Assessing Adversarial Robustness of Large Language Models: An Empirical Study

Zeyu Yang
zeyu.yang@telepathy.ai
Telepathy Labs
Zürich, Switzerland

Xiaochen Zheng
xzheng@ethz.ch
ETH Zürich
Zürich, Switzerland

Zhao Meng
zhmeng@ethz.ch
ETH Zürich
Zürich, Switzerland

Roger Wattenhofer
wattenhofer@ethz.ch
ETH Zürich
Zürich, Switzerland

Abstract

Large Language Models (LLMs) have revolutionized natural language processing, but their robustness against adversarial attacks remains a critical concern. We presents a novel white-box style attack approach that exposes vulnerabilities in leading open-source LLMs, including Llama, OPT, and T5. We assess the impact of model size, structure, and fine-tuning strategies on their resistance to adversarial perturbations. Our comprehensive evaluation across five diverse text classification tasks establishes a new benchmark for LLM robustness. The findings of this study have far-reaching implications for the reliable deployment of LLMs in real-world applications and contribute to the advancement of trustworthy AI systems.

CCS Concepts

• **Computing methodologies** → **Natural language processing**.

Keywords

Evaluation, Robustness, Adversarial Attack, Scaling Law, LLMs Fine-tuning

ACM Reference Format:
Zeyu Yang, Xiaochen Zheng, Zhao Meng, and Roger Wattenhofer. 2024. Assessing Adversarial Robustness of Large Language Models: An Empirical Study. In . ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 Introduction

In recent years, the field of artificial intelligence has witnessed a remarkable surge in the development and application of Large Language Models (LLMs). These models, such as ChatGPT [23], GPT-4 [22], and Llama-2 [29], have demonstrated exceptional performance in various natural language understanding and generation tasks [42]. The success of LLMs can be attributed to the innovative training techniques employed, including instruction tuning, prompt tuning, Low-Rank Adaptor (LoRA) [6, 11]. These advances have made it possible to fine-tune and infer models like Llama-2-7B on consumer-level devices, thereby increasing their accessibility and potential for integration into daily life.

However, despite their impressive capabilities, LLMs are not without limitations. One significant challenge is their susceptibility to variations in input types, which can lead to inconsistencies in output and potentially undermine their reliability in real-world applications. For example, when faced with ambiguous or provocative prompts, LLMs may generate inconsistent or inappropriate responses. To address this issue, several studies have been conducted to assess the robustness of LLM models [35, 44]. However, these efforts often overlook the importance of re-fine-tuning the models and conducting comprehensive studies of adversarial attacks with known adversarial sample generation mechanisms when full access to the model weights, architecture, and training pipeline is available [10, 30].

In this paper, we present an extensive study of three leading open-source LLMs: Llama, OPT, and T5. We evaluate the robustness of various sizes of these models across five distinct NLP classification datasets. To assess their vulnerability to input perturbations, we employ the adversarial geometry attack technique and measure the impact on model accuracy. Furthermore, we investigate the effectiveness of commonly used methods in LLM training, such as LoRA, different precision levels, and variations in model architecture and tuning approaches.

Our work makes several notable contributions to the field of LLM evaluation and robustness:

(1) We introduce a novel white-box style attack approach that leverages output logits and gradients to expose potential vulnerabilities and assess the robustness of LLMs.
(2) We establish a benchmark for evaluating the robustness of LLMs by focusing on their training strategies, setting the stage for future research in this domain.
(3) Our comprehensive evaluation spans five text classification tasks, providing a broad perspective on the capabilities and limitations of the models across diverse applications.

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

2 Related Work

2.1 The Evaluation of LLMs

In recent years, the LLM domain has experienced significant advances [5, 36]. A large number of exemplary large-scale models such as GPT-4, have emerged, showcasing exceptional performance across various sectors.

Given the remarkable capabilities and broad applications of these models, evaluating their performance has become paramount. Consequently, a significant portion of the research is dedicated to NLP tasks. For instance, [35] examines ChatGPT's performance in sentiment analysis, while [40] offers a comparative analysis with other LLMs. Numerous studies have explored the capabilities of LLM in natural language understanding, including text classification as highlighted by [15], and inference as demonstrated by [25]. Additionally, extensive research has been conducted to assess LLMs in generation tasks, encompassing areas such as translation [34], question answering [3, 15], and summarization [4].

Due to the impressive performance of LLMs, there has been widespread attention to their safety and stability. [33] conducted an early investigation into ChatGPT and other LLMs and utilized existing benchmarks like AdvGLUE [32], ANLI [21], and DDXPlus [28] for their evaluations. [43] assessed the performance of LLMs in visual inputs and their transferability to other visual-language models. On the topic of adversarial robustness, [31] introduced the AdvGLUE++ benchmark and proposed an approach to examine machine ethics through system prompts. [44] proposed a unified benchmark named PromptBench to evaluate the resilience of LLMs against prompts.

However, a significant limitation of the aforementioned studies on robustness is that they focus only on inference-based evaluations. They largely overlook the intricacies of the model's weights and output logits. Furthermore, these studies do not discuss the performance of various LLM techniques. In contrast, in our research, we not only conduct attacks based on the model's parameters and logits, but also actively participate in the model's tuning. Additionally, we place a special emphasis on studying the model size and specific techniques used in LLMs, aspects that previous works have not addressed.

2.2 Robustness in NLP

With the rapid advancement in NLP research, its applications have become increasingly prevalent. This ubiquity underscores the growing need for reliable NLP systems that can effectively counteract malign content and misinformation. A seminal work [12] highlighted the vulnerabilities of NLP systems to adversarial attacks.

There are some works about various input perturbations, which could be categorized into three groups: character level, word level, and sentence level [13, 16, 37]. At the character level, adversarial attacks focus on altering individual characters within a given text. [9] introduced HotFlip, utilizing gradient information to manipulate characters within text. [14] took a different approach by identifying words and modifying their characters. At the word level, adversarial strategies revolve around replacing specific words within the content. For instance, [1] employs evolutionary algorithms to swap out words with their synonyms. [38] utilized probabilistic sampling to generate adversarial examples. Furthermore, some researchers have explored adversarial tactics at the sentence level. [12] suggested a method that introduces an extraneous sentence to the primary content, aiming to mislead reading models. On the other hand, [24] adopted a new approach, where they employed an encoder-decoder framework to rephrase entire sentences. Recently, [31] evaluated the robustness and trustworthiness of GPT-3.5 and GPT-4 models, revealing vulnerabilities such as the ease of generating toxic and biased outputs and leaking private information. Despite GPT-4's improved performance on standard benchmarks, it is more susceptible to adversarial prompts, highlighting the need for rigorous trustworthiness guarantees and robust safeguards against new adaptive attacks.

