
Proceedings Paper:
Wang, Y. and Lin, C. (2025) Tougher text, smarter models: Raising the bar for adversarial defence benchmarks. In: Rambow, O., Wanner, L., Apidianaki, M., Al-Khalifa, H., Di Eugenio, B. and Schockaert, S., (eds.) Proceedings of the 31st International Conference on Computational Linguistics. 31st International Conference on Computational Linguistics (COLING), 19-24 Jan 2025, Abu Dhabi, UAE. Association for Computational Linguistics , pp. 6475-6491.

Tougher Text, Smarter Models: Raising the Bar for Adversarial Defence Benchmarks

Yang Wang[1,3] **and Chenghua Lin**[1,2]
[1]Department of Computer Science, The University of Sheffield, UK
[2]Department of Computer Science, The University of Manchester, UK
[3]Automated Analytics, UK
y.wang4@sheffield.ac.uk chenghua.lin@manchester.ac.uk

Abstract

Recent advancements in natural language processing have highlighted the vulnerability of deep learning models to adversarial attacks. While various defence mechanisms have been proposed, there is a lack of comprehensive benchmarks that evaluate these defences across diverse datasets, models, and tasks. In this work, we address this gap by presenting an extensive benchmark for textual adversarial defence that significantly expands upon previous work. Our benchmark incorporates a wide range of datasets, evaluates state-of-the-art defence mechanisms, and extends the assessment to include critical tasks such as single-sentence classification, similarity and paraphrase identification, natural language inference, and commonsense reasoning. This work not only serves as a valuable resource for researchers and practitioners in the field of adversarial robustness but also identifies key areas for future research in textual adversarial defence. By establishing a new standard for benchmarking in this domain, we aim to accelerate progress towards more robust and reliable natural language processing systems.

1 Introduction

Recent advancements in natural language processing (NLP) have led to impressive performance on various tasks, but also exposed the vulnerability of deep learning models to adversarial attacks (Wang et al., 2021; Han et al., 2022; Wang et al., 2022a; Ranjan et al., 2023; Zeng et al., 2023; Goyal et al., 2023; Shayegani et al., 2023; Huang et al., 2024). While numerous defence mechanisms have been proposed to counter these threats, there is a lack of comprehensive benchmarks to evaluate their effectiveness across diverse settings.

The advent of adversarial training (Goodfellow et al., 2014a) has demonstrated notable success in enhancing model robustness against small adversarial perturbations in computer vision. Traditional approaches adapt the training process to minimise empirical risk based on a *robustness loss*, as opposed to the standard loss applied to clean input samples (Madry et al., 2018). The robustness loss refers to the standard loss applied to the worst-case (i.e. loss-maximising) adversarial example for each training sample. In the context of NLP, however, adversarial training poses unique challenges due to the discrete nature of text. Specifically, the inner maximisation step required in the min-max formulation of adversarial training becomes computationally expensive (Yoo and Qi, 2021). To address this, various methods have been proposed in the literature, ranging from augmenting the training set with adversarial examples tailored to a specific model (Si et al., 2021; Dong et al., 2021; Zhou et al., 2021a), to more sophisticated optimisations in token-embedding space for the inner maximisation step (Zhu et al., 2020; Li and Qiu, 2021; Goyal et al., 2023).

In parallel, other studies focus on structure-free regularisation methods for adversarial robustness. Yang et al. (2023b) argue that encouraging higher entropy (i.e. uncertainty) in model outputs can enhance adversarial robustness. They emphasise the need to understand the *inherent robustness* properties of models, focusing on those that are flexible, simple, and not overly specialised for specific types of text adversarial attacks, as well as the interplay between a model's confidence and robustness. Building on this idea, they highlight that entropy regularisation techniques, such as label smoothing (Szegedy et al., 2015, 2016), can implicitly contribute to adversarial robustness by addressing model overconfidence. Similarly, Raina et al. (2024) proposed training-time temperature scaling as a defence mechanism. They empirically demonstrated that highly miscalibrated models (Guo et al., 2017) interfere with an adversarial attacker's ability to find meaningful search directions due to the little sensitivity in the predicted probabilities.

Proceedings of the 31st International Conference on Computational Linguistics, pages 6475–6491
January 19–24, 2025.

Unlike those adversarial training-based methods, which rely on manipulating the token-embedding space for inner maximisation to enhance adversarial robustness, regularisation-based approaches offer a more attractive and effective alternative. These regularisation-based methods are synonyms-agnostic and structure-free, which can be seamlessly applied across a broad spectrum of NLP tasks, extending beyond traditional text classification. To the best of our knowledge, the most recent benchmark in this area was established by Li et al. (2021b). They offered foundational insights but limited their focus to text classification tasks, evaluating only two datasets with defence methods developed prior to 2021. In contrast, our work broadens the evaluation by emphasising synonyms-agnostic and structure-free methods, ensuring broad applicability and relevance to a wider array of NLP challenges and tasks. Our contributions include:

1. We argue that the existing adversarial defence benchmark, as established by Li et al. (2021b), is limited in scope. In response, we extend the evaluation to include more NLP datasets, tasks, models, and recent advanced adversarial defence techniques.

2. We propose TTSO++, a variant of training-time temperature scaling that incorporates dynamic confidence adjustment through an entropy term. This adaptation enhances robustness against adversarial attacks, especially under TextFooler and TextBugger scenarios.

Our code is available at `https://github.com/PuReDefence/AdvBench4Text`.

2 Background

The vulnerability of deep learning models to adversarial attacks has become a significant concern in NLP. This section provides an overview of adversarial attacks and defences in NLP, with a particular focus on flexible defence methods that can be adapted to various NLP tasks.

2.1 Adversarial Attacks

Adversarial attacks in NLP aim to manipulate input text in ways that preserve semantic meaning but cause model misclassification. Following notation in Raina and Gales (2023) the distance between the benign sample x and the adversarial example $\tilde{x}$ can be measured via a proxy function $\mathcal{G}(x, \tilde{x}) \leq \epsilon$, where ϵ represents the maximum imperceptibility threshold. Goyal et al. (2023) categorised these attacks based on the attacker's knowledge (white-box vs. black-box), the perturbation level (character, word, or sentence-level), and the attack goal (targeted vs. untargeted). Common attack methods include word substitution (Ren et al., 2019; Zang et al., 2020; Li et al., 2020b; Garg and Ramakrishnan, 2020; Jin et al., 2020; Maheshwary et al., 2021; Waghela et al., 2024; Lu et al., 2024), character manipulation (Gao et al., 2018; Eger et al., 2019a,b; Pruthi et al., 2019; Liu et al., 2022a; Rocamora et al., 2024), and sentence paraphrasing (Ribeiro et al., 2018; Iyyer et al., 2018; Zhao et al., 2018; Li et al., 2020a, 2021a). Many of these popular attack methods are implemented in the TextAttack library (Morris et al., 2020).

2.2 Adversarial Defences

In this section, we will discuss two different types of adversarial defence methods.

2.2.1 Adversarial Training-based Methods

Numerous defence methods have been proposed to counter adversarial threats. In computer vision, adversarial training (Goodfellow et al., 2014b) minimises empirical risk from worst-case adversarial examples, but its inner maximisation step is computationally expensive for NLP models. To address this, a group of adversarial training methods like PGD (Madry et al., 2018), FreeLB (Zhu et al., 2020), and TAVAT (Li and Qiu, 2021) accelerate optimisation by identifying adversarial examples in the token-embedding space.

Despite their efficiency, the limited success of these methods is often attributed to perturbations in the embedding space, which may not adequately represent true adversarial examples in natural language. To mitigate this issue, approaches such as ASCC (Dong et al., 2021) and DNE (Zhou et al., 2021b) proposed a more meaningful embedding perturbation space, defining it as the convex hull of word synonyms. While these methods offer improved robustness, they require precomputation of synonyms, limiting their adaptability and effectiveness against diverse adversarial attacks. In light of these challenges, we emphasise the need for synonyms-agnostic and structure-free defence strategies, which provide broader applicability across NLP tasks. In practical scenarios, defenders should not rely on prior knowledge of the adversary's mechanisms for generating synonyms, as this can limit the robustness of the defence.

2.2.2 Regularisation-based Methods

Regularisation-based methods have emerged as a more flexible and generalisable approach to adversarial defence in NLP, particularly because they do not rely on model structures or synonym sets, making them adaptable across a wide range of tasks. Methods such as Flooding-X (Liu et al., 2022c), adversarial label smoothing (Yang et al., 2023b), and temperature scaling (Raina et al., 2024) have demonstrated notable effectiveness in enhancing adversarial robustness.

Flooding-X (Liu et al., 2022c) aims to prevent overconfidence in model predictions by maintaining the loss around a pre-defined "flood" level (Ishida et al., 2020), thereby mitigating the model's susceptibility to adversarial perturbations. Label smoothing (Szegedy et al., 2016), on the other hand, modifies the training objective by softening the hard labels, distributing a small amount of probability mass across all classes, which helps in reducing the model's confidence in incorrect predictions. Yang et al. (2023b) extensively studied standard label smoothing and its adversarial variant (Ren et al., 2022), and showed that label smoothing can improve robustness to textual adversarial attacks (both black-box and white-box) and mitigate overconfident errors on adversarial examples. Additionally, Raina et al. (2024) highlighted that the extreme class confidence exhibited by miscalibrated models (Guo et al., 2017) creates an illusion of robustness (IOR). To address this, they proposed training-time temperature scaling as a defence mechanism to improve *true* robustness against unseen attacks. Their empirical results showed that highly miscalibrated models impede adversarial attackers by reducing sensitivity in predicted probabilities, thereby limiting the attacker's ability to identify meaningful search directions.

