Guided Perturbation Sensitivity (GPS): Detecting Adversarial Text via Embedding Stability and Word Importance

Bryan E. Tuck, Rakesh M. Verma

University of Houston, Houston, TX 77004 USA
betuck@uh.edu, rmverma2@uh.edu

Abstract

Adversarial text attacks remain a persistent threat to transformer models, yet existing defenses are typically attack-specific or require costly model retraining, leaving a gap for attack-agnostic detection. We introduce Guided Perturbation Sensitivity (GPS), a detection framework that identifies adversarial examples by measuring how embedding representations change when important words are masked. GPS first ranks words using importance heuristics, then measures embedding sensitivity to masking top-k critical words, and processes the resulting patterns with a BiLSTM detector. Experiments show that adversarially perturbed words exhibit disproportionately high masking sensitivity compared to naturally important words. Across three datasets, three attack types, and two victim models, GPS achieves over 85% detection accuracy and demonstrates competitive performance compared to existing state-of-the-art methods, often at lower computational cost. Using Normalized Discounted Cumulative Gain (NDCG) to measure perturbation identification quality, we demonstrate that gradient-based ranking significantly outperforms attention, hybrid, and random selection approaches, with identification quality strongly correlating with detection performance for word-level attacks ($\rho = 0.65$). GPS generalizes to unseen datasets, attacks, and models without retraining, providing a practical solution for adversarial text detection.

Extended version — https://arxiv.org/abs/2508.11667

1 Introduction

A single word substitution can fool a state-of-the-art transformer into classifying a positive movie review as negative, or trick a spam filter into allowing malicious content through. While transformer models achieve remarkable performance on NLP benchmarks, they remain surprisingly brittle to *adversarial examples*—subtle, meaning-preserving perturbations that flip predictions (Goodfellow, Shlens, and Szegedy 2014). As these models are deployed in high-stakes applications from healthcare diagnostics to financial fraud detection, such vulnerabilities pose serious risks where even occasional failures can erode trust, enable manipulation, or trigger costly security incidents (Tuck 2025).

The fundamental challenge in adversarial text detection lies in distinguishing malicious edits from natural language variation. Unlike vision, where perturbations are continuous pixel modifications, text attacks operate in discrete lexical space, requiring manipulations that preserve semantics while remaining imperceptible to human readers (Morris et al. 2020). Word substitutions like changing "excellent" to "great" or character-level edits like "moive" for "movie" can completely flip model predictions while appearing innocuous.

Current detection approaches face a critical limitation: they either assume knowledge of specific attack patterns (Jones et al. 2020; Wang et al. 2021) or require expensive model retraining (Ye, Gong, and Liu 2020; Zeng et al. 2023). Methods that analyze output-layer signals often overfit to particular attacks and fail to generalize (Mosca et al. 2022), while gradient-based detectors overlook the rich sequential structure of adversarial manipulations (Shen et al. 2023). We need a detection approach that exploits the fundamental instability of adversarial examples without requiring attack-specific knowledge.

We build on a crucial theoretical foundation: adversarial examples reside near decision boundaries in regions of high curvature, where small perturbations cause dramatic classification changes (Fawzi et al. 2018; Bell et al. 2024). We hypothesize that this instability extends beyond decision boundaries into the representation space itself. By strategically masking important words, adversarial examples should exhibit disproportionate sensitivity compared to naturally important words in benign text, revealing their artificial nature through instability patterns.

This insight motivates **Guided Perturbation Sensitivity (GPS)**: a detection framework that identifies adversarial examples by measuring how embedding representations change when important words are masked. GPS first ranks words using importance-based methods, then measures embedding sensitivity to masking top-k critical words, and processes these sensitivity patterns with a BiLSTM detector. GPS detects adversarial examples without requiring knowledge of specific attack models or model retraining, generalizing across attacks, datasets, and models.

Our contributions are as follows:

- We introduce **Guided Perturbation Sensitivity (GPS)** (§3), a detection method that identifies adversarial examples by measuring embedding instability under targeted word masking, requiring no modification of the target model.

- We provide empirical evidence that **adversarial examples exhibit approximately 2× higher embedding sensitivity** (§4) to strategic word masking compared to benign inputs, empirically linking decision boundary theory to the representation level.
- Through **comprehensive evaluation across 18 experimental configurations** (§5–§7), we demonstrate that GPS achieves 85%+ detection accuracy across three datasets, three attack types, and two models, with superior generalization and computational efficiency reaching 98% performance at just $K = 5$ words.
- We reveal **fundamental differences in detection mechanisms** (§8, §9) showing gradient-based importance ranking achieves strong correlation ($\rho > 0.65$) between perturbation identification and detection performance for word-level attacks, while character-level attacks require different strategies.

2 Related Work

Adversarial vulnerability. Adversarial machine learning emerged in Huang et al. (2011) and gained prominence in computer vision (Szegedy et al. 2014), attributed to neural network linearity (Goodfellow, Shlens, and Szegedy 2014). This led to optimization-based attacks like Carlini-Wagner (Carlini and Wagner 2017). While these concepts extend to sequential data (Papernot et al. 2016), text requires semantic and grammatical preservation during perturbation.

Adversarial Text Attacks. Text attacks operate at character, word, and sentence level. Character methods include gradient-guided flips (Ebrahimi et al. 2018), importance-based edits (Gao et al. 2018), and Charmer (Rocamora et al. 2024), which achieves high success rates against both BERT and LLMs. Word-level approaches evolved from genetic algorithms (Alzantot et al. 2018) to importance-ranking systems (Jin et al. 2020) and contextual substitutions (Li et al. 2020). Recent advances include GBDA (Guo et al. 2021), optimizing distributions of adversarial examples, and ATGSL (Li et al. 2023), which balances attack effectiveness with text quality using simulated annealing and fine-tuned language models. These attacks are still effective with success rates often exceeding 90% against state-of-the-art transformers while maintaining semantic preservation, making them particularly challenging targets for detection systems (Mehdi Gholampour and Verma 2023).