However, the landscape of NLP research is ever-evolving. With the introduction of more sophisticated models boasting novel architectures and training methodologies, there is an growing need to assess the robustness of these newer models. This is especially true for LLMs, which present unique challenges and opportunities in the realm of robustness research.

3 Preliminaries

3.1 Open-source Large Language Models

We evaluate the following open-source large language models used in our experiments, as shown in Table 1.

T5	The Text-to-Text Transfer Transformer (T5) developed by Google Research redefines natural language processing (NLP) tasks by treating them uniformly as text-to-text conversions [26].
OPT	Open Pretrained Transformers (OPT) range from 125M to 175B parameters and are decoder-only models [39]. These models are trained on a diverse pre-training corpus.
Llama	Meta AI's Llama is a series of models ranging from 7B to 65B parameters. Llama is trained on a corpus comprising trillions of tokens from publicly available datasets.

Table 1: The open-source large language models in our study

3.2 Fine-tuning Techniques

We apply the following fine-tuning techniques in our study.

LoRA. LoRA innovates in fine-tuning pretrained language models for specific tasks, addressing the inefficiency of full fine-tuning in increasingly large models. By inserting trainable rank decomposition matrices into each layer and freezing the original model weights, LoRA significantly reduces the number of parameters requiring training.

Quantization. Quantization in large language models (LLMs) reduces the model size by lowering weight precision, with 8-bit precision presenting challenges due to errors from quantizing large-value vectors. These errors are pronounced in transformer architectures, requiring mixed-precision decomposition. This involves identifying outliers using a threshold, processing them in fp16, and quantizing the rest of the matrix at 8-bit precision. The two parts are then combined. The approach, exemplified by `LLM.int8()`, aims to make large models more accessible, trading off some performance for significant size reduction.

QLoRA. Quantized Low-Rank Adapters (QLoRA) [7] introduce an efficient technique to fine-tune large language models by significantly lowering memory requirements. QLoRA combines 4-bit quantization with Low-Rank Adapters, freezing the parameters of a compressed pretrained language model.

4 Methods

4.1 Adversarial Attack

In this study, our primary concern is text classification. Following the work by [18], we consider a sample sentence $S_i = \{w_1, w_2, \ldots, w_L\}$ containing L words, and its corresponding category label is y_i. Our textual classification system is built upon n LLMs, represented as $f(\cdot)$, coupled with a prompt indicating the categorization task, denoted as P_i. In a formal sense: $\hat{y}_i = f(S_i; P_i)$, where $\hat{y}_i$ stands for the given answer. The prediction is accurate when $\hat{y}_i$ equals y_i.

An adversarial attack based on word replacement processes the original sample S_i to produce an adversarial version S_i^{adv} by replacing the k-th word w_k in S_i with an alternative word w_{adv_k}. To ensure that the original sample S_i and its adversarial counterpart S_i^{adv} maintain semantic similarity, prevalent methodologies typically employ synonymous terms for replacements.

4.2 Geometry Attack Methodology

In our research, we extend the basic principles of adversarial attacks in the context of LLMs. Our focus is on exploiting geometric attacks [18, 19] to assess the vulnerability of LLMs to adversarial perturbations. We propose a systematic methodology grounded in geometric attack insights. The following sections detail the steps of our approach:

1) Gradient Computation for Influence Analysis: We commence by calculating the gradients of the generation loss $\mathcal{L}_i$ with respect to the embeddings e_i of input sentence S_i. The cross entropy loss $\mathcal{L}_i$ measures the dissimilarity between the prediction and label examples in the output space. This computation is essential for all words, including those segmented into sub-tokens. For such words, gradients are computed for each sub-token and subsequently averaged. This initial step is crucial for identifying the words that exert significant influence on $\mathcal{L}_i$. We determine the gradient of $\mathcal{L}_i$ with respect to the embedding vector e_i. This step determines the direction in which e_i should be adjusted to maximize the increase in the loss $\mathcal{L}_i$. The resulting gradient vector is denoted as $v_{e_i} = \nabla_{e_i} \mathcal{L}_i$.

2) Selection of Candidate Words: Suppose we select a target word w_t from step 1. Utilizing the DeepFool algorithm [20], we identify potential replacement words, forming a candidate set $\{w_{t_1}, w_{t_2}, \ldots, w_{t_T}\}$. Candidates are filtered based on their cosine similarity to w_t, with those below a defined threshold ϵ being excluded. This process ensures that only semantically similar and relevant candidates are considered.

3) Optimal Word Replacement and Projection Analysis: After replacing w_t with each candidate word, we compute the new text vectors $\{e_{i_1}, e_{i_2}, \ldots, e_{i_T}\}$. For each vector, we define the delta vector r_{i_j} as $e_{i_j} - e_i$. The projection of r_{i_j} onto v_{e_i} is calculated as $p_i = r_{i_j} \cdot v_{e_i}$. The optimal replacement candidate w_{t_m} is selected based on criterion $m = \arg\max_j \frac{|p_{i_j}|}{||v_{e_i}||}$. This ensures that the chosen word w_{t_m} induces the largest possible projection p_{i_m} onto the gradient vector v_{e_i}.

4) Iterative Process for Enhanced Adversarial Strength: The selected word w_{t_m} replaces w_t in S_i, updating e_i to e_{i_m}. This iterative procedure is repeated for N cycles, where N is an adjustable parameter in our methodology. Throughout these iterations, an increase in $\mathcal{L}_i$ should be observed, indicating a continuous enhancement in the adversarial effectiveness of the altered input.

Through this methodically structured process, our research aims to uncover and analyze potential vulnerabilities in LLMs. We refined our methodology to enable prompt fine-tuning for attack generation tasks, expanding its application beyond the previously limited scope of classification tasks.

5 Experiment Settings

5.1 Experiment Pipeline

This section introduces our methodology for evaluating the robustness of pre-trained LLMs against adversarial attacks. The procedure comprises three principal stages:

1) Model Fine-Tuning: We fine-tune a pre-trained language model on target dataset with different fine-tuning techniques as described in Sec. 3.2, evaluating its accuracy on the corresponding validation set to establish a performance baseline.