Together, these regularisation-based methods provide a synonyms-agnostic and structure-free framework for adversarial defence, making them well-suited for diverse NLP tasks without requiring prior knowledge of adversarial strategies.

3 Experiments

3.1 Datasets

Experiments are carried out on six NLP datasets (statistics summarised in Table 1), including different tasks: single-sentence classification, similarity and paraphrase identification, natural language inference, and commonsense reasoning.

Dataset	# Classes	Train	Validation	Test	Task
SST2	2	6920	872	1821	single-sentence classification
MR	2	8530	1066	1066	single-sentence classification
MRPC	2	3668	408	1725	paraphrase identification
SciTail	2	23088	2126	1304	natural language inference
SIQA	3	33410	1954	-	commonsense reasoning
CSQA	5	9741	1221	-	commonsense reasoning

Table 1: Dataset statistics.

SST2 (Socher et al., 2013) is a binary sentiment classification task where each sample consists of a single sentence from movie reviews. The objective is to predict whether a given sentence expresses positive or negative sentiment. MR (Pang and Lee, 2005) is another binary sentiment classification dataset similar to SST-2, based on movie reviews. Each sentence is labelled as expressing either positive or negative sentiment. MRPC (Dolan and Brockett, 2005) is a binary classification dataset for similarity and paraphrase identification, where the task is to determine whether two sentences in a pair are semantically equivalent. SciTail (Khot et al., 2018) is a natural language inference (NLI) dataset designed to test a model's ability to recognise entailment. SIQA (Sap et al., 2019) is a commonsense reasoning dataset where the goal is to choose the most appropriate answer from three options to questions about everyday social situations. SIQA presents a challenge in understanding social dynamics and reasoning beyond explicit facts. CSQA (Talmor et al., 2019) is another multiple-choice question answering dataset that requires different types of commonsense knowledge to predict the correct answers.

These datasets cover a range of tasks, including single-sentence classification, similarity and paraphrase identification, natural language inference, and commonsense reasoning, enabling comprehensive evaluation across multiple dimensions of language understanding. Each dataset was carefully selected to ensure diversity in task complexity and linguistic phenomena, providing a robust benchmark for assessing model performance in various natural language understanding (NLU) tasks.

3.2 Models

We follow existing adversarial robustness literature (Raina et al., 2024; Zhao et al., 2024; Moraffah et al., 2024) and use Transformer (Vaswani, 2017) encoders, which are state-of-the-art on many NLP tasks[1]. Specifically, we consider the base variants

[1]Appendix A shows the performance of encoder-only models relative to generative LLMs for many classification tasks.

Model	Checkpoint	Params
BERT-base	google-bert/bert-base-uncased	109M
RoBERTa-base	FacebookAI/roberta-base	124M
DeBERTa-base	microsoft/deberta-v3-base	184M
BGE-M3	BAAI/bge-m3	567M

Table 2: Pre-trained language models (PLMs) checkpoints from HuggingFace Hub. **Model**: Lists the names of different PLMs. **Checkpoint**: Specifies the HuggingFace checkpoint name with each model. **Params**: Indicates the number of parameters in each model.

of BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019b), and DeBERTa (He et al., 2020), which are the most commonly used baseline models in prior adversarial defence studies. To extend this evaluation, we also assess adversarial robustness using a more recent state-of-the-art embedding model BGE-M3 (Chen et al., 2024). A summary of all evaluated models is presented in Table 2.

While generative large language models (LLMs) such as Llama (Dubey et al., 2024) and ChatGPT (OpenAI, 2023) have demonstrated impressive capabilities in various NLP tasks, their inclusion in adversarial robustness evaluations for our benchmark datasets is not appropriate. Our preliminary experiments (summarised in Table 7) on some of our benchmark datasets show that generative LLMs like Llama3-8B (Dubey et al., 2024) and Phi3-3.8B (Abdin et al., 2024) perform poorly on clean accuracy compared to smaller, discriminative models such as BERT, RoBERTa, and DeBERTa. Despite their large parameter counts, these models consistently underperform on clean (without attack) classification tasks, which undermines the significance of their adversarial robustness, as robustness should be evaluated in the context of maintaining high accuracy on before-attack data. Given the high computational cost and lower clean accuracy of these models, it is misleading to report a high after-attack accuracy and a low attack success rate without considering their poor baseline before-attack performance.

3.3 Adversarial Defence Approaches

We consider seven defence baselines in our benchmark: PGD (Madry et al., 2018), FreeLB (Zhu et al., 2020), TAVAT (Li and Qiu, 2021), Flooding-X (Liu et al., 2022c), standard label smoothing (SLS) and adversarial label smoothing (ALS) (Yang et al., 2023b), and training-time temperature scaling optimisation (TTSO) (Raina et al., 2024).

We further create a simple variant of the baseline TTSO that uses entropy-based temperature scaling during training, named TTSO++. This approach adjusts the temperature based on the entropy of the prediction distribution. High entropy indicates that the model is uncertain, so a lower temperature can be applied to sharpen the distribution. Conversely, low entropy (high certainty) can be smoothed by applying a higher temperature. The temperature T is adjusted according to the entropy $H(\cdot)$ of the softmax distribution p:

$$T = T_{base} + \alpha \cdot H(p) \tag{1}$$

where $H(p)$ is the entropy of the softmax probabilities p, T_{base} is the base temperature, and α is a scaling factor controlling how strongly the temperature reacts to uncertainty. By adding entropy $H(p)$ to the temperature scaling formula, we introduce *dynamic confidence adjustment* based on the model's uncertainty. Note that Balanya et al. (2024) also proposed an entropy-based temperature scaling method, but we introduce a simpler one that does not need any learnable parameters.

3.4 Evaluation Metrics

We follow the conventions in the literature (Li et al., 2021b; Liu et al., 2022c; Lee et al., 2022) to evaluate our benchmark. We leverage TextFooler (Jin et al., 2020) and TextBugger (Li et al., 2018) to attack the victim models and measure the empirical performance. Both attackers are implemented using the default settings from the TextAttack library (Morris et al., 2020). While we acknowledge the advancements in attack techniques, TextAttack currently provides limited support for newer methods and only includes adversarial attack methods developed prior to 2021. Similarly, another widely-used OpenAttack (Zeng et al., 2021) library only covers adversarial attack methods up to 2020. Therefore, we focused on three well-established, general-purpose attack methods that are widely recognised for evaluating adversarial robustness (Wang et al., 2022b; Yang et al., 2023a; Zhan et al., 2023; Hu et al., 2023; Yang et al., 2023b; Lu et al., 2024; Ji et al., 2024; Zhang et al., 2024; Zhao et al., 2024).

To quantify the impact of each adversarial attack, we follow prior works (Li et al., 2021b; Liu et al., 2022c; Hu et al., 2023) and report the following metrics: accuracy under attack (AUA), attack success rate (ASR), and the average number of queries (AVGQ) required to successfully attack a model.

Additionally, we provide the before-attack accuracy to offer a baseline for comparison, and quantify the relative performance decline using performance drop rate (PDR).

Clean Accuracy (ACC) measures the accuracy of the model on the before-attack dataset. It provides a baseline for how well the model performs without adversarial interference.

Accuracy Under Attack (AUA) evaluates the accuracy of the model when subjected to adversarial examples. A higher AUA indicates better robustness against adversarial attacks.

Attack Success Rate (ASR) is the percentage of adversarial attacks that successfully cause the model to misclassify. A lower ASR signifies a more robust model.

Number of Queries (AVGQ) quantifies the average number of queries made to the model by an adversarial attack to achieve success. A higher number implies the model is harder to attack (Li et al., 2021c).

Performance Drop Rate (PDR) quantifies the relative performance decline, and provides a normalised measure for comparing different attacks (Zhu et al., 2023). APDR stands for average PDR across different attacks.

In contrast to prior work (Dong et al., 2021; Bao et al., 2021; Zheng et al., 2022; Liu et al., 2022b; Wang et al., 2022b; Hu et al., 2023; Zeng et al., 2023; Zhan et al., 2023; Wang et al., 2023), which often limits evaluations to a small subset of test samples from their datasets, we advocate for the inclusion of the *entire* test set across all datasets. This comprehensive evaluation ensures a more robust assessment of the defence methods' effectiveness. Such an approach contrasts with the prevailing practice in the field, where evaluations may be restricted to a small portion of the available test data, potentially leading to an incomplete representation of a model's performance across diverse scenarios.