Defense strategies against adversarial attacks. Existing NLP defenses fall into three camps: adversarial training (Miyato, Dai, and Goodfellow 2017), certified robustness (Jia et al. 2019; Ye, Gong, and Liu 2020; Zhang et al. 2024), and post-hoc detection. Detection approaches range from surface statistics like word frequency (Mozes et al. 2021) and logit irregularities (Mosca et al. 2022) to attribution signals (Alhazmi et al. 2025) and ensemble methods combining multiple gradient-based importance measures like TextShield (Shen et al. 2023). Recent work explores loss landscape geometry: (Zheng et al. 2023) measures sharpness by maximizing local loss increments, while TextDefense (Shen et al. 2025) uses dispersion of word-importance scores to flag suspicious inputs. These detectors either assume attack-specific artifacts, require additional optimization loops, or treat model outputs as static signals; none directly measure how the model's internal representations respond to targeted input modifications. GPS addresses this gap by probing dynamic embedding responses to guided masking, cleanly separating word-level from character-level behaviors and offering a scalable alternative to sharpness and ensemble-based methods.

3 Methodology

Our methodology detects adversarial text by analyzing how targeted word perturbations affect transformer embedding stability. We hypothesize that adversarially manipulated words exhibit unusual importance patterns and cause disproportionate embedding shifts compared to naturally important words. GPS identifies influential words using importance heuristics, measures embedding stability by sequentially masking top-ranked words, and processes the resulting sensitivity patterns with a detector. Our evaluation of gradient, attention, hybrid, and random selection reveals that gradient-based methods outperform attention-based approaches for word-level attacks.

3.1 Reference Embeddings

Given an input text $\mathcal{T} = (w_1, \ldots, w_N)$ (either benign or potentially adversarial) and a frozen transformer model f, we first compute its reference sentence embedding. This is obtained by averaging the final hidden states of its non-special subtokens:

$$\mathbf{e}(\mathcal{T}) = \frac{1}{|\Omega|} \sum_{i \in \Omega} \mathbf{h}_i^{(L)} \in \mathbb{R}^d, \tag{1}$$

where $\mathbf{h}_i^{(L)}$ is the final layer hidden state for the i-th subtoken, $\Omega = \{i \mid \text{subtoken } i \text{ is not a special token}\}$, and d is the embedding dimension. For adversarial detection tasks, we compute reference embeddings $\mathbf{e}_{\text{ben}}$ and $\mathbf{e}_{\text{adv}}$ for the benign and adversarial versions, respectively.

3.2 Identifying Influential Words with Importance Heuristics

To focus our sensitivity analysis on the most relevant words, we first rank all words w_k within the text $\mathcal{T}$ by their predicted importance to the model's decision. We compute an importance score α_k for each word using one of four post-hoc heuristics that require no model modification. The choice of heuristic significantly impacts detection performance, as it determines which words undergo sensitivity testing.

We evaluate four importance ranking strategies to identify the most effective approach for adversarial detection:

Gradient Attribution. Based on the intuition that words critical to the model's prediction exhibit large gradients, we compute importance scores following Simonyan, Vedaldi, and Zisserman (2013). We backpropagate the gradient of the cross-entropy loss $\ell(\mathcal{T})$ with respect to input embeddings $\mathbf{e}_j$ while keeping model f frozen. The importance score for word w_k sums the ℓ_2-norms of gradients across its constituent subtokens $j \in \mathcal{S}_k$:

$$\alpha_k^{\text{sal}} = \sum_{j \in \mathcal{S}_k} \left\| \nabla_{\mathbf{e}_j} \ell(\mathcal{T}) \right\|_2, \tag{2}$$

where $\ell(\mathcal{T})$ is the cross-entropy loss with respect to the predicted class, and $\mathcal{S}_k$ represents subtokens comprising word w_k. We sum over subtokens so that morphologically complex words contribute proportionally to the importance signal. This requires white-box access; surrogate-based saliency can substitute in black-box settings without modifying the detector. Our experiments demonstrate this approach most effectively identifies adversarially perturbed words.

Attention Rollout. To capture information flow through transformer layers, we employ attention rollout (Abnar and Zuidema 2020), which aggregates attention patterns across layers to estimate overall token attention. For each layer ℓ, we compute the head-averaged attention matrix $\mathbf{A}^{(\ell)}_{\text{avg}}$, incorporate residual connections, and row-normalize: $\hat{\mathbf{A}}^{(\ell)} = \text{row-normalize}(0.5 \cdot \mathbf{A}^{(\ell)}_{\text{avg}} + 0.5 \cdot \mathbf{I})$. These matrices are recursively multiplied across layers: $\mathbf{R} = \hat{\mathbf{A}}^{(1)} \times \cdots \times \hat{\mathbf{A}}^{(L)}$. The attention mass flowing to token j is $a_j = \sum_i R_{ij}$, yielding word scores:

$$\alpha^{\text{roll}}_k = \sum_{j \in \mathcal{S}_k} a_j. \tag{3}$$

Grad-SAM. Inspired by Grad-CAM (Selvaraju et al. 2017) for vision, Grad-SAM (Barkan et al. 2021) combines gradient information with attention weights to highlight tokens that are both attended to and important for the prediction. We capture attention weights $\mathbf{A}^{(\ell)}$ and their gradients $\nabla\mathbf{A}^{(\ell)}$ for each layer ℓ (obtained by backpropagating the predicted logit). We compute the element-wise product $\mathbf{G}^{(\ell)} = \nabla\mathbf{A}^{(\ell)} \odot \mathbf{A}^{(\ell)}$ for each layer. We aggregate these layer-wise products by averaging across all L transformer layers, followed by averaging over all H attention heads in the model. The resulting scores are finally summed for the subtokens $j \in \mathcal{S}_k$ corresponding to word w_k:

$$\alpha^{\text{gatt}}_k = \sum_{j \in \mathcal{S}_k} \sum_i \left(\frac{1}{H} \sum_{h=1}^{H} \frac{1}{L} \sum_{\ell=1}^{L} (\mathbf{G}^{(\ell)}_h)_{ij} \right). \tag{4}$$

Random Baseline. As a control, we randomly select K distinct words and assign them importance score $\alpha^{\text{rand}}_k = 1$, with remaining words receiving $\alpha_k = 0$. This baseline helps isolate the contribution of targeted word selection versus random masking.