2) Adversarial Attack Assessment: The fine-tuned model undergoes adversarial attacks described in Sec. 4.2, and its performance is assessed on a test dataset altered with adversarial examples.

3) Robustness Evaluation: We compare the model's accuracy before and after the adversarial attacks to assess its robustness and vulnerability to such manipulations.

5.2 Datasets

To evaluate the model's performance under various tasks and its resilience to attacks, we employed five classification datasets, categorized into binary and multi-class classifications. For binary classification, the datasets include IMDB [17], MRPC [8], and SST-2 [27], and for multiclass classification, AGNews [41] and DBpedia [2] are used. We will provide a more detailed introduction to these tasks/datasets in Appendix A.2

5.3 Evaluation Metrics

We assess our model's robustness and efficacy using four principal metrics, as described in Table 2.

6 Experimental Results

In this section, we conduct extensive experiments to evaluate the robustness of LLMs across five different datasets. These investigations are guided by three key research questions (RQ):

RQ1: How does the robustness of variously sized models differ under adversarial attacks across distinct tasks?

RQ2: Do contemporary training techniques for LLMs influence their performance and robustness?

RQ3: How does the model architecture (e.g., fine-tuning with a classification head vs. prompt tuning), affect the robustness of the model?

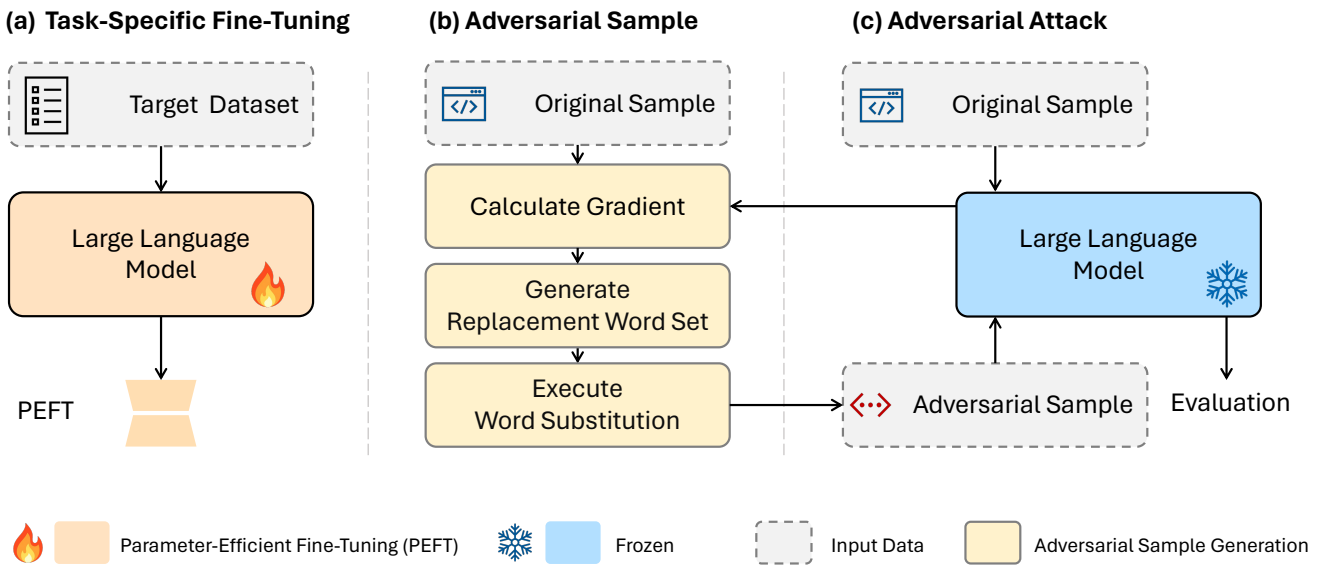


Figure 1: The framework of our adversarial robustness assessment

Metrics	Description
Acc	Accuracy: the model's correct classification rate of untouched input
Acc/attack	Accuracy Under Attack: post-attack classification accuracy revealing adversarial defense
ASR	Attack Success Rate: the frequency of accurate predictions turned false by attacks
Replacement	Replacement Rate: the extent of input alteration needed to change the model's prediction, indicating sensitivity to perturbations

Table 2: The evaluation metrics

6.1 Model Size (RQ1)

In this section, we analyze the performance metrics of various models across multiple tasks. The datasets under examination include IMDB, SST-2, MRPC, AGNews, and DBPedia. We measure the performance and robustness of LLMs with the metrics Acc, Acc/attack, ASR, and Replacement Rate described in Table 2.

The results from the IMDB dataset in Table 3 reveal distinct performance variations among different model architectures. In the T5 Series, accuracy generally improves with increasing model size, from 60m to 11b parameters, but the relationship is nonlinear. This suggests that while larger models tend to be more accurate, the accuracy does not increase uniformly with model size. Furthermore, the resilience of these models to adversarial attacks does not follow a simple inverse relationship with model size. The larger T5-11b model shows a more noticeable decrease in accuracy under attack conditions.

For the OPT models, a similar upward trend in accuracy is observed with increasing model size, but the Attack Success Rate (ASR) is lower, suggesting better resistance to attacks. In comparison, the Llama models demonstrate superior performance in both accuracy and robustness against attacks.

	Acc	Acc/attack ↑	ASR ↓	Replacement
T5-60m	0.8484	0.1256	0.8491	0.0929
T5-220m	0.8011	0.0436	0.9463	0.0722
T5-770m	0.9048	0.1536	0.8312	0.1143
T5-3b	0.9146	0.3259	0.6436	0.1413
T5-11b	0.9122	0.3098	0.6604	0.1330
OPT-125m	0.8616	0.6637	0.2297	0.0365
OPT-350m	0.8564	0.6924	0.1915	0.0305
OPT-1.3b	0.9231	0.7515	0.1859	0.0421
OPT–2.7b	0.9198	0.7651	0.1682	0.0396
OPT-6.7b	0.9408	0.7864	0.1641	0.0528
OPT-13b	0.9431	0.8016	0.1500	0.0671
Llama-7b	0.9483	0.8203	0.1350	0.0816
Llama-13b	0.9472	0.8237	0.1304	0.0875