3.5 Implementation Details

All experiments were conducted using the HuggingFace framework (Wolf et al., 2020) to leverage pre-trained model weights. For the adversarial training-based methods, including PGD, FreeLB, and TAVAT, we used the default hyper-parameters provided by the TextDefender library (Li et al., 2021b). The default hyper-parameters for each adversarial training baseline are: adversarial iterations of 5, adversarial learning rate of 0.03, adversarial initialisation magnitude of 0.05, adversarial maximum norm of 1, adversarial norm type of l_2. For experiments involving SLS and ALS, we performed a hyper-parameter search for the label smoothing coefficient from the set $\{0.1, 0.2, 0.3, 0.4, 0.5\}$. In experiments involving TTSO and TTSO++, we applied the same high temperature $T = 10$ to every instance and scaling factor $\alpha = 0.5$ by default. The learning rate was optimised by selecting the model that achieved the highest validation accuracy after fine-tuning for four epochs, with candidate values for the learning rate drawn from the set $\{1e-5, 2e-5, 5e-5\}$. For commonsense reasoning datasets, we follow Branco et al. (2021) to fine-tune the pre-trained model, converting the multiple-choice task into a sequence-ranking problem, as outlined in Liu et al. (2019a). We process the elements of input pairs separately, generating a score for each, with the maximum score corresponding to the selected answer. All experiments were executed on a single NVIDIA RTX 4090 GPU with 24GB of memory.

3.6 Results

3.6.1 Robustness in Classification Tasks

Table 3 presents the experimental results trained with various defence methods. Notably, TTSO and TTSO++ consistently outperform other baselines, achieving superior AUA across diverse attacks (TextFooler and TextBugger) and model architectures (BERT, RoBERTa, and DeBERTa). This robustness can be attributed to their ability to counteract the Illusion of Robustness (IOR) by addressing model miscalibration (Raina et al., 2024), a key factor behind overconfidence in adversarial scenarios. Unlike token-level embedding perturbation techniques such as PGD, FreeLB, and TAVAT, which often lead to overfitting specific attack patterns without enhancing overall model uncertainty, TTSO and TTSO++ effectively recalibrate model confidence by softening predictions, setting a new benchmark for adversarial defence strategies.

In comparison, methods like SLS and ALS emerge as flexible and lightweight alternatives to adversarial training-based methods. While approaches such as PGD, FreeLB, or TAVAT require computationally expensive inner maximisation steps during training and sometimes degrade performance under adversarial conditions, SLS and ALS offer significant improvements in adversarial robustness with minimal additional complexity. As shown in Table 3, Flooding-X consistently under-

Dataset	Defence	BERT			RoBERTa			DeBERTa		
		ACC↑	AUA↑		ACC↑	AUA↑		ACC↑	AUA↑	
			TF	TB		TF	TB		TF	TB
SST2	-	91.54	6.59	28.08	93.79	7.80	31.52	95.22	8.57	39.65
	PGD	**92.64**	8.73	34.27	94.29	8.79	34.21	94.67	8.73	38.88
	FreeLB	91.98	8.57	31.80	95.11	6.43	36.90	95.22	10.32	44.10
	TAVAT	**92.64**	10.71	34.32	**95.17**	9.23	37.40	**95.83**	11.04	47.01
	Flooding-X	89.84	11.64	30.37	94.95	5.44	29.65	95.22	7.36	31.74
	SLS	91.76	12.25	39.21	94.34	11.64	44.81	95.22	22.24	54.48
	ALS	91.21	15.54	39.37	94.51	23.50	52.50	95.55	15.76	52.22
	TTSO	91.71	41.63	50.85	94.67	45.63	56.84	95.66	55.02	65.84
	TTSO++	91.76	**43.27**	**53.27**	94.78	**48.93**	**59.91**	95.55	**56.07**	**66.23**
MR	-	85.55	3.94	22.80	88.65	6.75	31.80	90.71	9.94	34.24
	PGD	85.93	12.85	35.46	87.90	6.29	34.33	89.21	5.35	30.11
	FreeLB	86.30	6.66	28.71	**89.12**	5.53	33.58	91.18	8.63	34.80
	TAVAT	86.02	8.26	30.11	87.99	6.19	32.46	90.90	10.60	38.37
	Flooding-X	85.55	5.44	27.02	88.74	7.79	31.05	**91.37**	6.66	35.46
	SLS	**86.59**	13.60	36.49	88.09	20.83	42.59	90.62	17.45	45.31
	ALS	85.83	11.07	33.77	87.90	17.07	43.34	89.96	17.92	46.53
	TTSO	86.02	35.27	43.06	87.71	42.40	51.97	90.34	47.84	56.75
	TTSO++	86.02	**38.09**	**45.31**	87.62	**43.39**	**52.94**	90.24	**49.44**	**57.22**
MRPC	-	84.40	2.32	3.25	86.96	6.09	9.57	87.94	2.96	9.86
	PGD	84.06	9.86	11.25	87.48	5.28	11.54	87.83	3.71	10.55
	FreeLB	**85.45**	11.48	11.65	87.54	6.38	11.71	86.43	8.58	15.07
	TAVAT	84.29	8.70	10.43	**87.59**	9.97	15.07	88.06	6.61	16.23
	Flooding-X	82.67	7.48	7.88	87.48	4.75	8.41	88.35	3.30	10.20
	SLS	84.52	6.32	7.36	86.84	9.22	12.35	**88.46**	8.58	12.87
	ALS	82.96	5.97	7.83	86.03	10.61	13.28	**88.46**	6.14	17.45
	TTSO	83.77	41.62	39.71	86.78	46.38	41.91	87.54	50.78	54.38
	TTSO++	83.48	**42.49**	**40.92**	86.84	**50.43**	**42.90**	87.59	**50.99**	**55.12**
SciTail	-	92.80	44.45	32.22	93.60	42.80	31.70	95.53	47.04	33.82
	PGD	93.09	50.19	32.69	93.09	43.27	31.42	94.36	43.32	30.39
	FreeLB	93.60	47.04	32.60	**93.79**	44.21	31.51	95.67	47.04	33.68
	TAVAT	92.29	52.21	30.48	**93.79**	46.38	34.71	**96.52**	50.19	39.98
	Flooding-X	91.58	49.62	35.28	92.43	43.32	29.02	94.73	46.28	32.22
	SLS	92.33	48.02	34.85	93.60	45.58	35.79	95.67	48.64	35.04
	ALS	92.57	50.80	33.44	92.90	44.17	34.48	95.16	50.85	43.09
	TTSO	92.33	52.02	48.21	92.52	51.27	47.22	95.63	55.13	50.05
	TTSO++	**93.74**	**54.47**	**49.86**	93.41	**53.23**	**49.12**	95.25	**56.02**	**51.32**

Table 3: The experiment results of different defence methods. TF: TextFooler. TB: TextBugger. The best performance is marked in **bold**.

performs compared to other baselines. This poor performance aligns with the findings of Zhu and Rao (2023), who found flooding techniques ineffective for adversarial robustness. By maintaining the loss above a threshold, we argue that Flooding-X will hinder the model's ability to minimise adversarial loss and learn intricate decision boundaries. Its non-targeted regularisation treats all examples uniformly, lacking the specificity needed to counter adversarial attacks. While aimed at improving generalisation, Flooding-X appears to compromise the nuanced feature learning required for robust adversarial performance.

3.6.2 Evaluate on Embedding-based Model

While BERT, RoBERTa, and DeBERTa are the most commonly used encoder-based models in prior studies (Raina et al., 2024; Zhao et al., 2024; Moraffah et al., 2024), we extend this evaluation by assessing adversarial robustness using a more recent state-of-the-art embedding-based model BGE-M3 (Chen et al., 2024). Results are summarised in Table 4. TTSO++ consistently achieves superior robustness performance, excelling in all metrics across all datasets and attack types.

3.6.3 Robustness in Commonsense Reasoning

Table 5 highlights the adversarial robustness performance of all baseline defence methods on commonsense reasoning tasks using RoBERTa-base. TTSO++ achieves the best overall performance, with the highest AUA and lowest ASR across both datasets, demonstrating its strong defence capabilities. Flooding-X, however, consistently underperforms, reaffirming its limitations in adversarial settings. Notably, token-level embedding perturbation methods such as PGD, FreeLB, and TAVAT exhibit marginal improvements over the baseline but fail to achieve robustness comparable to TTSO++.