3.3 Sequential Sensitivity Profiling via Masking

After ranking words by importance $\boldsymbol{\alpha}$, we select the top K most important words, denoted $\mathcal{I}_K = \text{TOP-}K(\boldsymbol{\alpha})$. Through ablation studies across K values ranging from 5 to 50, we find that performance remains relatively stable across this range, with minimal degradation between $K = 5$ and $K = 50$ (§ 7). We select $K = 20$ for all experiments as it provides optimal accuracy while maintaining reasonable computational efficiency.

We then probe embedding stability with respect to these influential words through sequential masking. For each selected word index $k \in \mathcal{I}_K$, we create a masked version by replacing word w_k with the model's [MASK] token, yielding embedding $\tilde{\mathbf{e}}_k = \mathbf{e}(\mathcal{T} \text{ with } w_k \text{ masked})$. The sensitivity s_k quantifies the embedding space change caused by masking, measured as cosine distance between the reference embedding $\mathbf{e}(\mathcal{T})$ and masked embedding $\tilde{\mathbf{e}}_k$:

$$s_k = 1 - \frac{\mathbf{e}(\mathcal{T}) \cdot \tilde{\mathbf{e}}_k}{\|\mathbf{e}(\mathcal{T})\|_2 \, \|\tilde{\mathbf{e}}_k\|_2}. \tag{5}$$

Since we mask words individually, s_k captures the specific impact of each word w_k on the overall representation. We find that adversarially perturbed words exhibit disproportionately high sensitivity values compared to naturally important words, as they represent artificial manipulations that create unstable embedding regions.

3.4 GPS Feature Tensor for Detection

The sensitivity profiling yields a sensitivity score s_k for each word w_k among the top K most important words. We combine these sensitivity scores with the corresponding importance scores α_k while preserving the original word order of the text $\mathcal{T}$. This results in two aligned sequences of length N (the original number of words):

- The sensitivity sequence $\mathbf{s} = (s_1, \ldots, s_N)$, where s_k is the computed sensitivity if $k \in \mathcal{I}_K$, and $s_k = 0$ otherwise.
- The importance sequence $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_N)$, where α_k is the computed importance score if $k \in \mathcal{I}_K$, and $\alpha_k = 0$ otherwise.

We stack these two sequences column-wise to form an $N \times 2$ feature tensor $\mathbf{Z} = [\mathbf{s} \parallel \boldsymbol{\alpha}]$. This tensor retains the original positional information of each word while highlighting the sensitivity and importance of the words identified as most influential by the chosen heuristic. The resulting GPS features $\mathbf{Z}$ can serve as input to any classifier, from simple linear models to neural architectures.

4 Sensitivity Analysis Results

GPS tests whether adversarial examples exhibit measurably different embedding stability than benign text. Table 1 validates this: adversarial examples demonstrate 1.89× higher sensitivity on average, with 88.9% of experiments showing increased instability. Instability ratios cluster tightly across methods (1.836–1.932). Notably, random word selection achieves comparable performance (1.880×) to gradient-based and attention-based methods. This empirically extends established decision boundary instability from classification to

Importance Method	Benign Mean	Adversarial Mean	Ratio
Gradients	0.014	0.028	1.932
Attention Rollout	0.014	0.028	1.912
Grad-SAM	0.014	0.027	1.836
Random	0.013	0.026	1.880

Table 1: Mean sensitivity values s_k across importance methods with K=20. We compute sensitivity s_k as cosine distance between original and masked embeddings (Eq. 5), then take the mean across all masked positions. Columns show averages across experiments (18 per method). Ratio is the mean of per-experiment ratios (adversarial/benign for each experiment). Results are averaged across 3 datasets, 3 attacks, and 2 models.

Component	Options
Datasets	IMDB (Maas et al. 2011) (binary sentiment) AG News (4-way topic) (Zhang, Zhao, and LeCun 2015) Yelp Polarity (binary review) (Zhang, Zhao, and LeCun 2015)
Attacks	TextFooler (Jin et al. 2020) (word substitution) BERT-Attack (Li et al. 2020) (contextual) DeepWordBug (Gao et al. 2018) (char-level)
Models	RoBERTa-base (Liu et al. 2019) DeBERTa-V3-base (He, Gao, and Chen 2021)
Importance	Gradient Attribution (Simonyan, Vedaldi, and Zisserman 2013)
Heuristics	Attention-Rollout (Abnar and Zuidema 2020) Grad-SAM (Barkan et al. 2021) Random selection
Baselines	TextShield (Shen et al. 2023) Sharpness-based detection (Zheng et al. 2023)

Table 2: Experimental matrix. For data, attack, and model configuration, we generate 5,000 balanced training (20% held out for validation) and 1,000 test samples. Our generated adversarial examples exclusively comprise true adversarial samples that successfully deceived the target model; failed perturbation attempts that did not achieve misclassification are excluded from the corpus.

representation space. Embedding instability is an intrinsic property of adversarial examples rather than dependent on any particular importance heuristic, which lets GPS achieve reliable detection.

5 Experimental Setup

We evaluate GPS across the comprehensive experimental matrix shown in Table 2, designed to assess robustness across diverse adversarial scenarios while controlling for architectural, linguistic, and attack-specific variations. Our model selection strategy targets generalization across different transformer architectures: we leverage architectural differences between the models, with DeBERTa's disentangled attention mechanisms separating content and position information and its larger parameter count providing a contrast to standard attention and smaller capacity used in RoBERTa.

For adversarial sample generation, we employ TextAttack (Morris et al. 2020) across most attack methods, with the exception of BERT-Attack.[1] Our importance heuristic evaluation spans gradient-based, attention-based, and hybrid approaches, with random selection serving as a lower-bound control to validate that GPS performance stems from meaningful semantic signal rather than dataset artifacts.

For baselines, we utilize state-of-the-art adversarial detection methods with proven superiority over multiple contemporary approaches. TextShield and sharpness-based detection have demonstrated effectiveness against 8+ established methods, including MD (Lee et al. 2018), DISP (Zhou et al. 2019), FGWS (Mozes et al. 2021), and WDR (Mosca et al. 2022), providing strong baselines for GPS evaluation.

[1]BERT-Attack's search over up to $K = 48$ substitutions per subword can explode combinatorially (e.g., 48^4 candidates for a four-piece token), driving runtimes prohibitively high. We therefore use the TextDefender (Li et al. 2021) implementation, which employs word-level swaps to maintain computational feasibility.

