# Quantized Versions of Llama Robustness Against Prompt-Based Adversarial Attacks

**Mahdi Saberi, Alireza Sharbafchi**
{saber032,sharb019}@umn.edu
Group Name: Artificial Language Processing (ALP)

## Abstract

Modern natural language processing (NLP) models such as BERT, GPT-3, and LLaMA have achieved impressive language understanding capabilities but remain vulnerable to adversarial attacks. Even minor input modifications can lead to misclassifications or unintended outputs, posing a significant challenge for their reliable deployment in critical applications. Although recent efforts to optimize and quantize these models enable efficient deployment on resource-constrained devices like smartphones — exemplified by Meta's quantized LLaMA variants, which offer substantial reductions in model size and memory usage — such optimizations often come at the cost of decreased robustness. In this work, we systematically investigate the vulnerabilities of quantized NLP models to adversarial perturbations and develop techniques to enhance their resilience. By identifying the factors that contribute to robustness degradation in smaller, more efficient models, we propose strategies to strengthen their defenses without compromising their efficiency benefits. Our findings aim to improve the trustworthiness of NLP models across a wide spectrum of hardware platforms, ultimately fostering safer and more reliable language technologies for real-world deployments.

## 1 Introduction

Modern NLP models excel at various tasks but are susceptible to adversarial attacks, where small input modifications can manipulate predictions. This issue is particularly critical for models optimized for mobile or low-resource environments, such as Meta's quantized LLaMA models. These vulnerabilities challenge the deployment of NLP systems in security-critical applications like content moderation, AI decision-making, and more.

Adversarial attacks in NLP have seen significant advancements through innovative methods targeting text classifiers and transformer models. One notable approach is HotFlip, introduced by (Ebrahimi et al., 2017), a white-box adversarial attack method for character-level neural text classifiers. HotFlip uses gradient-based optimization to make small changes, or "flips," to individual characters in the input text. These perturbations are designed to maximize the model's loss while preserving semantic meaning and fluency. The efficiency of HotFlip lies in its ability to generate adversarial examples with minimal modifications, effectively fooling classifiers at both character and word levels.

Building on these advancements, Gradient-Based Distributional Attack (GBDA) was proposed by Guo et al. (Guo et al., 2021) as a novel framework specifically for transformer models in NLP. GBDA overcomes the challenges of discrete text data by formulating a continuous matrix representation of adversarial examples. Using Gumbel-softmax, it ensures semantic similarity and fluency in the generated examples. This approach enables efficient gradient-based attacks while achieving high attack success rates in both white-box and black-box settings, setting a new standard for adversarial attack frameworks.

In another significant contribution, (Maheshwary et al., 2021) introduced a decision-based attack strategy tailored for hard-label black-box scenarios. These attacks focus on text classification and entailment tasks where only the top predicted label is available, without access to gradients or probability distributions. Their method employs a population-based optimization algorithm to generate adversarial examples that maintain high semantic similarity and grammatical correctness with minimal word perturbations. This approach not only achieves high attack success rates but also ensures lower perturbation rates compared to prior methods, demonstrating its effectiveness in crafting robust adversarial examples. These works highlight the need for defenses that preserve semantic integrity while improving robustness.

The aim of this work is to identify vulnerabilities in LLMs, particularly optimized versions, and to develop robust defense mechanisms. By doing so, we address the trade-off between efficiency and robustness, ensuring reliability across devices ranging from high-performance systems to mobile phones.

## 2 Motivation

Despite their efficiency, optimized LLMs are more susceptible to adversarial perturbations. For instance, Meta's quantized LLaMA models achieve a 56% size reduction but exhibit increased vulnerability to attacks due to their compact architectures. Addressing these vulnerabilities is essential for real-world applications where robustness and reliability are paramount.

## 3 Problem Statement

Deploying LLMs on resource-constrained devices faces challenges due to the susceptibility of optimized models to adversarial attacks. This project investigates:

- Susceptibility of different versions of LLaMA 3.2 to adversarial perturbations.
- Fine-Tuning and robustness evaluation across sentiment analysis Stanford Sentiment Treebank (SST-2 dataset), as well as paraphrase detection on Quora Question Pairs (QQP dataset).
- Incorporating methods used in AdvGLUE dataset for enhancing robustness through adversarial training.