Table 3: IMDB Dataset Results

From Table 4 on the SST-2 dataset, there are distinct performance trends. The T5-11b, achieves the highest accuracy of 0.9656. However, its persistence to adversarial attacks is not highest. Notably, the highest ASR within the T5 series is recorded for the T5-770m model, indicating a trade-off as model size increases. In the case of the OPT series, the OPT-6.7b model stands out. However, similar to the IMDB dataset, this model also shows a significant decline in accuracy but more robust than T5 models. One more observation in the OPT series is the overall decrease in ASR with increasing model size, but this trend is disrupted at the 13b parameter mark, where an anomalous increase in ASR is observed. The Llama models, demonstrate consistently high accuracy. It also presents lower ASR

	Acc	Acc/attack ↑	ASR ↓	Replacement
T5-60m	0.9083	0.2419	0.7304	0.1428
T5-220m	0.8884	0.1228	0.8622	0.1611
T5-770m	0.8739	0.0534	0.9395	0.1785
T5-3b	0.9495	0.1563	0.8437	0.1987
T5-11b	0.9656	0.2248	0.7672	0.2043
OPT-125m	0.8807	0.7409	0.1587	0.0607
OPT-350m	0.9011	0.7716	0.1437	0.0598
OPT-1.3b	0.9443	0.8227	0.1288	0.0733
OPT–2.7b	0.8897	0.7921	0.1097	0.0476
OPT-6.7b	0.9693	0.8682	0.1043	0.0550
OPT-13b	0.9656	0.7867	0.1853	0.0719
Llama-7b	0.9683	0.8203	0.1528	0.0916
Llama-13b	0.9632	0.8124	0.1566	0.0877

Table 4: SST-2 Dataset Results

	Acc	Acc/attack ↑	ASR ↓	Replacement
T5-60m	0.8048	0.0325	0.9594	0.0720
T5-220m	0.8035	0.0633	0.9203	0.1100
T5-770m	0.8924	0.1992	0.7771	0.1203
T5-3b	0.8584	0.1074	0.8739	0.0917
T5-11b	0.8877	0.0712	0.9196	0.0873
OPT-125m	0.8321	0.6504	0.2184	0.0332
OPT-350m	0.8956	0.6741	0.2473	0.0437
OPT-1.3b	0.9134	0.7721	0.1547	0.0419
OPT–2.7b	0.9128	0.7854	0.1396	0.0533
OPT-6.7b	0.9096	0.7902	0.1313	0.0579
OPT-13b	0.9254	0.8183	0.1157	0.0560
Llama-7b	0.9277	0.8256	0.1101	0.0637
Llama-13b	0.9198	0.8107	0.1186	0.0742

Table 5: MRPC Dataset Results

	Acc	Acc/attack ↑	ASR ↓	Replacement
T5-60m	0.8606	0.3608	0.5807	0.1740
T5-220m	0.9084	0.5370	0.4098	0.1864
T5-770m	0.9278	0.6597	0.2896	0.1860
T5-3b	0.9193	0.7267	0.2102	0.1834
T5-11b	0.9212	0.8469	0.1531	0.1867
OPT-125m	0.8152	0.6040	0.2591	0.0809
OPT-350m	0.8321	0.6036	0.2746	0.0864
OPT-1.3b	0.8806	0.6316	0.2828	0.0912
OPT–2.7b	0.9175	0.7028	0.2340	0.0833
OPT-6.7b	0.9341	0.7143	0.2353	0.0756
OPT-13b	0.9456	0.7745	0.1809	0.0941
Llama-7b	0.9328	0.7315	0.2158	0.0864
Llama-13b	0.9338	0.7688	0.1767	0.0837

Table 6: AGNews Dataset Results

	Acc	Acc/attack ↑	ASR ↓	Replacement
T5-60m	0.9817	0.3974	0.5952	0.1187
T5-220m	0.9765	0.4082	0.5819	0.1234
T5-770m	0.9921	0.7476	0.2464	0.1483
T5-3b	0.9914	0.8608	0.1317	0.1550
T5-11b	0.9919	0.8815	0.1113	0.1724
OPT-125m	0.9034	0.6512	0.2792	0.0534
OPT-350m	0.9511	0.6718	0.2937	0.0305
OPT-1.3b	0.9784	0.7046	0.2798	0.0496
OPT–2.7b	0.9822	0.7513	0.2351	0.0552
OPT-6.7b	0.9907	0.7780	0.2147	0.0641
OPT-13b	0.9916	0.7912	0.2021	0.0576
Llama-7b	0.9908	0.7596	0.2333	0.0881
Llama-13b	0.9921	0.7286	0.2656	0.0912

Table 7: DBPedia Dataset Results

compared to T5 models but similar performance to OPT models. For SST-2, the ASR of T5 models exhibit a trend entirely contrary to that observed for MRPC. It reaches its minimum at the T5-770m model. For the OPT models, although their ASR is much lower compared to the T5 series, there is a consistent decrease in ASR as the size of the OPT models increases. Regarding the Llama models, the 7b model slightly outperforms the 13b in terms of accuracy and ASR.

In Tables 6 and Tables 7, analyzing results from multi-class classification tasks, a distinct pattern emerges. These datasets reveal enhanced stability against synonym substitution attacks. For T5 models, the data shows a lower ASR on these tasks compared to binary datasets, suggesting a better resistance to attacks in complex classification scenarios. In contrast, OPT and Llama models exhibit a higher ASR on the AGNews and DBpedia14 datasets. Another result is that for both T5 and OPT series, there is a marked decline in ASR around the 770 million or 1 billion parameter threshold. This indicates an increased robustness and better handling of adversarial attacks with the scale-up of model size.

6.1.1 Analysis. When examining the accuracy of the model, we observed a trend where the accuracy gradually increases with the growth in model size. However, after reaching a certain size threshold, the accuracy tends to saturate, stabilizing around specific values. This phenomenon is particularly pronounced when tested on datasets like DBpedia. When comparing different models operating at the same parameter scale, their performances were found to be quite similar, without any significant disparities.

However, the experiments related to robustness revealed more distinct differences. From Figure 2a, Figure 2b and Figure 2c, we have more intuitive results. Observing the performance of a uniform model across various datasets, we made the following observations: **T5 Model**: As the size of the T5 model increases, its ASR gradually decreases. This suggests that larger models, with more parameters, tend to have a deeper understanding of language. As a result, they can maintain stronger stability in the face of various disturbances. However, on datasets like MRPC and SST-2, there were noticeable fluctuations in performance. One possible explanation for this is that as the model size grows, the words selected based on the model's gradient become more precise and have a more significant impact on the results. This introduces a trade-off related to model size.

OPT Model: For the OPT model, a similar trend was observed across most datasets. As the model size increased, its robustness generally improved, aligning with the observations made for the T5 model.