Dataset	Defence	Acc↑	TextFooler			TextBugger			Apdr↓
			Aua↑	Asr↓	AvgQ↑	Aua↑	Asr↓	AvgQ↑	
SST2	-	93.36	6.59	92.94	91.14	38.00	59.29	43.33	76.11
	PGD	92.86	6.59	92.90	94.72	40.14	56.77	44.30	74.84
	FreeLB	93.85	6.43	93.15	90.58	39.43	57.99	43.79	75.57
	TAVAT	**94.89**	6.70	92.94	92.85	40.47	57.35	44.15	75.14
	Flooding-X	93.03	4.50	95.16	85.47	34.43	62.99	42.60	79.08
	SLS	93.79	12.74	86.42	113.85	47.94	48.89	45.56	67.66
	ALS	93.90	12.96	86.20	114.51	50.58	46.14	45.78	66.17
	TTSO	93.63	45.96	50.91	162.66	57.94	38.12	104.55	44.52
	TTSO++	93.36	**47.61**	**49.00**	**163.92**	**58.05**	**37.82**	**105.47**	**43.41**
MR	-	87.15	5.16	94.08	94.13	33.30	61.79	47.11	77.94
	PGD	87.71	5.16	94.12	96.62	35.08	60.00	48.89	77.06
	FreeLB	88.27	4.50	94.90	97.93	35.37	59.94	47.54	77.42
	TAVAT	**89.59**	6.10	93.19	102.61	38.37	57.17	49.03	75.18
	Flooding-X	88.18	4.22	95.21	90.28	33.49	62.02	46.54	78.61
	SLS	87.99	13.32	84.86	124.05	43.71	50.32	49.95	67.59
	ALS	88.37	11.82	86.62	124.74	42.96	51.38	51.01	69.00
	TTSO	88.46	41.09	53.55	**171.37**	50.38	43.05	112.69	48.30
	TTSO++	88.37	**41.93**	**52.55**	170.61	**51.97**	**41.19**	**114.71**	**46.87**
MRPC	-	86.96	4.81	94.47	152.11	12.23	85.93	101.07	90.20
	PGD	86.55	4.58	94.71	161.32	11.88	86.27	104.67	90.49
	FreeLB	86.67	5.45	93.71	152.83	10.32	88.09	102.72	90.90
	TAVAT	**87.54**	4.87	94.44	174.73	15.54	82.25	112.24	88.34
	Flooding-X	87.07	4.81	94.47	149.09	12.64	85.49	100.20	89.98
	SLS	86.84	5.33	93.86	178.08	19.36	77.70	119.15	85.78
	ALS	86.49	8.41	90.28	175.10	15.94	81.57	106.76	85.92
	TTSO	85.74	43.94	48.75	380.00	52.46	38.81	254.36	43.78
	TTSO++	85.57	**44.64**	**47.83**	**385.33**	**52.64**	**38.48**	**257.86**	**43.16**
SciTail	-	94.54	44.97	52.44	105.04	35.32	62.64	96.17	57.54
	PGD	94.97	49.81	47.55	109.80	38.85	59.09	98.06	53.32
	FreeLB	94.78	50.66	46.55	109.44	40.36	57.42	104.09	51.98
	TAVAT	**95.11**	50.80	46.59	111.04	42.00	55.84	102.13	51.22
	Flooding-X	93.27	48.92	47.55	108.44	38.62	58.60	99.93	53.08
	SLS	94.36	48.31	48.80	112.80	40.59	56.98	100.81	52.89
	ALS	94.07	47.74	49.25	109.48	40.59	56.85	97.75	53.05
	TTSO	94.40	54.37	42.40	**123.80**	52.63	44.25	169.03	43.33
	TTSO++	94.31	**55.83**	**40.80**	123.65	**53.86**	**42.89**	**171.95**	**41.84**

Table 4: The experiment results of different defence methods using BGE-M3 model. The best performance is marked in **bold**.

4 Discussion

4.1 Dynamic Confidence Adjustment

From Table 3, we observe that TTSO++ consistently outperforms TTSO across datasets and models in terms of all evaluation metrics. A key factor in this improvement lies in the nuanced difference between the temperature-scaling mechanisms of TTSO and TTSO++. TTSO applies a uniform temperature ($T_{base} = 10$) to all instances during training, ensuring equal smoothing of logits across the dataset. While this strategy offers simplicity and improves model calibration, it is inherently limited. A fixed temperature does not account for variations in the difficulty of individual examples. For easy-to-classify examples (where the model is naturally confident), applying a slightly higher temperature can unnecessarily dampen the predictions, leading to a loss of useful certainty. Conversely, for hard-to-classify examples (where the model should be uncertain) or adversarial instances, applying a fixed high temperature may not be enough to capture the complexity of the example, leading to insufficient adjustment of the logits. In contrast, TTSO++ incorporates entropy-based temperature scaling, where the temperature is dynamically adjusted for each input instance based on the model's certainty. This approach leverages entropy as a proxy for uncertainty. Higher entropy (low certainty) leads to a higher temperature, while lower entropy (high certainty) results in a lower temperature. This adaptive mechanism allows TTSO++ to tailor the level of smoothing to the specific demands of each input, striking a better balance between preserving confidence for easy examples and enhancing robustness for challenging ones. As a result, TTSO++ achieves superior performance, where the ability to dynamically handle uncertain inputs is critical.

The effectiveness of TTSO++ is particularly evident in commonsense reasoning tasks like SIQA and CSQA (Table 5). Here, TTSO++ demonstrates

Dataset	Defence	Acc↑	TextFooler		
			Aua↑	Asr↓	AvgQ↑
SIQA	-	71.24	57.98	18.61	16.15
	PGD	71.24	58.39	18.03	16.18
	FreeLB	71.55	59.01	17.53	16.15
	TAVAT	71.19	62.33	12.44	16.32
	Flooding-X	70.37	57.98	17.60	16.13
	SLS	71.60	59.77	16.51	16.23
	ALS	72.16	59.77	17.16	16.12
	TTSO	71.08	59.21	16.70	16.23
	TTSO++	**72.22**	**63.01**	**11.36**	**17.01**
CSQA	-	59.87	48.24	19.43	11.94
	PGD	59.11	48.01	18.77	11.92
	FreeLB	60.77	48.73	19.81	11.99
	TAVAT	**60.81**	48.51	20.22	12.01
	Flooding-X	58.64	47.42	19.13	11.91
	SLS	59.46	48.48	18.46	12.04
	ALS	58.39	46.76	19.92	12.04
	TTSO	59.91	49.12	18.01	11.62
	TTSO++	58.94	**50.89**	**13.65**	**12.88**

Table 5: The experiment results on the commonsense reasoning tasks (SIQA and CSQA). Following Branco et al. (2021), we employed TextFooler (Jin et al., 2020) and evaluated adversarial performance with RoBERTa-base under the same experimental settings. The best performance is marked in **bold**.

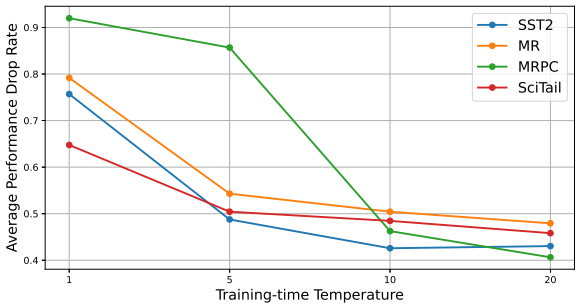

Figure 1: Average performance drop rate (Apdr) across two attackers using TTSO++ with RoBERTa-base as training temperature varies.

the highest Aua and lowest Asr across all models and datasets. The instance-wise temperature scaling provides the model with the flexibility to adapt to diverse question-answering scenarios, effectively mitigating the impact of adversarial attacks. TTSO++ sets a new benchmark, offering superior adversarial robustness and generalisability across datasets and tasks.

4.2 High Temperature Training

While tuning T_{base} could potentially enhance performance against adversarial attacks, we opted for a fixed temperature to ensure consistency and simplicity in our experimental setup. The choice of 10 as the fixed temperature was empirically validated across a range of NLP tasks and demonstrated robust performance across clean and adversarial examples. By using a fixed temperature, we reduce the need for extensive hyper-parameter tuning, which can introduce additional computational overhead and potential overfitting to specific datasets or adversarial attacks.

Defence	SST2	MRPC	SIQA
-	1	1	1
PGD	x6.3	x10.2	x4.1
FreeLB	x2.6	x3.1	x2.2
TAVAT	x4.2	x3.9	x2.6
Flooding-X	x1.2	x1.3	x1.1
SLS	x1.1	x1.1	x1.1
ALS	x1.2	x1.2	x1.1
TTSO	x1.1	x1.1	x1.0
TTSO++	x1.1	x1.1	x1.1

Table 6: Runtime comparison of training RoBERTa-base using different adversarial defence methods.

Figure 1 presents the changes in before- and after-attack accuracy of a model trained with the standard objective and various base temperatures (T_{base}, as described in §3.3) during training. While similar trends were observed across all models, we present the results specifically for RoBERTa-base in this section. The results indicate that higher temperatures during training generally enhance robustness against adversarial attacks. To quantify this, we use the average performance drop rate (Apdr) (Zhu et al., 2023), which averages the performance drop rate

$$\text{Pdr} = 1 - \frac{\sum_{(x;y)\in\mathcal{D}} \mathcal{M}[f_\theta(A(x)), y]}{\sum_{(x;y)\in\mathcal{D}} \mathcal{M}[f_\theta(x), y]} \quad (2)$$

across different adversarial attacks, where A is the adversarial attack applied to input text x, $\mathcal{M}[\cdot]$ is the evaluation function, and $f_\theta(\cdot)$ is the network. For classification task, $\mathcal{M}[\cdot]$ is the indicator function $\mathbb{1}[\hat{y}, y]$ which equals to 1 when $\hat{y} = y$, and 0 otherwise. Notably, on SST2 dataset, we observe a slight increase in Apdr when the training temperature is set to 20. This suggests that excessively high temperatures may overly smooth the predicted probability distribution, making it harder for the model to effectively learn from the training data.