In terms of Novelty, this project studies the vulnerabilities of the quantized versions of the LLaMA 3.2 versions that have been published recently for resource-constraint devices like cellphones. Our results are strong enough to show that the quantized versions are more venerable to small perturbations to various prompts. This study can be helpful to the people how are trying to publish their plug-ins for cellphone devices and may want to consider how prompt-perturbation can effect their efficiency.

## 4 Approach

Based on the attacks were used in the AdvGLUE (Wang et al., 2021) dataset, we've chosen to apply one of the adversary perturbations which is computationally much expensive and used it during the training. This makes the model familiar with various adversary smaples and robustious it agaist perturbations.

### 4.1 Tasks and Datasets

We evaluate quantized Llama 1B and 3B parameters models' robustness using the Stanford Sentiment Treebank (SST-2) (Wolf et al., 2020) and Quora Question Pairs (QQP) (Peinelt et al., 2019) dataset. We performed a binary classification on the SST-2 (Wolf et al., 2020) dataset, as well as QQP (Peinelt et al., 2019) dataset where classifies two questions as 1 if they are parapherase of each other, otherwise they classify as 0.

### 4.2 Adversarial Attacks

In here, we evaluated the quantized models with the AdvGLUE (Wang et al., 2021) dataset, which uses different attack strategies to perturb the prompts. We will explain some of these methods in the following:

**Typo-Based Perturbation:**We select TextBugger (Li et al., 2018) as the representative algorithm for generating typo-based adversarial examples. When performing the attack, TextBugger first identifies the important words and then replaces them with typos.

**Embedding-similarity-based Perturbation:** We choose TextFooler (Jin et al., 2020) as the representative adver sarial attack that considers embedding similarity as a constraint to generate semantically consistent adversarial examples. Essentially, TextFooler first performs word importance ranking, and then substitutes those important ones to their synonyms extracted according to the cosine similarity of word embeddings.

**Distraction-based Perturbation:**We integrate two attack strategies: (i) StressTest (Naik et al., 2018) appends three true statements ("and true is true", "and false is not true", "and true is true" for five times) to the end of the hypothesis sentence for NLI tasks. (ii) CheckList (Ribeiro et al., 2020) adds randomly generated URLs and handles to distract model attention. Since the aforementioned distraction-based perturbations may impact the linguistic acceptability and the understanding of semantic equivalence, we mainly apply these rules to part of the GLUE (Wang, 2018) tasks, including SST-2 (Wolf et al., 2020) and NLI tasks (MNLI, RTE, QNLI), to evaluate whether model can be easily misled by the strong negation words or such lexical similarity. Table 1 discussed about the all methods used to generate AdvGLUE (Wang et al., 2021) dataset.

### 4.3 Evaluation Metric

We use **Attack Success Rate (ASR)** and **Classification Accuracy** (for classification tasks) to quantify model vulnerability:

$$\text{ASR} = \frac{\text{Successful Attacks}}{\text{Total Attempts}} \times 100\%.$$

### 4.4 Proposed Method

Adversarial fine-tuning integrates adversarial algorithms during training, enabling models to adapt to perturbations. The AdvGLUE (Wang et al., 2021) dataset provides diverse adversarial discussed beriefley in Section 4.2.

## 5 Experiments and Results

All the experiments were performed on a server with an NVIDIA A100 GPU.

### 5.1 Experimental Setup

LLaMA 3.2 models (1B and 3B parameters) were fine-tuned on SST-2 (Wolf et al., 2020). TextBugger (Li et al., 2018) was chosen for adversarial training (AT) due to its low computational demand, in a way that in every iteration of training, the perturbed prompt goes to network and compared with the label at the end. Table 2 shows how the models' accuracy improves using this technique. The first two big models (13B and 7B LLaMA) were obtained from the (Yang et al., 2024). Since the adversary taininig demands a powerfull computation, and this couldn't happen on colab notebook, we haven't trained these big models with the perturbations.

| Model | Test (No Attack) | Test (With Attack) | AT |
|---|---|---|---|
| LLaMA 13B | 0.963 | 0.812 | - |
| LLaMA 7B | **0.968** | **0.820** | - |
| LLaMA 3B | 0.912 | 0.446 | **0.513** |
| LLaMA 1B | 0.887 | 0.350 | 0.492 |

Table 2: Accuracy percentage under adversarial conditions.