Llama Model: For the Llama model, the differences in performance between the two sizes were minimal. This suggests that the size variation did not significantly influence the model's robustness.

However, when comparing different models, the disparities become even more pronounced. It is obvious that the T5 model's ASR and replacement rate are significantly higher than those of OPT and Llama. This indicated that Decoder-only Causal LMs have higher robustness against encoder-decoder architectures under synonym substitution adversarial attacks.

6.2 LLMs Fine-tuning Techniques (RQ2)

6.2.1 Instruction Tuning. To study the impact of instruction tuning on model robustness, we compared the performance of Flan-T5 with the standard T5. The Flan-T5 is an advanced variant of T5 that has undergone instruction tuning across over a thousand downstream tasks. In contrast, the traditional T5 was not trained with such an extensive procedure.

Based on our experimental results, as shown in the table, there is a significant decline in accuracy for both T5 and Flan-T5 under adversarial attacks. This observation indicates that models, irrespective of whether they have undergone instruction tuning, remain susceptible to adversarial manipulations. Furthermore, consistent with our previous findings, we noticed that as the model size increases, the attack success rate tends to decline.

Interestingly, as shown in Fig 3, our results indicated that Flan-T5 exhibits a higher ASR than the standard T5. This suggests that models subjected to instruction tuning, like Flan-T5, can be more easily compromised. We hypothesize the primary reason for this observation:

The instruction tuning process for Flan-T5 encompassed datasets similar to IMDB. This might have rendered the model with a deeper understanding of tasks related to this data. As a result, attackers could more easily pinpoint words in the input that were influential and susceptible to replacement.

	Acc	Acc/attack ↑	ASR ↓	Replacement
T5-60m	0.8484	0.1256	0.8491	0.0929
Flan-T5-60m	0.8453	0.0882	0.8968	0.0820
T5-220m	0.8011	0.0436	0.9463	0.0722
Flan-T5-220m	0.8777	0.0996	0.8862	0.0978
T5-770m	0.9048	0.1536	0.8312	0.1143
Flan-T5-770m	0.9141	0.1171	0.8729	0.1106
T5-3b	0.9146	0.3259	0.6436	0.1413
Flan-T5-3b	0.9228	0.2328	0.7489	0.1261
T5-11b	0.9348	0.4904	0.4752	0.1326
Flan-T5-11b	0.9122	0.3098	0.6604	0.1330

Table 8: Experimental results comparing T5 and Flan-T5 after instruction tuning using the IMDB dataset.

6.2.2 Precisions. In machine learning, balancing model size with precision is crucial. Model size indicates capacity, and precision affects information granularity. Larger models typically perform better but require more computational resources. Techniques like quantization and precision adjustments help deploy these models more efficiently. We studied the impact of precision settings on the robustness of T5-770m and OPT-1.3b models by comparing their performance under various precisions.

For the T5-770m and OPT-1.3b models, it's clear that as precision changes from fp16 to int4, there isn't a significant drop in their inherent accuracy. This indicates that models can handle reduced precision without compromising their general performance drastically. What's more, across different precision settings, the attack success rate for the T5-770m models remains fairly higher, which shows the same conclusion as in 6.1. However, the precision settings do not show a consistent pattern of influence on the ASR and replacement rate.

In essence, while different models exhibit different robustness against adversarial attacks, the precision settings do not play a significant role in this robustness.

	Acc	Acc/attack ↑	ASR ↓	Replacement
T5-770m-fp16	0.9106	0.1631	0.8208	0.1196
T5-770m-int8	0.9048	0.1536	0.8312	0.1143
T5-770m-int4	0.9210	0.1725	0.8127	0.1211
OPT-1.3b-fp16	0.9218	0.7496	0.1868	0.0536
OPT-1.3b-int8	0.9231	0.7515	0.1859	0.0421
OPT-1.3b-int4	0.9207	0.7531	0.1820	0.0498

Table 9: Results of T5-770m and OPT-1.3b models under different precision settings, including fp16, int8 and int4. The performance is evaluated with IMDB dataset.

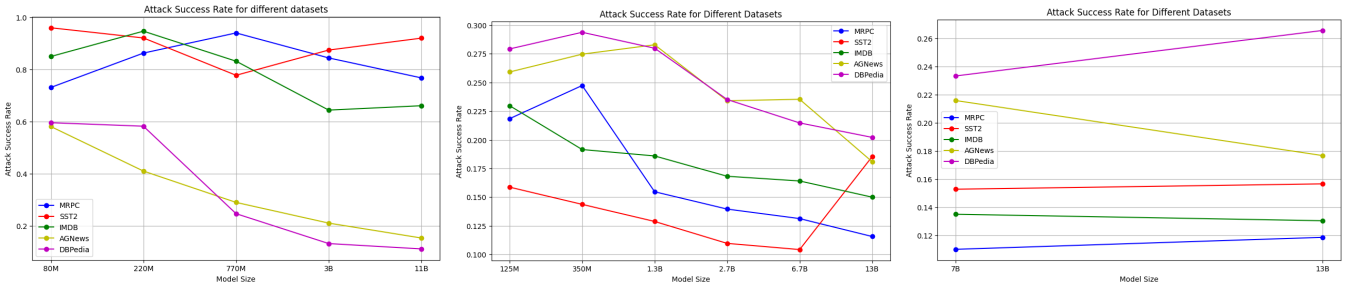
	Acc	Acc/attack ↑	ASR ↓	Replacement
T5-770m	0.9067	0.1499	0.8347	0.1036
T5-770m-Lora	0.9048	0.1536	0.8312	0.1143
OPT-1.3b	0.9135	0.7448	0.1847	0.0366
OPT-1.3b-LoRA	0.9231	0.7515	0.1859	0.0421
OPT-2.7b	0.9266	0.7741	0.1646	0.0452
OPT-2.7b-LoRA	0.9198	0.7651	0.1682	0.0396

Table 10: Results of the T5-770m, OPT-1.3b, and OPT-2.7b models' performance with and without the application of LoRA, using the IMDB dataset.

6.2.3 LoRA. As mention in Sec 3.2, LoRA has been a groundbreaking approach, bringing about significant reductions in memory requirements during model training. In this case, the potential trade-off in question is model robustness.

For our investigation, we selected the T5-770m, OPT-1.3b, and OPT-2.7b models. Experiments were conducted under two conditions for each model: with and without the application of LoRA. The IMDB dataset served as our benchmark for this analysis.