4.3 Runtime Analysis

Table 6 presents the runtime comparison of training the RoBERTa-base model on the SST2, MRPC, and SIQA datasets using different adversarial defence methods. The baseline model (no defence) has a normalised runtime of 1 across all datasets, serving as the standard for comparison.

Adversarial training-based methods such as PGD, FreeLB, and TAVAT introduce significant runtime overhead due to the inclusion of inner maximisation steps. PGD, which requires multiple gradient updates per iteration to approximate adversarial perturbations, is the most computationally expensive, resulting in training times that are x6.3, x10.2, and x4.1 longer for SST2, MRPC, and SIQA, respectively. Although adversarial training-based methods offer slight improvements in adversarial robustness, the runtime cost makes them impractical for large-scale training tasks. In contrast, regularisation-based methods like Flooding-X, SLS, ALS, TTSO, and TTSO++ impose minimal runtime overhead, with training times ranging from x1.1 to x1.3 across all datasets. These methods are particularly appealing for large-scale scenarios, as they do not involve the computationally expensive inner maximisation step. Among these, TTSO++ stands out by combining strong adversarial robustness with a minimal runtime impact. Its entropy-based temperature scaling mechanism effectively adjusts model predictions without requiring extensive computational resources, making it an ideal defence for both efficiency and robustness.

5 Conclusion

In this work, we investigated adversarial defence techniques that are broadly applicable across diverse NLP tasks, focusing on synonym-agnostic and structure-free approaches. By establishing a comprehensive benchmark, we evaluated state-of-the-art adversarial defence strategies developed prior to 2024, extending the evaluation beyond traditional text classification to encompass single-sentence classification, similarity and paraphrase identification, natural language inference, and commonsense reasoning tasks.

Our systematic exploration of regularisation-based methods revealed valuable insights into their potential for textual adversarial defence. Based on these findings, we proposed TTSO++, a simple yet effective variant of temperature scaling that leverages entropy-based adjustments during training. TTSO++ achieves state-of-the-art robustness under adversarial attacks while maintaining strong performance on clean examples. Its minimal computational overhead makes it highly practical for real-world applications. By extending adversarial evaluation to a broader spectrum of NLP tasks, we aim to inspire the development of more flexible, generalisable, and efficient defence mechanisms. We believe this study provides a robust foundation for future research, bridging the gap between task-specific defences and universally applicable solutions for adversarial robustness in NLP.

Limitations

Our study presents empirical results using state-of-the-art encoder-based Transformer models, which are widely regarded as the most appropriate for classification-based NLP tasks (Raina et al., 2024; Zhao et al., 2024). However, the rapidly growing field of LLMs opens new avenues for exploration. Future work could examine the susceptibility of decoder-based LLMs to adversarial attacks and evaluate the performance of the defence methods discussed in this paper in such settings. Additionally, while our research focuses on defence methods that can be uniformly applied across all benchmark datasets, it remains unexplored whether more specialised techniques, such as contrastive-based methods (Pan et al., 2022; he et al., 2023) or prompt-based methods (Xu and Wang, 2024; Yang et al., 2024), could be adapted to provide universal adversarial defences. Investigating these methods' applicability to a broader range of tasks could further enhance the scope of adversarial robustness research. Finally, we proposed TTSO++ as an improvement over the fixed-temperature TTSO method by introducing entropy-based temperature scaling. While TTSO++ demonstrates significant advancements, further optimisation of temperature scaling strategies could yield additional improvements. For example, dynamically adjusting the temperature based on training progression (e.g., curriculum-based or confidence-based scaling) may better align with the evolving complexity of the task during training. Future research could explore these methods to develop more adaptive and effective defences.

Acknowledgments

This work was supported by the Innovate UK Knowledge Transfer Partnership (KTP) grant (Grant Number: 13320). We thank the University of Manchester for providing the computing resources required to conduct the experiments.

References

Marah Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Harkirat Behl, et al. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*.

Sergio A. Balanya, Juan Maroñas, and Daniel Ramos. 2024. Adaptive temperature scaling for robust calibration of deep neural networks. *Neural Comput. Appl.*, 36(14):8073–8095.

Rongzhou Bao, Jiayi Wang, and Hai Zhao. 2021. Defending pre-trained language models from adversarial word substitutions without performance sacrifice. *arXiv preprint arXiv:2105.14553*.

Ruben Branco, António Branco, Joao Rodrigues, and Joao Silva. 2021. Shortcutted commonsense: Data spuriousness in deep learning of commonsense reasoning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1504–1521.

Jianlyu Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. M3-embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2318–2335, Bangkok, Thailand. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Bill Dolan and Chris Brockett. 2005. Automatically constructing a corpus of sentential paraphrases. In *Third International Workshop on Paraphrasing (IWP2005)*.

Xinshuai Dong, Anh Tuan Luu, Rongrong Ji, and Hong Liu. 2021. Towards robustness against natural language word substitutions. *arXiv preprint arXiv:2107.13541*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Steffen Eger, Gözde Gül Şahin, Andreas Rücklé, Ji-Ung Lee, Claudia Schulz, Mohsen Mesgar, Krishnkant Swarnkar, Edwin Simpson, and Iryna Gurevych. 2019a. Text processing like humans do: Visually attacking and shielding NLP systems. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1634–1647, Minneapolis, Minnesota. Association for Computational Linguistics.

Steffen Eger, Gözde Gül Şahin, Andreas Rücklé, Ji-Ung Lee, Claudia Schulz, Mohsen Mesgar, Krishnkant Swarnkar, Edwin Simpson, and Iryna Gurevych. 2019b. Text processing like humans do: Visually attacking and shielding nlp systems. *arXiv preprint arXiv:1903.11508*.

Ji Gao, Jack Lanchantin, Mary Lou Soffa, and Yanjun Qi. 2018. Black-box generation of adversarial text sequences to evade deep learning classifiers. In *2018 IEEE Security and Privacy Workshops (SPW)*, pages 50–56. IEEE.

Tianyu Gao, Adam Fisch, and Danqi Chen. 2021. Making pre-trained language models better few-shot learners. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3816–3830, Online. Association for Computational Linguistics.

Siddhant Garg and Goutham Ramakrishnan. 2020. Bae: Bert-based adversarial examples for text classification. *ArXiv*, abs/2004.01970.

Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. 2014a. Explaining and harnessing adversarial examples. *CoRR*, abs/1412.6572.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. 2014b. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*.

Shreya Goyal, Sumanth Doddapaneni, Mitesh M Khapra, and Balaraman Ravindran. 2023. A survey of adversarial defenses and robustness in nlp. *ACM Computing Surveys*, 55(14s):1–39.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. 2017. On calibration of modern neural networks. In *International conference on machine learning*, pages 1321–1330. PMLR.

Xu Han, Ying Zhang, Wei Wang, and Bin Wang. 2022. Text adversarial attacks and defenses: Issues, taxonomy, and perspectives. *Security and Communication Networks*, 2022(1):6458488.

Jonathan Hayase, Ema Borevkovic, Nicholas Carlini, Florian Tramèr, and Milad Nasr. 2024. Query-based adversarial prompt generation. *arXiv preprint arXiv:2402.12329*.

Jia-long he, Xiao-Lin zhang, Yong-Ping wang, Huan-Xiang zhang, Lu gao, and En-Hui xu. 2023. Contrastive adversarial learning in text classification tasks. *J. Intell. Fuzzy Syst.*, 45(2):3473–3484.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced bert with disentangled attention. *ArXiv*, abs/2006.03654.

Xinrong Hu, Ce Xu, Junlong Ma, Zijian Huang, Jie Yang, Yi Guo, and Johan Barthelemy. 2023. [MASK] insertion: a robust method for anti-adversarial attacks. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1058–1070, Dubrovnik, Croatia. Association for Computational Linguistics.

Xiaowei Huang, Wenjie Ruan, Wei Huang, Gaojie Jin, Yi Dong, Changshun Wu, Saddek Bensalem, Ronghui Mu, Yi Qi, Xingyu Zhao, et al. 2024. A survey of safety and trustworthiness of large language models through the lens of verification and validation. *Artificial Intelligence Review*, 57(7):175.

Takashi Ishida, Ikko Yamane, Tomoya Sakai, Gang Niu, and Masashi Sugiyama. 2020. Do we need zero training loss after achieving zero training error? In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org.

Mohit Iyyer, John Wieting, Kevin Gimpel, and Luke Zettlemoyer. 2018. Adversarial example generation with syntactically controlled paraphrase networks. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1875–1885, New Orleans, Louisiana. Association for Computational Linguistics.

Jiabao Ji, Bairu Hou, Zhen Zhang, Guanhua Zhang, Wenqi Fan, Qing Li, Yang Zhang, Gaowen Liu, Sijia Liu, and Shiyu Chang. 2024. Advancing the robustness of large language models through self-denoised smoothing. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 246–257, Mexico City, Mexico. Association for Computational Linguistics.

Di Jin, Zhijing Jin, Joey Tianyi Zhou, and Peter Szolovits. 2020. Is bert really robust? a strong baseline for natural language attack on text classification and entailment. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 8018–8025.