Optimized models are disproportionately vulnerable to white-box attacks due to their compact architectures. For instance, LLaMA 1B exhibited significant drops in accuracy under adversarial perturbation, underscoring the need for resource-efficient robustness mechanisms.

## 6 Ethics and Discussion

Adversarial attack research can be misused. However, it is essential for developing defenses against malicious actors. Ensuring responsible use through transparency and ethical guidelines is critical for advancing the field without societal harm.

## 7 Limitations & Challenges

The size of the models makes fine-tuning time-consuming, even for optimized models like LLaMA 3.2 with 1B parameters. AdvGLUE (Wang et al., 2021) was generated using several algorithms to attack well-known models, with human annotation applied to certain parts. This combination makes it challenging to improve the robustness of models against this test bench using only one algorithm (due to resource limitations) and without comprehensive human annotation across the entire dataset.

## 8 Conclusion and Future Work

Adversarial training improves robustness against token manipulation attacks. In terms of Novelty, this project studies the vulnerabilities of the quantized versions of the LLaMA 3.2 versions that have been published recently for resource-constraint devices like cellphones. Our results are strong enough to show that the quantized versions are more venerable to small perturbations to various prompts. This study can be helpful to the people how are trying to publish their plug-ins for cellphone devices and may want to consider how prompt-perturbation can effect their efficiency.

## References

Javid Ebrahimi, Anyi Rao, Daniel Lowd, and Dejing Dou. 2017. Hotflip: White-box adversarial examples for NLP. *CoRR*, abs/1712.06751.

Chuan Guo, Alexandre Sablayrolles, Hervé Jégou, and Douwe Kiela. 2021. Gradient-based adversarial attacks against text transformers. *CoRR*, abs/2104.13733.

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.

Jinfeng Li, Shouling Ji, Tianyu Du, Bo Li, and Ting Wang. 2018. Textbugger: Generating adversarial

| Method | Type | Computation(1-5) |
|---|---|---|
| TextBugger | Word-level | 2 |
| TextFooler | Word-level | 3 |
| BERT-ATTACK | Word-level | 4 |
| SememePSO | Word-level | 5 |
| CompAttack | Word-level | 5 |
| SCPN | Sentence-level | 4 |
| T3 | Sentence-level | 5 |
| AdvFever | Sentence-level | 4 |
| CheckList | Sentence-level | 3 |
| StressTest | Sentence-level | 3 |

Table 1: Computational demand ratings for adversarial attack methods in NLP.

text against real-world applications. *arXiv preprint arXiv:1812.05271*.

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.

Aakanksha Naik, Abhilasha Ravichander, Norman Sadeh, Carolyn Rose, and Graham Neubig. 2018. Stress test evaluation for natural language inference. *arXiv preprint arXiv:1806.00692*.

Nicole Peinelt, Maria Liakata, and Dong Nguyen. 2019. Aiming beyond the obvious: Identifying non-obvious cases in semantic similarity datasets. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 2792–2798.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of nlp models with checklist. *arXiv preprint arXiv:2005.04118*.

Alex Wang. 2018. Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*.

Boxin Wang, Chejian Xu, Shuohang Wang, Zhe Gan, Yu Cheng, Jianfeng Gao, Ahmed Hassan Awadallah, and Bo Li. 2021. Adversarial GLUE: A multi-task benchmark for robustness evaluation of language models. *CoRR*, abs/2111.02840.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 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, Online. Association for Computational Linguistics.

Zeyu Yang, Zhao Meng, Xiaochen Zheng, and Roger Wattenhofer. 2024. Assessing adversarial robustness of large language models: An empirical study. *arXiv preprint arXiv:2405.02764*.