The experiments show that adversarial attacks significantly reduce accuracy across all models, regardless of LoRA's use. However, crucially, both the attack success rate and replacement rate, key measures of resilience against adversarial tactics, were unaffected by LoRA. This indicates that while LoRA enhances optimization, it doesn't negatively impact the model's defense against adversarial attacks, providing optimization benefits without sacrificing robustness.



(a) T5 model **(b) OPT model** **(c) Llama model**

Figure 2: The experimental results of different models on various datasets.

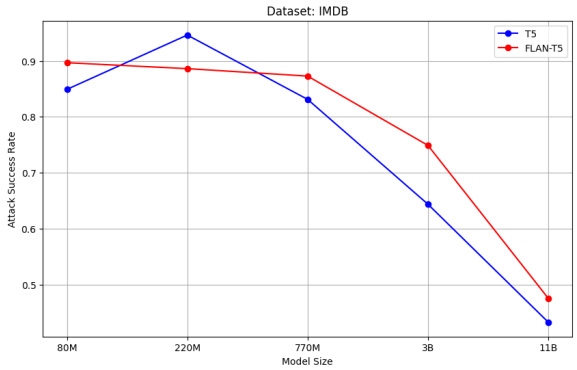


Figure 3: The experimental results of T5 and Flan-T5 on IMDB dataset

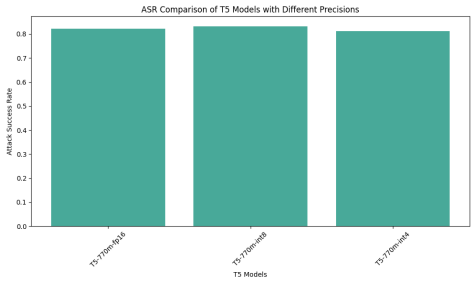


Figure 4: Different precisions on the IMDB dataset (T5 model)

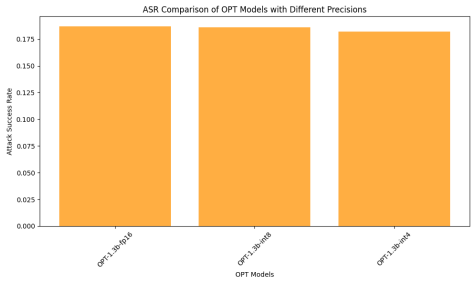


Figure 5: Different precisions on the IMDB dataset (OPT Model)

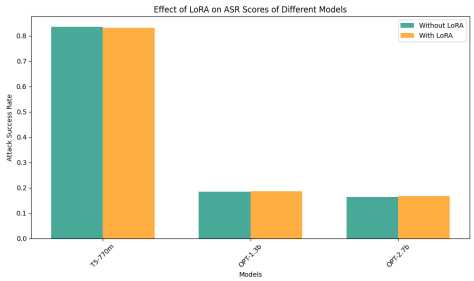


Figure 6: Results of LoRA on IMDB dataset

Figure 7: Comparison of different models and precisions on the IMDB dataset

6.3 Model Architectures (RQ3)

The architecture of a model's output space significantly influences its performance and resilience against adversarial attacks. For models with a classification head, the output is simplified to a binary decision, contrasting with OPT models without such a head, which must identify 'negative' or 'positive' labels from a vast vocabulary. This distinction impacts the model's accuracy.

Our data shows that smaller models with a classification head are more accurate than headless ones due to their simplified output space, which aids decision-making, especially in models with limited processing power. Moreover, models with a head reach their peak performance faster, achieving accuracy saturation more quickly.

However, an intriguing observation is the heightened attack success rate for models with a classification head. On the surface, this suggests that launching adversarial attacks against these models is a more straightforward task.One main factor contributes to this vulnerability is DeepFool's efficacy with last layer FFN: In such models, DeepFool can more readily discern the optimal direction for launching its attack, amplifying the ASR. This marked efficiency underscores a reduced robustness in these models against adversarial intrusions.

	Acc	Acc/attack ↑	ASR ↓	Replacement
OPT-125m	0.8616	0.6637	0.2297	0.0365
OPT-125m-head	0.9074	0.6215	0.3151	0.0476
OPT-350m	0.8564	0.6924	0.1915	0.0305
OPT-350m-head	0.9152	0.6643	0.2741	0.0682
OPT-1.3b	0.9231	0.7515	0.1859	0.0421
OPT-1.3b-head	0.9316	0.7621	0.1819	0.0533
OPT–2.7b	0.9198	0.7651	0.1682	0.0396
OPT–2.7b-head	0.9367	0.7704	0.1775	0.0516
OPT-6.7b	0.9408	0.7864	0.1641	0.0528
OPT-6.7b-head	0.9422	0.7765	0.1759	0.0627
OPT-13b	0.9431	0.8016	0.1500	0.0671
OPT-13b-head	0.9427	0.7877	0.1644	0.0641

Table 11: Experimental results of OPT models with/without classification head with IMDB dataset.

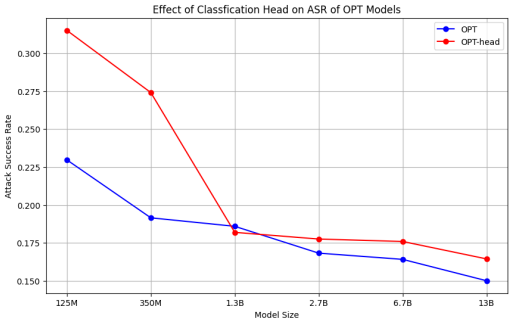


Figure 8: Results of classfication head on IMDB dataset

6.4 Attack Examples

In this section, we are going to show some adversarial examples for different tasks and models in Table 12

7 Conclusion

This paper utilized a novel geometric adversarial attack method to assess the robustness of leading LLMs, utilizing advanced fine-tuning techniques for task-specific model adaptation. Our groundbreaking approach revealed that these models exhibit variable sensitivity to adversarial attacks, influenced by their size and architectural differences. This indicates inherent vulnerabilities in LLMs, yet suggests potential resilience in certain configurations. Contrary to expectations, LLM-specific techniques did not markedly reduce robustness. Future research could explore models like RLHF and model parallelism approaches within this framework. Additionally, the evolution of more complex adversarial attacks promises deeper insights into LLM strengths and weaknesses.