Tushar Khot, Ashish Sabharwal, and Peter Clark. 2018. Scitail: A textual entailment dataset from science question answering. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

Deokjae Lee, Seungyong Moon, Junhyeok Lee, and Hyun Oh Song. 2022. Query-efficient and scalable black-box adversarial attacks on discrete sequential data via bayesian optimization. In *International Conference on Machine Learning*, pages 12478–12497. PMLR.

Dianqi Li, Yizhe Zhang, Hao Peng, Liqun Chen, Chris Brockett, Ming-Ting Sun, and Bill Dolan. 2020a. Contextualized perturbation for textual adversarial attack. *arXiv preprint arXiv:2009.07502*.

Dianqi Li, Yizhe Zhang, Hao Peng, Liqun Chen, Chris Brockett, Ming-Ting Sun, and Bill Dolan. 2021a. Contextualized perturbation for textual adversarial attack. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5053–5069, Online. Association for Computational Linguistics.

Jinfeng Li, Shouling Ji, Tianyu Du, Bo Li, and Ting Wang. 2018. Textbugger: Generating adversarial text against real-world applications. *ArXiv*, abs/1812.05271.

Linyang Li, Ruotian Ma, Qipeng Guo, Xiangyang Xue, and Xipeng Qiu. 2020b. BERT-ATTACK: Adversarial attack against BERT using BERT. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6193–6202, Online. Association for Computational Linguistics.

Linyang Li and Xipeng Qiu. 2021. Token-aware virtual adversarial training in natural language understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 8410–8418.

Zongyi Li, Jianhan Xu, Jiehang Zeng, Linyang Li, Xiaoqing Zheng, Qi Zhang, Kai-Wei Chang, and Cho-Jui Hsieh. 2021b. Searching for an effective defender: Benchmarking defense against adversarial word substitution. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3137–3147, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Zongyi Li, Jianhan Xu, Jiehang Zeng, Linyang Li, Xiaoqing Zheng, Qi Zhang, Kai-Wei Chang, and Cho-Jui Hsieh. 2021c. Searching for an effective defender: Benchmarking defense against adversarial word substitution. *arXiv preprint arXiv:2108.12777*.

Aiwei Liu, Honghai Yu, Xuming Hu, Shu'ang Li, Li Lin, Fukun Ma, Yawen Yang, and Lijie Wen. 2022a. Character-level white-box adversarial attacks against transformers via attachable subwords substitution. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7664–7676, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):1–35.

Qin Liu, Rui Zheng, Bao Rong, Jingyi Liu, Zhihua Liu, Zhanzhan Cheng, Liang Qiao, Tao Gui, Qi Zhang,

and Xuan-Jing Huang. 2022b. Flooding-x: Improving bert's resistance to adversarial attacks via loss-restricted fine-tuning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5634–5644.

Qin Liu, Rui Zheng, Bao Rong, Jingyi Liu, ZhiHua Liu, Zhanzhan Cheng, Liang Qiao, Tao Gui, Qi Zhang, and Xuanjing Huang. 2022c. Flooding-X: Improving BERT's resistance to adversarial attacks via loss-restricted fine-tuning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5634–5644, Dublin, Ireland. Association for Computational Linguistics.

Xiaodong Liu, Pengcheng He, Weizhu Chen, and Jianfeng Gao. 2019a. Multi-task deep neural networks for natural language understanding. *arXiv preprint arXiv:1901.11504*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019b. Roberta: A robustly optimized bert pretraining approach. *ArXiv*, abs/1907.11692.

Ning Lu, Shengcai Liu, Zhirui Zhang, Qi Wang, Haifeng Liu, and Ke Tang. 2024. Less is more: Understanding word-level textual adversarial attack via n-gram frequency descend. In *2024 IEEE Conference on Artificial Intelligence (CAI)*, pages 823–830. IEEE.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. 2018. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*.

Rishabh Maheshwary, Saket Maheshwary, and Vikram Pudi. 2021. Generating natural language attacks in a hard label black box setting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 13525–13533.

Raha Moraffah, Shubh Khandelwal, Amrita Bhattacharjee, and Huan Liu. 2024. Adversarial text purification: A large language model approach for defense. In *Advances in Knowledge Discovery and Data Mining - 28th Pacific-Asia Conference on Knowledge Discovery and Data Mining, PAKDD 2024, Proceedings*, Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), pages 65–77, Germany. Springer Science and Business Media Deutschland GmbH. Publisher Copyright: © The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2024.; 28th Pacific-Asia Conference on Knowledge Discovery and Data Mining, PAKDD 2024 ; Conference date: 07-05-2024 Through 10-05-2024.

John Morris, Eli Lifland, Jin Yong Yoo, Jake Grigsby, Di Jin, and Yanjun Qi. 2020. Textattack: A framework for adversarial attacks, data augmentation, and adversarial training in nlp. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 119–126.

OpenAI. 2023. `https://chat.openai.com.chat`.

Lin Pan, Chung-Wei Hang, Avirup Sil, and Saloni Potdar. 2022. Improved text classification via contrastive adversarial training. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11130–11138.

Bo Pang and Lillian Lee. 2005. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the ACL*.

Francesco Periti, Haim Dubossarsky, and Nina Tahmasebi. 2024. (chat) gpt v bert: Dawn of justice for semantic change detection. *arXiv preprint arXiv:2401.14040*.

Danish Pruthi, Bhuwan Dhingra, and Zachary C. Lipton. 2019. Combating adversarial misspellings with robust word recognition. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5582–5591, Florence, Italy. Association for Computational Linguistics.

Vyas Raina and Mark Gales. 2023. Sample attackability in natural language adversarial attacks. *arXiv preprint arXiv:2306.12043*.

Vyas Raina, Samson Tan, Volkan Cevher, Aditya Rawal, Sheng Zha, and George Karypis. 2024. Extreme miscalibration and the illusion of adversarial robustness. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2500–2525, Bangkok, Thailand. Association for Computational Linguistics.

Sudhanshu Ranjan, Chung-En Sun, Linbo Liu, and Tsui-Wei Weng. 2023. Fooling GPT with adversarial in-context examples for text classification. In *R0-FoMo:Robustness of Few-shot and Zero-shot Learning in Large Foundation Models*.

Steve Rathje, Dan-Mircea Mirea, Ilia Sucholutsky, Raja Marjieh, Claire E Robertson, and Jay J Van Bavel. 2024. Gpt is an effective tool for multilingual psychological text analysis. *Proceedings of the National Academy of Sciences*, 121(34):e2308950121.

Qibing Ren, Liangliang Shi, Lanjun Wang, and Junchi Yan. 2022. Adversarial robustness via adaptive label smoothing.

Shuhuai Ren, Yihe Deng, Kun He, and Wanxiang Che. 2019. Generating natural language adversarial examples through probability weighted word saliency. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 1085–1097.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2018. Semantically equivalent adversarial rules for debugging nlp models. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (volume 1: long papers)*, pages 856–865.

Elias Abad Rocamora, Yongtao Wu, Fanghui Liu, Grigorios G Chrysos, and Volkan Cevher. 2024. Revisiting character-level adversarial attacks. *arXiv preprint arXiv:2405.04346*.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan LeBras, and Yejin Choi. 2019. Socialiqa: Commonsense reasoning about social interactions. *arXiv preprint arXiv:1904.09728*.

Erfan Shayegani, Md Abdullah Al Mamun, Yu Fu, Pedram Zaree, Yue Dong, and Nael Abu-Ghazaleh. 2023. Survey of vulnerabilities in large language models revealed by adversarial attacks. *arXiv preprint arXiv:2310.10844*.

Chenglei Si, Zhengyan Zhang, Fanchao Qi, Zhiyuan Liu, Yasheng Wang, Qun Liu, and Maosong Sun. 2021. Better robustness by more coverage: Adversarial and mixup data augmentation for robust finetuning. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1569–1576, Online. Association for Computational Linguistics.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1631–1642.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. 2015. Rethinking the inception architecture for computer vision. corr abs/1512.00567 (2015).

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. CommonsenseQA: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4149–4158, Minneapolis, Minnesota. Association for Computational Linguistics.

A Vaswani. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*.

Hetvi Waghela, Sneha Rakshit, and Jaydip Sen. 2024. A modified word saliency-based adversarial attack on text classification models. *arXiv preprint arXiv:2403.11297*.

Boxin Wang, Shuohang Wang, Yu Cheng, Zhe Gan, Ruoxi Jia, Bo Li, and Jingjing Liu. 2021. Infobert: Improving robustness of language models from an information theoretic perspective. In *International Conference on Learning Representations*.

Jia Wang, Min Gao, Zongwei Wang, Chenghua Lin, Wei Zhou, and Junhao Wen. 2022a. Ada: Adversarial learning based data augmentation for malicious users detection. *Applied Soft Computing*.

Jiayi Wang, Rongzhou Bao, Zhuosheng Zhang, and Hai Zhao. 2022b. Rethinking textual adversarial defense for pre-trained language models. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:2526–2540.

Zhaoyang Wang, Zhiyue Liu, Xiaopeng Zheng, Qinliang Su, and Jiahai Wang. 2023. Rmlm: A flexible defense framework for proactively mitigating word-level adversarial attacks. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2757–2774.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pages 38–45.

Yue Xu and Wenjie Wang. 2024. Linkprompt: Natural and universal adversarial attacks on prompt-based language models. *arXiv preprint arXiv:2403.16432*.