Dataset	Model	Attack	Rand	Attn	GS	Grad	TS	Sharp
AG News	RoBERTa	BA	0.717	0.714	0.788	0.845	**0.846**	0.837
		TF	0.775	0.796	0.843	0.887	**0.893**	0.874
		DWB	0.781	0.772	0.864	**0.895**	0.883	0.860
	DeBERTa	BA	0.729	0.744	0.801	0.839	**0.840**	0.786
		TF	0.798	0.804	0.821	**0.884**	0.883	0.832
		DWB	0.782	**0.902**	0.860	0.897	0.878	0.812
IMDB	RoBERTa	BA	0.741	0.755	0.830	0.845	0.783	**0.873**
		TF	0.846	0.846	0.913	**0.919**	0.870	0.888
		DWB	0.936	0.935	**0.959**	0.958	0.813	0.859
	DeBERTa	BA	0.606	0.621	0.698	0.755	0.731	**0.797**
		TF	0.731	0.757	0.803	**0.859**	0.756	0.800
		DWB	0.929	0.930	0.938	**0.968**	0.775	0.775
Yelp	RoBERTa	BA	0.694	0.714	0.812	0.836	0.832	**0.910**
		TF	0.774	0.783	0.860	0.899	0.849	**0.912**
		DWB	0.781	0.771	0.878	**0.927**	0.874	0.895
	DeBERTa	BA	0.772	0.804	0.844	0.870	0.826	**0.905**
		TF	0.771	0.773	0.865	**0.917**	**0.917**	0.911
		DWB	0.793	0.815	0.856	**0.931**	0.902	0.893

Table 3: Accuracy across datasets, models, attacks, and strategies with K=20. Best accuracy per condition is shown in **bold**. Attacks: BA (BERT-Attack), DWB (DeepWordBug), TF (TextFooler). Our Strategies: Rand (Random), Attn (Attention), GS (Grad-SAM), Grad (Gradient). Baselines: TS (TextShield), Sharp (Sharpness-based).

6 Detection Performance

Having established that adversarial examples exhibit embedding instability, we now demonstrate that this translates into practical detection performance (Table 3). We evaluate GPS using a BiLSTM model to classify sensitivity-importance traces $\mathbf{Z} \in \mathbb{R}^{N\times 2}$ as benign or adversarial. BiLSTM efficiently captures sequential dependencies while handling variable-length texts and maintaining low parameter counts (257,154 parameters); we train using AdamW optimizer (lr=5×10^{-4}), batch size 32, with 10% of training data reserved for validation, and early stopping on validation F1-score with 5-epoch patience over a maximum of 40 epochs. We derive traces from our four importance heuristics with K=20 words and compare against TextShield (Shen et al. 2023), an ensemble of four LSTMs processing gradient-based features,[2] and Sharp (Zheng et al. 2023), a sharpness-based detector measuring local loss landscape curvature.

Gradient-based heuristics consistently outperform attention-based methods, echoing critiques of attention as a direct proxy for predictive importance (Jain and Wallace 2019). Gradient Attribution and Grad-SAM match or exceed state-of-the-art baselines, confirming that embedding instability is most evident when perturbing words critical to model predictions. GPS's superior performance over TextShield's ensemble approach and Sharp's loss landscape

[2]As official code was unavailable, we reimplemented TextShield following the original paper's specifications.

Transfer Setting	GPS (Grad)		TS		Sharp	
	In	Out	In	Out	In	Out
R1: Dataset Shift						
Yelp → IMDB	0.883	**0.878**	0.838	0.806	**0.913**	0.755
IMDB → Yelp	**0.902**	**0.855**	0.822	0.848	0.886	0.765
R2: Attack Shift						
TF → DWB	**0.904**	**0.915**	0.841	0.799	0.829	0.836
DWB → TF	**0.943**	**0.875**	0.831	0.816	0.835	0.829
TF → BA	**0.904**	**0.839**	0.841	0.794	0.829	0.839
BA → TF	**0.844**	**0.875**	0.807	0.855	0.839	0.829
DWB → BA	**0.943**	0.649	0.831	0.729	0.835	**0.839**
BA → DWB	**0.844**	**0.862**	0.807	0.815	0.839	0.836
R3: Model Shift						
RoBERTa → DeBERTa	**0.886**	**0.827**	0.835	0.822	0.835	0.758
DeBERTa → RoBERTa	**0.852**	**0.880**	0.813	0.798	0.841	0.774

Table 4: Generalization performance of adversarial text detection methods across transfer settings with K=20. We evaluate three detection approaches: GPS (Ours, using Grad importance), TS (TextShield), and Sharp (sharpness-based detection). R1 tests cross-dataset generalization, R2 evaluates cross-attack generalization, and R3 examines cross-encoder transferability. In/Out columns show in-domain and out-of-domain F1 scores, respectively. BA (BERT-Attack), DWB (DeepWordBug), TF (TextFooler).

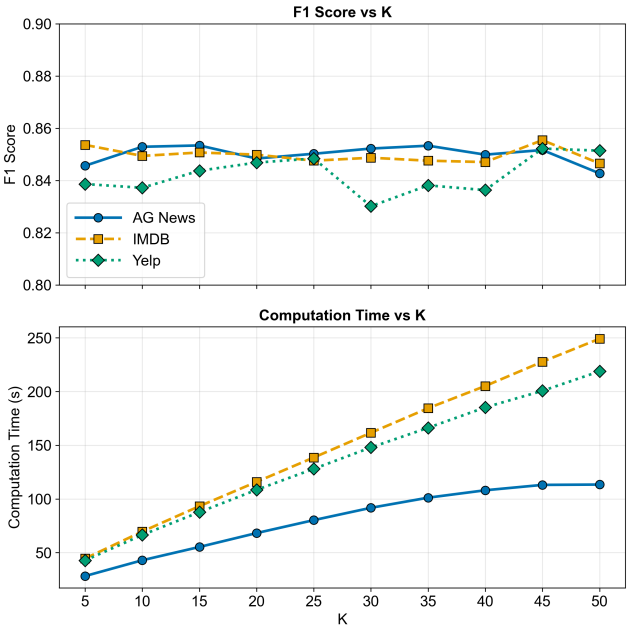

Figure 1: Performance vs efficiency trade-off for GPS across different K values on BERT-Attack adversarial examples using RoBERTa. The annotation box shows baseline computation times for comparison. GPS (28–249s) provides competitive timing with Sharp (51–86s) while significantly outperforming TextShield (111–233s), with the flexibility to trade computation time for detection accuracy via the K parameter.

