy, which avoids costly LLM-based generation used in prior methods such as SelfDenoise. Overall, our approach achieves a favorable balance between robustness and computational efficiency.

6 Conclusion

We propose CluCERT, a clustering-guided and efficient smoothing framework for certifying the robustness of large language models. By combining fast synonym-based perturbation with semantic clustering, CluCERT focuses on meaningful substitutions while improving sampling efficiency, enabling stronger certified bounds with significantly lower computational cost. Extensive experiments on text classification and mathematical reasoning tasks demonstrate its effectiveness and generalizability.

Acknowledgements

This work was supported by The Royal Society Grant (Ensuring Trustworthy AI: Robustness Certification for Large Language Models)[Reference RGS\R2\252444]. GJ's contribution is supported by the NVIDIA Academic Grant Program. The authors would like to acknowledge the use of the University of Exeter High-Performance Computing (HPC) facility in carrying out this work.

References

Carlini, N.; Tramer, F.; Dvijotham, K. D.; Rice, L.; Sun, M.; and Kolter, J. Z. 2022. (certified!!) Adversarial robustness for free! *arXiv preprint arXiv:2206.10550*.

Chao, P.; Robey, A.; Dobriban, E.; Hassani, H.; Pappas, G. J.; and Wong, E. 2025. Jailbreaking black box large language models in twenty queries. In *2025 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, 23–42. IEEE.

Chen, T.; Liu, S.; Chang, S.; Cheng, Y.; Amini, L.; and Wang, Z. 2020. Adversarial robustness: From self-supervised pre-training to fine-tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 699–708.

Chen, Y.; Hu, Z.; Zhi, C.; Han, J.; Deng, S.; and Yin, J. 2024. Chatunitest: A framework for llm-based test generation. In *Companion Proceedings of the 32nd ACM International Conference on the Foundations of Software Engineering*.

Cheng, Y.; Jiang, L.; Macherey, W.; and Eisenstein, J. 2020. Advaug: Robust adversarial augmentation for neural machine translation. *arXiv preprint arXiv:2006.11834*.

Clopper, C. J.; and Pearson, E. S. 1934. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*.

Cobbe, K.; Kosaraju, V.; Bavarian, M.; Chen, M.; Jun, H.; Kaiser, L.; Plappert, M.; Tworek, J.; Hilton, J.; Nakano, R.; Hesse, C.; and Schulman, J. 2021. Training Verifiers to Solve Math Word Problems. *arXiv preprint arXiv:2110.14168*.

Cohen, J. M.; Rosenfeld, E.; and Kolter, J. Z. 2019. Certified Adversarial Robustness via Randomized Smoothing. arXiv:1902.02918.

Fang, H.; Kong, J.; Zhuang, T.; Qiu, Y.; Gao, K.; Chen, B.; Xia, S.-T.; Wang, Y.; and Zhang, M. 2025. Your Language Model Can Secretly Write Like Humans: Contrastive Paraphrase Attacks on LLM-Generated Text Detectors. arXiv:2505.15337.

Gao, J.; Lanchantin, J.; Soffa, M. L.; and Qi, Y. 2018. Black-box Generation of Adversarial Text Sequences to Evade Deep Learning Classifiers. arXiv:1801.04354.

Huang, X.; Ruan, W.; Huang, W.; Jin, G.; Dong, Y.; Wu, C.; Bensalem, S.; Mu, R.; Qi, Y.; Zhao, X.; et al. 2024. A survey of safety and trustworthiness of large language models through the lens of verification and validation. *Artificial Intelligence Review*.

Ji, J.; Hou, B.; Zhang, Z.; Zhang, G.; Fan, W.; Li, Q.; Zhang, Y.; Liu, G.; Liu, S.; and Chang, S. 2024. Advancing the Robustness of Large Language Models through Self-Denoised Smoothing. arXiv:2404.12274.

Jia, J.; Wang, B.; Cao, X.; Liu, H.; and Gong, N. Z. 2022. Almost Tight L0-norm Certified Robustness of Top-k Predictions against Adversarial Perturbations. arXiv:2011.07633.

Jia, R.; Raghunathan, A.; Göksel, K.; and Liang, P. 2019. Certified robustness to adversarial word substitutions. *arXiv preprint arXiv:1909.00986*.

Jin, D.; Jin, Z.; Zhou, J. T.; and Szolovits, P. 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*.

Jin, G.; Yi, X.; Huang, W.; Schewe, S.; and Huang, X. 2025. $S^2$2O: Enhancing Adversarial Training With Second-Order Statistics of Weights. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(10): 8630–8641.

Jin, G.; Yi, X.; and Huang, X. 2025. Reconcile Certified Robustness and Accuracy for DNN-based Smoothed Majority Vote Classifier. *arXiv preprint arXiv:2509.25979*.

Kumar, A.; Agarwal, C.; Srinivas, S.; Li, A. J.; Feizi, S.; and Lakkaraju, H. 2023. Certifying llm safety against adversarial prompting. *arXiv preprint arXiv:2309.02705*.

Levine, A.; and Feizi, S. 2020. Robustness certificates for sparse adversarial attacks by randomized ablation. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Li, J.; Ji, S.; Du, T.; Li, B.; and Wang, T. 2019. TextBugger: Generating Adversarial Text Against Real-world Applications. In *Proceedings 2019 Network and Distributed System Security Symposium*, NDSS 2019. Internet Society.