Ethics statement

In our research, we employ adversarial attack methodologies to generate text, aiming to evaluate the robustness of LLMs against inputs. However, we acknowledge the ethical implications associated with the use of adversarial attacks. One primary concern is the potential generation of harmful information. This includes text that may be offensive, misleading, or harmful in other ways. Therefore, one should be cautious when taking such methods into practical use.

References

[1] Moustafa Alzantot, Yash Sharma, Ahmed Elgohary, Bo-Jhang Ho, Mani Srivastava, and Kai-Wei Chang. 2018. Generating Natural Language Adversarial Examples. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Brussels, Belgium, 2890–2896. https://doi.org/10.18653/v1/D18-1316

[2] Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. 2007. Dbpedia: A nucleus for a web of open data. In *international semantic web conference*. Springer, 722–735.

[3] Yushi Bai, Jiahao Ying, Yixin Cao, Xin Lv, Yuze He, Xiaozhi Wang, Jifan Yu, Kaisheng Zeng, Yijia Xiao, Haozhe Lyu, Jiayin Zhang, Juanzi Li, and Lei Hou. 2023. Benchmarking Foundation Models with Language-Model-as-an-Examiner. arXiv:2306.04181 [cs.CL]

[4] Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, Quyet V. Do, Yan Xu, and Pascale Fung. 2023. A Multitask, Multilingual, Multimodal Evaluation of ChatGPT on Reasoning, Hallucination, and Interactivity. arXiv:2302.04023 [cs.CL]

[5] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258* (2021).

[6] Tim Dettmers, Mike Lewis, Younes Belkada, and Luke Zettlemoyer. 2022. LLM.int8(): 8-bit Matrix Multiplication for Transformers at Scale. arXiv:2208.07339 [cs.LG]

[7] Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. QLoRA: Efficient Finetuning of Quantized LLMs. arXiv:2305.14314 [cs.LG]

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

[9] Javid Ebrahimi, Anyi Rao, Daniel Lowd, and Dejing Dou. 2018. HotFlip: White-Box Adversarial Examples for Text Classification. arXiv:1712.06751 [cs.CL]

[10] Chuan Guo, Alexandre Sablayrolles, Hervé Jégou, and Douwe Kiela. 2021. Gradient-based Adversarial Attacks against Text Transformers. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 5747–5757.

[11] Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2021. LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations*.

[12] Robin Jia and Percy Liang. 2017. Adversarial Examples for Evaluating Reading Comprehension Systems. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Copenhagen, Denmark, 2021–2031. https://doi.org/10.18653/v1/D17-1215

[13] Akbar Karimi, Leonardo Rossi, and Andrea Prati. 2021. AEDA: An Easier Data Augmentation Technique for Text Classification. In *Findings of the Association for Computational Linguistics: EMNLP 2021*. Association for Computational Linguistics, Punta Cana, Dominican Republic, 2748–2754. https://doi.org/10.18653/v1/2021.findings-emnlp.234

[14] Jinfeng Li, Shouling Ji, Tianyu Du, Bo Li, and Ting Wang. 2019. TextBugger: Generating Adversarial Text Against Real-world Applications. In *Proceedings 2019 Network and Distributed System Security Symposium*. Internet Society. https://doi.org/10.14722/ndss.2019.23138

[15] Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Cosgrove, Christopher D. Manning, Christopher Ré, Diana Acosta-Navas, Drew A. Hudson, Eric Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue Wang, Keshav Santhanam, Laurel Orr, Lucia Zheng, Mert Yuksekgonul, Mirac Suzgun, Nathan Kim, Neel Guha, Niladri Chatterji, Omar Khattab, Peter Henderson, Qian Huang, Ryan Chi, Sang Michael Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas Icard, Tianyi Zhang, Vishrav Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Koreeda. 2022. Holistic Evaluation of Language Models. arXiv:2211.09110 [cs.CL]

[16] Edward Ma. 2019. NLP Augmentation. https://github.com/makcedward/nlpaug.

[17] Andrew Maas, Raymond E Daly, Peter T Pham, Dan Huang, Andrew Y Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*. 142–150.

[18] Zhao Meng, Yihan Dong, Mrinmaya Sachan, and Roger Wattenhofer. 2022. Self-Supervised Contrastive Learning with Adversarial Perturbations for Defending Word Substitution-based Attacks. In *Findings of the Association for Computational Linguistics: NAACL 2022*. 87–101.

[19] Zhao Meng and Roger Wattenhofer. 2020. A geometry-inspired attack for generating natural language adversarial examples. *arXiv preprint arXiv:2010.01345* (2020).

Table 12: Adversarial Examples for Various Tasks and Models

Task Description and Adversarial Example
Dataset: IMDB *Model:* T5-780M **Original:** "Choose the sentiment of this review? Expensive lunch meals. Fried pickles were good. Waitress messed up 2 orders out of 4. "Dont" think "Ill" return. Asked for no cheese waitress joked extra cheese then brought my meal with cheese. Better places to eat in area. " **Adversarial:** "Choose the sentiment of this review? Expensive brunch meals. Fried marinated were good. Waitress eventhough up 2 orders out of 4. "Dont" imagining "Ill" return. Asked for no cheese miss jokingly extra cheese then brought my meal with cheese. Better place to devoured in area."
Dataset: MRPC *Model:* OPT-2.7B **Original:** "Do these two sentences mean the same thing? Crews worked to install a new culvert and prepare the highway so motorists could use the eastbound lanes for travel as storm clouds threatened to dump more rain. Crews worked to install a new culvert and repave the highway so motorists could use the eastbound lanes for travel." **Adversarial:** "Do these two sentences mean the same thing? Crews acted to mount a nouvelle drains and prepare the avenue so chauffeurs could use the eastbound routing for voyage as stormy haze threatens to spill more rainfall. Crews worked to install a newer septic and repave the highway so motorists could use the eastbound lanes for tours"
Dataset: AGNews *Model:* Llama-7B **Original:** "Dial M for Music Mobile-phone makers scored a surprising hit four years ago when they introduced handsets equipped with tiny digital cameras. Today, nearly one-third of the cell phones sold worldwide do double duty as cameras." **Adversarial:** "Dial M for Melody Mobile-phone manufacturers scored a surprising hit four years ago when they unveiled handsets equipped with tiny digital cameras. Today, nearly one-third of the cell phones sold worldwide do double duty as cameras."