analysis validates that targeted embedding perturbations provide reliable adversarial detection signals. Contrasting behaviors reveal key insights: gradient-based GPS consistently surpasses attention-based approaches on semantic substitution attacks, while attention-guided GPS remains competitive against character-level DeepWordBug attacks on IMDB. TextShield and Sharp experience significant performance degradation on IMDB's character-level attacks, performing worse than random selection, yet maintain stronger performance on AG News and Yelp. Effective adversarial detection requires aligning detection mechanisms with both specific embedding disruptions and dataset characteristics.

6.1 Generalization Evaluation

Real-world deployment requires detectors that generalize beyond training conditions; if GPS only works on familiar datasets, attacks, or models, its practical utility is severely limited. We evaluate GPS's generalization capabilities across three dimensions using gradient attribution, our best-performing heuristic, and compare against TextShield and Sharp (Table 4). Dataset shifts (R1) involve training on adversarial examples from one dataset (combining TextFooler, DeepWordBug, and BERT-Attack on RoBERTa) and testing on another dataset. Attack shifts (R2) train detectors on one attack type across Yelp and IMDB datasets and test on a different attack type. Model shifts (R3) train on one transformer architecture and test on another, using all datasets and attacks.

GPS shows robust generalization across most transfer scenarios. Embedding sensitivity patterns reflect adversarial manipulation properties rather than dataset-specific artifacts, showing that gradient-based importance captures universal adversarial signatures. Cross-attack generalization shows GPS effectively transfers between word-level attacks (TextFooler, BERT-Attack), though performance drops significantly when transferring from character-level to contextualized semantic substitution attacks (DWB→BA). This reflects model-attack asymmetry: DeepWordBug is the weakest attack (Table 3), and transfer from weaker to stronger attacks is inherently limited, consistent with stronger-to-weaker transfer observed. Cross-architecture results reveal an asymmetry also: positive transfer from DeBERTa to RoBERTa reflects DeBERTa's larger parameter capacity and disentangled attention mechanisms providing richer adversarial representations that generalize to smaller architectures. Sharp's fixed-threshold loss landscape method shows particular vulnerability to dataset and model shifts, maintaining stable performance only where loss sharpness ordering between benign and adversarial samples remains consistent.

7 Computation Trade-Offs

A critical hyperparameter for GPS is determining the optimal number of words K to mask, as too few may miss important adversarial signals, while too many incur unnecessary computational overhead. We evaluate how detection performance varies with K, measuring both F1 scores and computational costs to identify the optimal K that balances detection performance with efficiency (Figure 1).

GPS achieves remarkable efficiency: with K=5 capturing over 98% of the performance observed at K=50, perfor-

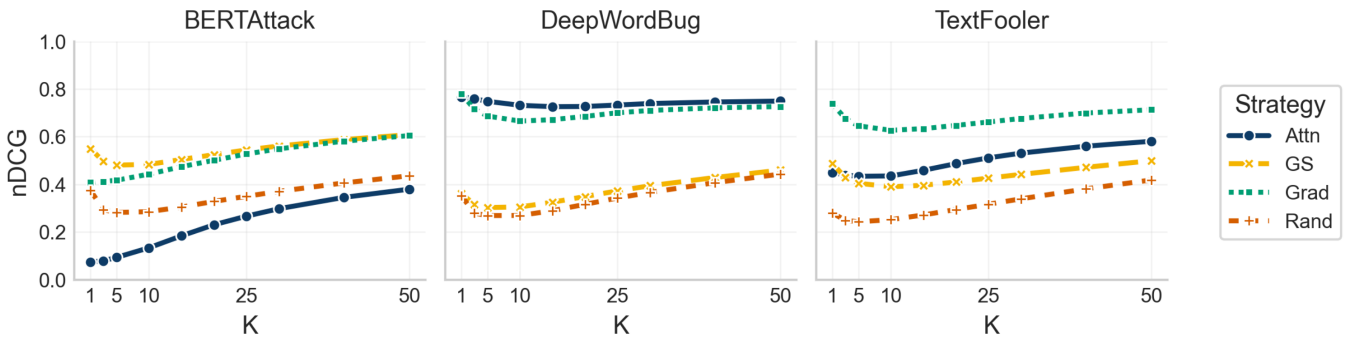

Figure 2: NDCG@k performance for ranking perturbed words on Yelp with RoBERTa across BERT-Attack, DeepWordBug, and TextFooler. Higher NDCG values indicate better ranking quality of truly perturbed words. Strategies: Rand (Random), Attn (Attention), GS (Grad-SAM), Grad (Gradient). Similar patterns hold across other dataset-model combinations.

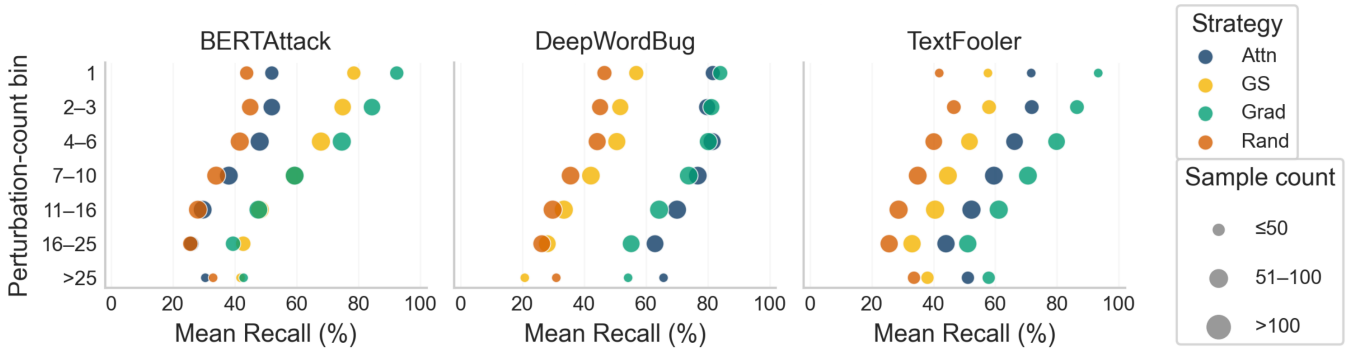

Figure 3: Recall of perturbed words in top-20 rankings by perturbation count bins on Yelp with RoBERTa. Dot size indicates sample count per bin. Higher recall indicates better identification of truly perturbed words. Strategies: Rand (Random), Attn (Attention), GS (Grad-SAM), Grad (Gradient). Similar patterns hold across other dataset-model combinations.