Kexin Yang, Dayiheng Liu, Wenqiang Lei, Baosong Yang, Xiangpeng Wei, Zhengyuan Liu, and Jun Xie. 2023a. Fantastic expressions and where to find them: Chinese simile generation with multiple constraints. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 468–486, Toronto, Canada. Association for Computational Linguistics.

Yahan Yang, Soham Dan, Dan Roth, and Insup Lee. 2023b. In and out-of-domain text adversarial robustness via label smoothing. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 657–669, Toronto, Canada. Association for Computational Linguistics.

Yuting Yang, Pei Huang, Juan Cao, Jintao Li, Yun Lin, and Feifei Ma. 2024. A prompt-based approach to adversarial example generation and robustness enhancement. *Frontiers of Computer Science*, 18(4):184318.

Jin Yong Yoo and Yanjun Qi. 2021. Towards improving adversarial training of NLP models. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 945–956, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yuan Zang, Fanchao Qi, Chenghao Yang, Zhiyuan Liu, Meng Zhang, Qun Liu, and Maosong Sun. 2020. Word-level textual adversarial attacking as combinatorial optimization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6066–6080, Online. Association for Computational Linguistics.

Guoyang Zeng, Fanchao Qi, Qianrui Zhou, Tingji Zhang, Zixian Ma, Bairu Hou, Yuan Zang, Zhiyuan Liu, and Maosong Sun. 2021. OpenAttack: An open-source textual adversarial attack toolkit. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 363–371, Online. Association for Computational Linguistics.

Jiehang Zeng, Jianhan Xu, Xiaoqing Zheng, and Xuanjing Huang. 2023. Certified robustness to text adversarial attacks by randomized [mask]. *Computational Linguistics*, 49(2):395–427.

Pengwei Zhan, Jing Yang, He Wang, Chao Zheng, Xiao Huang, and Liming Wang. 2023. Similarizing the influence of words with contrastive learning to defend word-level adversarial text attack. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 7891–7906, Toronto, Canada. Association for Computational Linguistics.

Zeliang Zhang, Wei Yao, Susan Liang, and Chenliang Xu. 2024. Random smooth-based certified defense against text adversarial attack. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 1251–1265, St. Julian's, Malta. Association for Computational Linguistics.

Jiahao Zhao, Wenji Mao, and Daniel Dajun Zeng. 2024. Disentangled text representation learning with information-theoretic perspective for adversarial robustness. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.

Zhengli Zhao, Dheeru Dua, and Sameer Singh. 2018. Generating natural adversarial examples. In *International Conference on Learning Representations*.

Rui Zheng, Rong Bao, Qin Liu, Tao Gui, Qi Zhang, Xuanjing Huang, Rui Xie, and Wei Wu. 2022. PlugAT: A plug and play module to defend against textual adversarial attack. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 2873–2882, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Qihuang Zhong, Liang Ding, Juhua Liu, Bo Du, and Dacheng Tao. 2023a. Can chatgpt understand too? a comparative study on chatgpt and fine-tuned bert. *arXiv preprint arXiv:2302.10198*.

Qihuang Zhong, Liang Ding, Juhua Liu, Bo Du, and Dacheng Tao. 2023b. Can chatgpt understand too? a comparative study on chatgpt and fine-tuned bert. *Preprint*, arXiv:2302.10198.

Yi Zhou, Xiaoqing Zheng, Cho-Jui Hsieh, Kai-Wei Chang, and Xuanjing Huang. 2021a. Defense against synonym substitution-based adversarial attacks via Dirichlet neighborhood ensemble. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5482–5492, Online. Association for Computational Linguistics.

Yi Zhou, Xiaoqing Zheng, Cho-Jui Hsieh, Kai-Wei Chang, and Xuanjing Huang. 2021b. Defense against synonym substitution-based adversarial attacks via dirichlet neighborhood ensemble. In *ACL*.

Bin Zhu and Yanghui Rao. 2023. Exploring robust overfitting for pre-trained language models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 5506–5522.

Chen Zhu, Yu Cheng, Zhe Gan, Siqi Sun, Tom Goldstein, and Jingjing Liu. 2020. Freelb: Enhanced adversarial training for natural language understanding. In *International Conference on Learning Representations*.

Kaijie Zhu, Jindong Wang, Jiaheng Zhou, Zichen Wang, Hao Chen, Yidong Wang, Linyi Yang, Wei Ye, Yue Zhang, Neil Zhenqiang Gong, et al. 2023. Promptbench: Towards evaluating the robustness of large language models on adversarial prompts. *arXiv preprint arXiv:2306.04528*.

A Generative LLMs

With the advent of powerful generative large language models (LLMs), such as ChatGPT (OpenAI, 2023), their usage has become increasingly widespread. However, similar to recent studies (Zhong et al., 2023a; Raina et al., 2024; Periti et al., 2024), we find that these popular generative LLMs are not suitable for inclusion in our benchmark for several key reasons. A comparative analysis of their performance with state-of-the-art generative LLMs, using 0-shot and few-shot prompting, is presented in Table 7 for some datasets considered in this paper. First, fine-tuned encoder-based models (e.g., BERT-based models) continue to demonstrate competitive, if not superior, performance on each task, which has led to their extensive adoption in many industry applications. These models are not only lightweight (possessing far fewer parameters compared to generative LLMs) but also cost-effective while achieving strong performance across a wide range of tasks. Second, the use of generative LLMs

Model	Params	SST2 (%)	MR (%)	MRPC (%)
Mistral-7B (0-shot)†	7B	-	86.47	67.15
Mistral-7B (5-shot)†	7B	-	88.92	76.21
ChatGPT-3.5 (0-shot)‡	-	92.00	-	66.00
ChatGPT-3.5 (1-shot)‡	-	96.00	-	66.00
ChatGPT-3.5 (5-shot)‡	-	98.00	-	76.00
ChatGPT-3.5 (0-shot CoT)‡	-	96.00	-	78.00
BERT-base	110M	91.54	85.55	84.40
RoBERTa-base	110M	93.79	88.65	86.96
DeBERTa-base	110M	95.22	90.71	87.94
Phi3-3.8B (0-shot)	3.8B	85.93	81.57	74.01
Phi3-3.8B (0-shot CoT)	3.8B	87.19	83.44	74.29
Llama3-8B (0-shot)	8B	89.46	83.80	76.90
Llama3-8B (0-shot CoT)	8B	90.12	84.13	78.49

Table 7: Comparison of model performance with popular generative LLMs. † Figures given in Raina et al. (2024). ‡ Figures given in Zhong et al. (2023b).

complicates the standardisation of input-output formats across diverse tasks and datasets, which introduces potential bias into the evaluation process (Liu et al., 2023). This challenge makes it difficult to ensure reproducibility and to facilitate fair comparisons across different research contexts (Hayase et al., 2024; Rathje et al., 2024). Moreover, disentangling a model's intrinsic performance from artifacts introduced by the prompting strategy is non-trivial, as model outcomes are influenced by both the design of the prompts and the generated responses (Gao et al., 2021; Liu et al., 2023). Finally, the adversarial attack and defence literature, which forms the basis of our contributions, predominantly focuses on encoder-based models. Aligning our experimental setup with this body of work enables us to build on existing attack and defence mechanisms.

In Table 7, we provide a comparative analysis of the performance of several state-of-the-art generative LLMs, utilising both zero-shot and five-shot prompting. Additionally, we present performance comparisons from Zhong et al. (2023b) and Raina et al. (2024), and evaluate encoder-based models (BERT, RoBERTa, and DeBERTa) alongside generative LLMs, including Phi3 (Abdin et al., 2024) and Llama3 (Dubey et al., 2024), on some of the datasets covered in this work.

B Detailed Performance Breakdown

In this section, we provide the detailed breakdown of performances for the different Transformer encoders across each dataset: Table 8 for BERT-base model, Table 9 for RoBERTa-base model, and Table 10 for DeBERTa-base model. Each Table presents the adversarial robustness performance trained with different defence methods. The best performance is marked in **bold**.