[20] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. 2016. DeepFool: a simple and accurate method to fool deep neural networks. arXiv:1511.04599 [cs.LG]

[21] Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2020. Adversarial NLI: A New Benchmark for Natural Language Understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 4885–4901. https://doi.org/10.18653/v1/2020.acl-main.441

[22] OpenAI. 2023. GPT-4 Technical Report. (2023). arXiv:2303.08774 [cs.CL]

[23] OpenAI. 2023. OpenAI Chatbot. https://chat.openai.com

[24] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep Contextualized Word Representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. Association for Computational Linguistics, New Orleans, Louisiana, 2227–2237. https://doi.org/10.18653/v1/N18-1202

[25] Chengwei Qin, Aston Zhang, Zhuosheng Zhang, Jiaao Chen, Michihiro Yasunaga, and Diyi Yang. 2023. Is ChatGPT a General-Purpose Natural Language Processing Task Solver? arXiv:2302.06476 [cs.CL]

[26] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. arXiv:1910.10683 [cs.LG]

[27] 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*. 1631–1642.

[28] Arsene Fansi Tchango, Rishab Goel, Zhi Wen, Julien Martel, and Joumana Ghosn. 2022. DDXPlus: A New Dataset For Automatic Medical Diagnosis. arXiv:2205.09148 [cs.CL]

[29] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models. arXiv:2307.09288 [cs.CL]

[30] Eric Wallace, Shi Feng, Nikhil Kandpal, Matt Gardner, and Sameer Singh. 2019. Universal Adversarial Triggers for Attacking and Analyzing NLP. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, 2153–2162.

[31] Boxin Wang, Weixin Chen, Hengzhi Pei, Chulin Xie, Mintong Kang, Chenhui Zhang, Chejian Xu, Zidi Xiong, Ritik Dutta, Rylan Schaeffer, Sang T. Truong, Simran Arora, Mantas Mazeika, Dan Hendrycks, Zinan Lin, Yu Cheng, Sanmi Koyejo, Dawn Song, and Bo Li. 2023. DecodingTrust: A Comprehensive Assessment of Trustworthiness in GPT Models. arXiv:2306.11698 [cs.CL]

[32] Boxin Wang, Chejian Xu, Shuohang Wang, Zhe Gan, Yu Cheng, Jianfeng Gao, Ahmed Hassan Awadallah, and Bo Li. 2022. Adversarial GLUE: A Multi-Task Benchmark for Robustness Evaluation of Language Models. arXiv:2111.02840 [cs.CL]

[33] Jindong Wang, Xixu Hu, Wenxin Hou, Hao Chen, Runkai Zheng, Yidong Wang, Linyi Yang, Haojun Huang, Wei Ye, Xiubo Geng, Binxin Jiao, Yue Zhang, and Xing Xie. 2023. On the Robustness of ChatGPT: An Adversarial and Out-of-distribution Perspective. arXiv:2302.12095 [cs.AI]

[34] Longyue Wang, Chenyang Lyu, Tianbo Ji, Zhirui Zhang, Dian Yu, Shuming Shi, and Zhaopeng Tu. 2023. Document-Level Machine Translation with Large Language Models. arXiv:2304.02210 [cs.CL]

[35] Zengzhi Wang, Qiming Xie, Zixiang Ding, Yi Feng, and Rui Xia. 2023. Is ChatGPT a Good Sentiment Analyzer? A Preliminary Study. arXiv:2304.04339 [cs.CL]

[36] Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022. Emergent Abilities of Large Language Models. arXiv:2206.07682 [cs.CL]

[37] Jason Wei and Kai Zou. 2019. EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, Hong Kong, China, 6382–6388. https://doi.org/10.18653/v1/D19-1670

[38] Huangzhao Zhang, Hao Zhou, Ning Miao, and Lei Li. 2019. Generating Fluent Adversarial Examples for Natural Languages. In *Proceedings of the 57th Annual*

Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, Florence, Italy, 5564–5569. https://doi.org/10.18653/v1/P19-1559

[39] Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. 2022. OPT: Open Pre-trained Transformer Language Models. arXiv:2205.01068 [cs.CL]

[40] Wenxuan Zhang, Yue Deng, Bing Liu, Sinno Jialin Pan, and Lidong Bing. 2023. Sentiment Analysis in the Era of Large Language Models: A Reality Check. arXiv:2305.15005 [cs.CL]

[41] Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. *Advances in neural information processing systems* 28 (2015).

[42] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. 2023. A Survey of Large Language Models. arXiv:2303.18223 [cs.CL]

[43] Yunqing Zhao, Tianyu Pang, Chao Du, Xiao Yang, Chongxuan Li, Ngai-Man Cheung, and Min Lin. 2023. On Evaluating Adversarial Robustness of Large Vision-Language Models. arXiv:2305.16934 [cs.CV]

[44] Kaijie Zhu, Jindong Wang, Jiaheng Zhou, Zichen Wang, Hao Chen, Yidong Wang, Linyi Yang, Wei Ye, Neil Zhenqiang Gong, Yue Zhang, and Xing Xie. 2023. PromptBench: Towards Evaluating the Robustness of Large Language Models on Adversarial Prompts. arXiv:2306.04528 [cs.CL]

A Research Methods

A.1 Experiments Setup

In our study, we employ pretrained weights from HuggingFace and use int8 quantization for GPU memory optimization. We also standardize the use of LoRA to reduce training parameters. For models under 3 billion parameters, experiments are conducted on an NVIDIA RTX 3090 (24GB), whereas models above 3 billion parameters are tested on NVIDIA RTX A6000 (48GB) or NVIDIA A100 (40GB), catering to both fine-tuning and attack simulations.

A.2 Dataset

In this section, we will provide more details about the datasets used in this work.

IMDB: This dataset contains 50,000 movie reviews for sentiment analysis, equally divided between positive and negative sentiments.

SST-2: An extension of the original SST, it focuses on the binary classification of sentiments in movie review sentences.

MRPC: A corpus for paraphrase identification, it includes sentence pairs from online news sources, annotated for semantic equivalence.

AGNews: This news categorization dataset comprises 120,000 training and 7,600 test samples across four categories: World, Sports, Business, and Science/Technology.

DBpedia: A large-scale, multi-class dataset from the DBpedia knowledge base, featuring 560,000 training and 70,000 test samples across 14 categories.

The statistics of these five datasets are presented in Table 13.

Table 13: Statistics of the Datasets

Dataset	Labels	Avg Length	Train	Test
IMDB	2	279.48	25000	25000
MRPC	2	52.89	3670	1730
SST-2	2	19.81	67300	1820
AGNews	4	43.93	120000	7600
DBPedia	14	55.14	560000	7000