mance variations beyond this point are negligible (< 0.015 F1). Computation time scales linearly with K, with AG News showing earlier saturation due to shorter document lengths. We find that $K \in [5, 10]$ provides an optimal performance-efficiency balance, enabling GPS to operate in resource-constrained environments while maintaining detection quality. The predictable linear scaling allows practitioners to adjust K based on computational budgets without sacrificing reliability, positioning GPS as a practical solution where existing methods may be computationally prohibitive.

8 Ranking Quality Evaluation

Importance heuristics must accurately prioritize words that were actually perturbed during attacks; poor ranking would mean GPS wastes computational resources masking irrelevant words while missing true adversarial modifications. We evaluate each heuristic's effectiveness in ranking perturbed words using Normalized Discounted Cumulative Gain (NDCG) (Wang et al. 2013), which penalizes relevant items appearing lower in the ranked list (Figure 2). This analysis directly tests whether the advantages of gradient-based methods translate to practical perturbation identification.

Using the top-20 candidate words, Gradient Attribution consistently outperforms other heuristics across attack types, showing superior sensitivity to word-level adversarial perturbations. Attention-Rollout performs notably worse against word-level attacks but remains competitive against character-level perturbations. The characteristic spike-dip-recovery pattern in NDCG curves occurs when highly relevant perturbed words appear at top ranks, followed by a drop as irrelevant words enter, then gradual recovery as more perturbed words are found at lower ranks. Gradient-Attribution consistently places perturbed words at the highest ranks. We find these patterns remain consistent across datasets and model variants, with ranking performance mirroring detection results. Effective adversarial detection fundamentally depends on accurate perturbation identification.

8.1 Robustness to Perturbation Density

Real-world adversarial attacks vary significantly in intensity: some make minimal edits to avoid detection, while others heavily modify text to ensure success (Figure 3). Understanding how each heuristic performs across this spectrum is critical for robust detection, as a method that only works on lightly perturbed examples has limited practical value. We sort samples into perturbation-count bins and compute the mean recall of the top-20 candidate words in each bin to quantify how identification quality scales with attack intensity.

Gradient-Attribution exceeds 80% in the sparse 1-6 range

Dataset	ρ	p-value	q-value	n
Global	0.365	0.002	–	72
AG News	**0.903**	<0.001	<**0.001**	24
Yelp	**0.723**	<0.001	<**0.001**	24
IMDB	0.255	0.230	0.230	24

Table 5: Spearman's correlation (ρ) between detection accuracy and NDCG@20. The global correlation across all configurations is moderate, although AG News and Yelp show strong, significant correlations. q-values are Benjamini-Hochberg FDR-corrected (Benjamini and Hochberg 1995) with significant values ($q < 0.05$) in bold.

Attack Type	ρ	p-value	q-value	n
BERT-Attack	**0.655**	<0.001	**0.002**	24
TextFooler	**0.517**	0.010	**0.015**	24
DeepWordBug	-0.103	0.633	0.633	24

Table 6: Spearman's correlation (ρ) between detection accuracy and NDCG@20 by attack type. Word-level attacks show significant positive correlations, while character-level attacks show slight negative correlation, suggesting different detection mechanisms operate for different attack types.

across BERT-Attack, DeepWordBug, and TextFooler. Attention drops sharply once perturbations surpass six words, falling below 40% for the most heavily perturbed samples; random shows similar decline. This performance gap directly accounts for the weaker detection scores reported in Section 6. Identical trends across all three attacks confirm that perturbation density, rather than attack mechanism, drives this failure mode. Gradient-based heuristics maintain higher word-level localization irrespective of perturbation budget, while attention-based methods lose discrimination as adversarial modifications accumulate. This explains why gradient attribution consistently outperforms other approaches across diverse attack scenarios.

9 Relationship Between Perturbation Identification and Detection Performance

A fundamental question in adversarial detection research is whether methods that excel at identifying specific perturbations necessarily translate to superior detection performance. Understanding this relationship is critical for developing principled approaches to adversarial defense and determining when explanation-based evaluation metrics like NDCG truly reflect detector quality. We investigate whether effective perturbation identification directly correlates with detection accuracy across different attack types and datasets.

9.1 Dataset Correlations

We compute Spearman's rank correlation (ρ) (Schober, Boer, and Schwarte 2018) between detection accuracy and perturbation identification quality (measured by NDCG@20) across all configurations (Table 5). AG News and Yelp show strong positive correlations, establishing that for these datasets, heuristics that better identify perturbations consistently achieve higher detection accuracy. Gradient-based heuristics excel in both perturbation identification and detection under these conditions. IMDB departs from this pattern, showing no significant correlation between perturbation identification and detection performance. The dataset-dependent patterns reveal that the relationship between explanation quality and detection effectiveness is not universal.

9.2 Attack-Specific Correlation Patterns

Analyzing correlations by attack type (Table 6) reveals that the relationship between perturbation identification and detection performance depends critically on attack type. Word-level attacks show strong positive correlations between perturbation identification and detection accuracy; when heuristics accurately rank these perturbations, detectors achieve better performance. DeepWordBug presents a fundamentally different pattern, showing no correlation. Character-level attacks operate through different mechanisms where NDCG-based perturbation identification becomes less relevant, and alternative detection mechanisms dominate.