Li, X.; Wang, R.; Cheng, M.; Zhou, T.; and Hsieh, C.-J. 2024. Drattack: Prompt decomposition and reconstruction makes powerful llm jailbreakers. *arXiv preprint arXiv:2402.16914*.

Lin, X.; Huang, Z.; Zhang, Z.; Zhou, J.; and Chen, E. 2025. Explore What LLM Does Not Know in Complex Question Answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Morris, J.; Lifland, E.; Yoo, J. Y.; Grigsby, J.; Jin, D.; and Qi, Y. 2020. TextAttack: A Framework for Adversarial Attacks, Data Augmentation, and Adversarial Training in NLP. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 119–126.

Mu, R.; Marcolino, L.; Ni, Q.; and Ruan, W. 2024. Enhancing robustness in video recognition models: Sparse adversarial attacks and beyond. *Neural Networks*.

Raghunathan, A.; Steinhardt, J.; and Liang, P. 2018. Certified defenses against adversarial examples. *arXiv preprint arXiv:1801.09344*.

Rocamora, E. A.; Chrysos, G. G.; and Cevher, V. 2025. Certified Robustness Under Bounded Levenshtein Distance. *arXiv preprint arXiv:2501.13676*.

Salman, H.; Sun, M.; Yang, G.; Kapoor, A.; and Kolter, J. Z. 2020. Denoised smoothing: A provable defense for pretrained classifiers. *Advances in Neural Information Processing Systems*, 33: 21945–21957.

Shayegani, E.; Mamun, M. A. A.; Fu, Y.; Zaree, P.; Dong, Y.; and Abu-Ghazaleh, N. 2023. Survey of vulnerabilities in large language models revealed by adversarial attacks. *arXiv preprint arXiv:2310.10844*.

Socher, R.; Perelygin, A.; Wu, J.; Chuang, J.; Manning, C. D.; Ng, A. Y.; and Potts, C. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, 1631–1642.

Sun, S.; and Ruan, W. 2023. TextVerifier: Robustness Verification for Textual Classifiers with Certifiable Guarantees. In Rogers, A.; Boyd-Graber, J.; and Okazaki, N., eds., *Findings of the Association for Computational Linguistics: ACL 2023*, 4362–4380. Toronto, Canada: Association for Computational Linguistics.

Sun, S.; Sen, P.; and Ruan, W. 2024. CROWD: Certified Robustness via Weight Distribution for Smoothed Classifiers against Backdoor Attack. In Al-Onaizan, Y.; Bansal, M.; and Chen, Y.-N., eds., *Findings of the Association for Computational Linguistics: EMNLP 2024*, 17056–17070. Miami, Florida, USA: Association for Computational Linguistics.

Tao, Y.; Shen, Y.; Zhang, H.; Shen, Y.; Wang, L.; Shi, C.; and Du, S. 2024. Robustness of Large Language Models Against Adversarial Attacks. arXiv:2412.17011.

Wang, W.; Tang, P.; Lou, J.; and Xiong, L. 2021. Certified Robustness to Word Substitution Attack with Differential Privacy. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Wang, Z.; Hu, J.; and Mu, R. 2025. Safety of embodied navigation: A survey. *arXiv preprint arXiv:2508.05855*.

Weng, L.; Zhang, H.; Chen, H.; Song, Z.; Hsieh, C.-J.; Daniel, L.; Boning, D.; and Dhillon, I. 2018. Towards fast computation of certified robustness for relu networks. In *International Conference on Machine Learning*, 5276–5285. PMLR.

Ye, M.; Gong, C.; and Liu, Q. 2020. SAFER: A structure-free approach for certified robustness to adversarial word substitutions. *arXiv preprint arXiv:2005.14424*.

Yi, D.; Mu, R.; Jin, G.; Qi, Y.; Hu, J.; Zhao, X.; Meng, J.; Ruan, W.; and Huang, X. 2024. Position: building guardrails for large language models requires systematic design. In *Forty-first International Conference on Machine Learning*.

Zeng, J.; Zheng, X.; Xu, J.; Li, L.; Yuan, L.; and Huang, X. 2021. Certified Robustness to Text Adversarial Attacks by Randomized [MASK]. arXiv:2105.03743.

Zhang, J.; Chen, Z.; Zhang, H.; Xiao, C.; and Li, B. 2023. DiffSmooth: Certifiably Robust Learning via Diffusion Models and Local Smoothing. In *32nd USENIX Security Symposium (USENIX Security 23)*. USENIX Association.

Zhang, T.; Zhang, Y.; Mu, R.; Liu, J.; Fieldsend, J.; and Ruan, W. 2024a. PRASS: probabilistic risk-averse robust learning with stochastic search. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, 559–567.

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

Zhang, Y.; Kouvaros, P.; and Lomuscio, A. 2025. Scalable Neural Network Geometric Robustness Validation via Hölder Optimisation. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Zhang, Y.; Wang, M.; Li, Q.; Tiwari, P.; and Qin, J. 2025. Pushing the limit of LLM capacity for text classification. In *Companion Proceedings of the ACM on Web Conference 2025*.

Zhang, Y.; Zhang, T.; Mu, R.; Huang, X.; and Ruan, W. 2024b. Towards fairness-aware adversarial learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 24746–24755.