Dataset	Defence	Acc↑	TextFooler (%)			TextBugger (%)			Apdr↓
			Aua↑	Asr↓	AvgQ↑	Aua↑	Asr↓	AvgQ↑	
SST2	-	91.54	6.59	92.80	89.53	28.08	69.35	41.30	81.06
	PGD	**92.64**	8.73	90.57	98.54	34.27	63.01	42.74	76.79
	FreeLB	91.98	8.57	90.69	98.95	31.80	65.43	42.35	78.06
	TAVAT	**92.64**	10.71	88.44	103.78	34.32	62.95	43.21	75.70
	Flooding-X	89.84	11.64	87.04	95.38	30.37	66.20	43.08	76.62
	SLS	91.76	12.25	86.65	108.85	39.21	57.27	44.74	71.96
	ALS	91.21	15.54	82.96	110.51	39.37	56.83	46.48	69.90
	TTSO	91.71	**41.63**	**54.61**	**148.27**	**50.85**	**44.55**	**95.83**	**49.58**
MR	-	85.55	3.94	95.39	82.70	22.80	73.36	42.85	84.37
	PGD	85.93	12.85	85.04	115.93	35.46	58.73	49.77	71.89
	FreeLB	86.30	6.66	92.28	99.94	28.71	66.74	45.51	79.51
	TAVAT	86.02	8.26	90.40	107.21	30.11	64.99	47.04	77.70
	Flooding-X	85.55	5.44	93.64	91.97	27.02	68.42	44.77	81.03
	SLS	**86.59**	13.60	84.29	118.28	36.49	57.85	47.95	71.08
	ALS	85.83	11.07	87.10	112.63	33.77	60.66	49.31	73.88
	TTSO	86.02	**35.27**	**59.00**	**157.94**	**43.06**	**49.95**	**103.85**	**54.47**
MRPC	-	84.40	2.32	97.25	124.00	3.25	96.15	72.84	96.70
	PGD	84.06	9.86	88.28	205.38	11.25	86.62	101.98	87.44
	FreeLB	**85.45**	11.48	86.57	212.41	11.65	86.36	107.19	86.47
	TAVAT	84.29	8.70	89.68	229.16	10.43	87.62	106.60	88.65
	Flooding-X	82.67	7.48	90.95	151.69	7.88	90.46	86.56	90.71
	SLS	84.52	6.32	92.52	170.88	7.36	91.29	88.79	91.91
	ALS	82.96	5.97	92.80	168.67	7.83	90.57	93.89	91.68
	TTSO	83.77	**41.62**	**50.31**	**370.89**	**39.71**	**52.60**	**220.98**	**51.46**
SciTail	-	92.80	44.45	52.10	106.17	32.22	65.28	95.52	58.69
	PGD	93.09	50.19	46.08	111.88	32.69	64.88	95.07	55.48
	FreeLB	**93.60**	47.04	49.75	109.00	32.60	65.18	96.17	57.46
	TAVAT	92.29	**52.21**	**43.43**	115.19	30.48	66.97	98.18	55.20
	Flooding-X	91.58	49.62	45.81	111.23	35.28	61.48	102.06	53.65
	SLS	92.33	48.02	47.99	110.77	34.85	62.25	94.86	55.12
	ALS	92.57	50.80	45.12	112.64	33.44	63.87	97.90	54.50
	TTSO	92.33	52.02	43.66	**123.38**	**48.21**	**47.78**	**157.87**	**45.72**

Table 8: The experiment results of different defence methods using BERT-base model.

Dataset	Defence	Acc↑	TextFooler			TextBugger			Apdr↓
			Aua↑	Asr↓	AvgQ↑	Aua↑	Asr↓	AvgQ↑	
SST2	-	93.79	7.80	91.69	89.15	31.52	66.39	43.31	79.04
	PGD	94.29	8.79	90.68	97.68	34.21	63.72	44.28	77.20
	FreeLB	95.11	6.43	93.24	94.88	36.90	61.20	44.23	77.22
	TAVAT	**95.17**	9.23	90.31	100.21	37.40	60.70	44.25	75.50
	Flooding-X	94.95	5.44	94.27	85.27	29.65	68.77	41.99	81.52
	SLS	94.34	11.64	87.66	111.72	44.81	52.50	45.69	70.08
	ALS	94.51	23.50	75.13	130.85	52.50	44.45	49.03	59.79
	TTSO	94.67	**45.63**	**51.80**	**159.93**	**56.84**	**39.97**	**101.98**	**45.88**
MR	-	88.65	6.75	92.38	101.06	31.80	64.13	48.91	78.26
	PGD	87.90	6.29	92.85	102.56	34.33	60.94	49.29	76.89
	FreeLB	**89.12**	5.53	93.79	98.94	33.58	62.32	47.84	78.06
	TAVAT	87.99	6.19	92.96	102.04	32.46	63.11	48.46	78.04
	Flooding-X	88.74	7.79	91.23	99.61	31.05	65.01	46.73	78.12
	SLS	88.09	20.83	76.36	136.77	42.59	51.65	53.54	64.00
	ALS	87.90	17.07	80.58	130.26	43.34	50.69	53.59	65.64
	TTSO	87.71	**42.40**	**51.66**	**171.72**	**51.97**	**40.75**	**114.56**	**46.20**
MRPC	-	86.96	6.09	93.00	163.57	9.57	89.00	96.71	91.00
	PGD	87.48	5.28	93.97	180.35	11.54	86.81	102.54	90.39
	FreeLB	87.54	6.38	92.72	191.58	11.71	86.62	105.26	89.67
	TAVAT	**87.59**	9.97	88.62	212.02	15.07	82.79	108.67	85.71
	Flooding-X	87.48	4.75	94.57	173.10	8.41	90.39	101.31	92.48
	SLS	86.84	9.22	89.39	208.54	12.35	85.78	111.62	87.58
	ALS	86.03	10.61	87.67	220.83	13.28	84.57	108.99	86.12
	TTSO	86.78	**46.38**	**46.56**	**389.29**	**41.91**	**51.70**	**226.51**	**49.13**
SciTail	-	93.60	42.80	54.27	101.92	31.70	66.13	92.84	60.20
	PGD	93.09	43.27	53.51	104.29	31.42	66.25	95.29	59.88
	FreeLB	**93.79**	44.21	52.86	104.06	31.51	66.40	96.19	59.63
	TAVAT	**93.79**	46.38	50.55	106.96	34.71	62.99	98.50	56.77
	Flooding-X	92.43	43.32	53.13	104.57	29.02	68.60	92.31	60.87
	SLS	93.60	45.58	51.31	106.60	35.79	61.76	103.88	56.53
	ALS	92.90	44.17	52.46	105.72	34.48	62.89	93.94	57.67
	TTSO	92.52	**51.27**	**44.59**	**121.66**	**47.22**	**48.96**	**158.72**	**46.77**

Table 9: The experiment results of different defence methods using RoBERTa-base model.

Dataset	Defence	Acc↑	TextFooler			TextBugger (%)			Apdr↓ (%)
			Aua↑	Asr↓	AvgQ↑	Aua↑	Asr↓	AvgQ↑	
SST2	-	95.22	8.57	91.00	96.80	39.65	58.36	44.45	74.68
	PGD	94.67	8.73	90.78	96.96	38.88	58.93	44.15	74.85
	FreeLB	95.22	10.32	89.16	108.13	44.10	53.69	45.09	71.42
	TAVAT	**95.83**	11.04	88.48	111.03	47.01	50.95	46.74	69.71
	Flooding-X	95.22	7.36	92.27	92.93	31.74	66.67	42.09	79.47
	SLS	95.22	22.24	76.64	130.72	54.48	42.79	47.70	59.71
	ALS	95.55	15.76	83.51	120.09	52.22	45.34	49.00	64.43
	TTSO	95.66	**55.02**	**42.48**	**173.84**	**65.84**	**31.17**	**109.65**	**36.83**
MR	-	90.71	9.94	89.04	101.00	34.24	62.25	47.71	75.65
	PGD	89.21	5.35	94.01	96.21	30.11	66.25	47.00	80.13
	FreeLB	91.18	8.63	90.53	105.07	34.80	61.83	48.37	76.18
	TAVAT	90.90	10.60	88.34	110.67	38.37	57.79	50.01	73.06
	Flooding-X	**91.37**	6.66	92.71	98.27	35.46	61.19	47.03	76.95
	SLS	90.62	17.45	80.75	135.75	45.31	50.00	52.15	65.37
	ALS	89.96	17.92	80.08	132.54	46.53	48.28	53.05	64.18
	TTSO	90.34	**47.84**	**47.04**	**179.78**	**56.75**	**37.18**	**117.93**	**42.11**
MRPC	-	87.94	2.96	96.64	155.80	9.86	88.79	97.04	92.71
	PGD	87.83	3.71	95.78	170.55	10.55	87.99	98.00	91.88
	FreeLB	86.43	8.58	90.07	186.01	15.07	82.56	107.91	86.32
	TAVAT	88.06	6.61	92.50	194.10	16.23	81.57	110.88	87.03
	Flooding-X	88.35	3.30	96.26	156.63	10.20	88.45	97.65	92.36
	SLS	**88.46**	8.58	90.30	158.34	12.87	85.45	101.25	87.88
	ALS	**88.46**	6.14	93.05	182.10	17.45	80.28	111.49	86.67
	TTSO	87.54	**50.78**	**41.99**	**394.67**	**54.38**	**37.88**	**253.36**	**39.94**
SciTail	-	95.53	47.04	50.76	107.60	33.82	64.60	85.38	57.68
	PGD	94.36	43.32	54.09	106.32	30.39	67.80	89.40	60.94
	FreeLB	95.67	47.04	50.84	107.88	33.68	64.80	90.24	57.81
	TAVAT	**96.52**	50.19	48.00	112.98	39.98	58.58	98.80	53.29
	Flooding-X	94.73	46.28	51.14	106.81	32.22	65.99	85.41	58.57
	SLS	95.67	48.64	49.16	110.09	35.04	63.37	92.54	56.27
	ALS	95.16	50.85	46.56	116.84	43.09	54.72	103.17	50.64
	TTSO	95.63	**55.13**	**42.35**	**126.89**	**50.05**	**47.66**	**164.70**	**45.01**

Table 10: The experiment results of different defence methods using DeBERTa-base model.