10 Conclusion

We introduced Guided Perturbation Sensitivity (GPS), an adversarial text detector that exploits a fundamental property of adversarial examples: their embedding representations are measurably less stable than those of benign text. Adversarial inputs exhibit approximately $2\times$ higher sensitivity to strategic word masking compared to benign text, a pattern consistent across importance heuristics. By measuring this embedding drift, GPS provides an empirical link between the theoretical instability of adversarial examples near decision boundaries and practical detection in NLP systems.

Our evaluation across 18 configurations reveals that effective adversarial detection is attack-type specific. Word-level attacks exhibit a strong correlation ($\rho > 0.65$) between perturbation identification quality and detection accuracy, validating that gradient-based importance ranking directly enables effective detection. Character-level attacks exhibit no such correlation, operating through different embedding disruption patterns. Cross-architecture transfer experiments further reveal that embedding sensitivity patterns learned on larger models transfer effectively to smaller architectures, indicating that adversarial signatures generalize across model capacities.

GPS achieves 85%+ detection accuracy while generalizing across datasets, architectures, and attack types without retraining. Its linear scaling with K enables practitioners to balance accuracy against computational cost, with $K{=}5$ capturing 98% of peak performance. Limitations include the requirement for white-box model access and labeled training data for the BiLSTM detector. Future work should explore adaptive selection of K based on input characteristics and ensemble strategies combining gradient and attention heuristics to capture both word-level and character-level attack signatures. Beyond detection, embedding instability analysis may inform the design of inherently robust architectures and more targeted adversarial training strategies.

Acknowledgments

Research partly supported by NSF grant 2244279, ARO grant W911NF-23-1-0191 and a US Department of Transportation grant for CyberCare. Verma is the founder of Everest Cyber Security and Analytics, Inc.

References

Abnar, S.; and Zuidema, W. 2020. Quantifying Attention Flow in Transformers. In Jurafsky, D.; Chai, J.; Schluter, N.; and Tetreault, J., eds., *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 4190–4197. Online: Association for Computational Linguistics.

Alhazmi, A.; Aljubairy, A.; Zhang, W.; Sheng, Q. Z.; and Alhazmi, E. 2025. Can Interpretability of Deep Learning Models Detect Textual Adversarial Distribution? *ACM Trans. Intell. Syst. Technol.* Just Accepted.

Alzantot, M.; Sharma, Y.; Elgohary, A.; Ho, B.-J.; Srivastava, M.; and Chang, K.-W. 2018. Generating Natural Language Adversarial Examples. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2890–2896.

Barkan, O.; Hauon, E.; Caciularu, A.; Katz, O.; Malkiel, I.; Armstrong, O.; and Koenigstein, N. 2021. Grad-SAM: Explaining Transformers via Gradient Self-Attention Maps. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, CIKM '21, 2882–2887. New York, NY, USA: Association for Computing Machinery. ISBN 9781450384469.

Bell, B.; Geyer, M.; Glickenstein, D.; Hamm, K.; Scheidegger, C. E.; Fernandez, A. S.; and Moore, J. 2024. Persistent Classification: Understanding Adversarial Attacks by Studying Decision Boundary Dynamics. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 18.

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

Carlini, N.; and Wagner, D. 2017. Towards Evaluating the Robustness of Neural Networks. In *IEEE Symposium on Security and Privacy (SP)*, 39–57.

Ebrahimi, J.; Rao, A.; Lowd, D.; and Dou, D. 2018. HotFlip: White-Box Adversarial Examples for Text Classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, Volume 2: Short Papers (ACL)*, 31–36.

Fawzi, A.; Moosavi-Dezfooli, S.-M.; Frossard, P.; and Soatto, S. 2018. Empirical Study of the Topology and Geometry of Deep Networks. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3762–3770.

Gao, J.; Lanchantin, J.; Soffa, M. L.; and Qi, Y. 2018. Black-Box Generation of Adversarial Text Sequences to Evade Deep Learning Classifiers. In *Proceedings of the IEEE Security and Privacy Workshops (SPW)*, 50–56.

Goodfellow, I. J.; Shlens, J.; and Szegedy, C. 2014. Explaining and Harnessing Adversarial Examples. *CoRR*, abs/1412.6572.

Guo, C.; Sablayrolles, A.; Jégou, H.; and Kiela, D. 2021. Gradient-based Adversarial Attacks against Text Transformers. In Moens, M.-F.; Huang, X.; Specia, L.; and Yih, S. W.-t., eds., *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 5747–5757. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics.

He, P.; Gao, J.; and Chen, W. 2021. DeBERTaV3: Improving DeBERTa using ELECTRA-Style Pre-Training with Gradient-Disentangled Embedding Sharing. *ArXiv*, abs/2111.09543.

Huang, L.; Joseph, A. D.; Nelson, B.; Rubinstein, B. I.; and Tygar, J. D. 2011. Adversarial machine learning. In *Proceedings of the 4th ACM Workshop on Security and Artificial Intelligence*, AISec '11, 43–58. New York, NY, USA: Association for Computing Machinery. ISBN 9781450310031.

Jain, S.; and Wallace, B. C. 2019. Attention is not Explanation. In *North American Chapter of the Association for Computational Linguistics*.

Jia, R.; Raghunathan, A.; Göksel, K.; and Liang, P. 2019. Certified Robustness to Adversarial Word Substitutions. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 4129–4142.

Jin, D.; Jin, Z.; Zhou, J.; and Szolovits, P. 2020. Is Bert Really Robust? A Strong Baseline for Natural Language Attack on Text Classification and Entailment. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34: 8018–8025.

Jones, E.; Jia, R.; Raghunathan, A.; and Liang, P. 2020. Robust Encodings: A Framework for Combating Adversarial Typos. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2752–2765.

Lee, K.; Lee, K.; Lee, H.; and Shin, J. 2018. A Simple Unified Framework for Detecting Out-of-Distribution Samples and Adversarial Attacks. In Bengio, S.; Wallach, H.; Larochelle, H.; Grauman, K.; Cesa-Bianchi, N.; and Garnett, R., eds., *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

Li, G.; Shi, B.; Liu, Z.; Kong, D.; Wu, Y.; Zhang, X.; Huang, L.; and Lyu, H. 2023. Adversarial Text Generation by Search and Learning. In Bouamor, H.; Pino, J.; and Bali, K., eds., *Findings of the Association for Computational Linguistics: EMNLP 2023*, 15722–15738. Singapore: Association for Computational Linguistics.

Li, L.; Ma, R.; Guo, Q.; Xue, X.; and Qiu, X. 2020. BERT-ATTACK: Adversarial Attack Against BERT Using BERT. In Webber, B.; Cohn, T.; He, Y.; and Liu, Y., eds., *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 6193–6202. Online: Association for Computational Linguistics.

Li, Z.; Xu, J.; Zeng, J.; Li, L.; Zheng, X.; Zhang, Q.; Chang, K.-W.; and Hsieh, C.-J. 2021. Searching for an Effective Defender: Benchmarking Defense against Adversarial Word Substitution. In *Conference on Empirical Methods in Natural Language Processing*.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019.

RoBERTa: A Robustly Optimized BERT Pretraining Approach. *ArXiv*, abs/1907.11692.

Maas, A. L.; Daly, R. E.; Pham, P. T.; Huang, D.; Ng, A. Y.; and Potts, C. 2011. Learning Word Vectors for Sentiment Analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 142–150. Portland, Oregon, USA: Association for Computational Linguistics.

Mehdi Gholampour, P.; and Verma, R. M. 2023. Adversarial Robustness of Phishing Email Detection Models. In *Proceedings of the 9th ACM International Workshop on Security and Privacy Analytics*, IWSPA '23, 67–76. New York, NY, USA: Association for Computing Machinery. ISBN 9798400700996.

Miyato, T.; Dai, A. M.; and Goodfellow, I. 2017. Adversarial Training Methods for Semi-Supervised Text Classification. In *International Conference on Learning Representations (ICLR)*.

Morris, J. X.; 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 *Conference on Empirical Methods in Natural Language Processing*.

Mosca, E.; Agarwal, S.; Rando, J.; and Groh, G. L. 2022. "That Is a Suspicious Reaction!": Interpreting Logits Variation to Detect NLP Adversarial Attacks. In *Annual Meeting of the Association for Computational Linguistics*.

Mozes, M.; Müller, B.; Nikolaev, V.; and Schuller, B. 2021. Frequency-Guided Word Substitutions for Detecting Textual Adversarial Examples. In *European Chapter of the Association for Computational Linguistics (EACL)*.

Papernot, N.; McDaniel, P.; Swami, A.; and Harang, R. 2016. Crafting Adversarial Input Sequences for Recurrent Neural Networks. *arXiv preprint arXiv:1604.08275*.

Rocamora, E. A.; Wu, Y.; Liu, F.; Chrysos, G.; and Cevher, V. 2024. Revisiting Character-level Adversarial Attacks for Language Models. In *Forty-first International Conference on Machine Learning*.

Schober, P.; Boer, C.; and Schwarte, L. A. 2018. Correlation Coefficients: Appropriate Use and Interpretation. *Anesthesia and Analgesia*, 126(5): 1763–1768.

Selvaraju, R. R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; and Batra, D. 2017. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. In *2017 IEEE International Conference on Computer Vision (ICCV)*, 618–626.

Shen, L.; Pu, Y.; Zhang, X.; Ge, C.; Yang, X.; Peng, H.; Wang, W.; and Ji, S. 2025. TextDefense: Adversarial Text Detection based on Word Importance Score Dispersion. *IEEE Transactions on Dependable and Secure Computing*, 1–15.

Shen, L.; Zhang, Z.; Jiang, H.; and Chen, Y. 2023. TextShield: Beyond Successfully Detecting Adversarial Sentences in text classification. In *The Eleventh International Conference on Learning Representations*.

Simonyan, K.; Vedaldi, A.; and Zisserman, A. 2013. Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. *CoRR*, abs/1312.6034.

Szegedy, C.; Zaremba, W.; Sutskever, I.; Bruna, J.; Erhan, D.; Goodfellow, I.; and Fergus, R. 2014. Intriguing properties of neural networks. In *International Conference on Learning Representations (ICLR)*.

Tuck, B. E. 2025. LLMs Under Attack: Understanding the Adversarial Mindset. In *Proceedings of the 10th ACM International Workshop on Security and Privacy Analytics*, IWSPA '25, 34–35. New York, NY, USA: Association for Computing Machinery. ISBN 9798400715013.

Wang, X.; Jin, H.; Yang, Y.; and He, K. 2021. Natural Language Adversarial Defense through Synonym Encoding. In *Conference on Uncertainty in Artificial Intelligence (UAI)*.

Wang, Y.; Wang, L.; Li, Y.; He, D.; and Liu, T.-Y. 2013. A Theoretical Analysis of NDCG Type Ranking Measures. In *Annual Conference Computational Learning Theory*.

Ye, M.; Gong, C.; and Liu, Q. 2020. SAFER: A Structure-free Approach for Certified Robustness to Adversarial Word Substitutions. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 3465–3475.

Zeng, J.; Xu, J.; Zheng, X.; and Huang, X. 2023. Certified Robustness to Text Adversarial Attacks by Randomized [MASK]. *Computational Linguistics*, 49(2): 395–427.

Zhang, X.; Zhao, J.; and LeCun, Y. 2015. Character-level convolutional networks for text classification. In *Proceedings of the 29th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'15, 649–657. Cambridge, MA, USA: MIT Press.

Zhang, Z.; Yao, W.; Liang, S.; and Xu, C. 2024. Random Smooth-based Certified Defense against Text Adversarial Attack. In *Findings of the European Chapter of the Association for Computational Linguistics (EACL)*.

Zheng, R.; Dou, S.; Zhou, Y.; Liu, Q.; Gui, T.; Zhang, Q.; Wei, Z.; Huang, X.; and Zhang, M. 2023. Detecting Adversarial Samples through Sharpness of Loss Landscape. In Rogers, A.; Boyd-Graber, J.; and Okazaki, N., eds., *Findings of the Association for Computational Linguistics: ACL 2023*, 11282–11298. Toronto, Canada: Association for Computational Linguistics.

Zhou, Y.; Jiang, J.-Y.; Chang, K.-W.; and Wang, W. 2019. Learning to Discriminate Perturbations for Blocking Adversarial Attacks in Text Classification. In Inui, K.; Jiang, J.; Ng, V.; and Wan, X., eds., *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 4904–4913. Hong Kong, China: Association for Computational Linguistics.