EasyChair Preprint

№ 12274

Robustness of Large Language Models: Mitigating Adversarial Attacks and Input Perturbations

Kurez Oroy and Jane Anderson

February 24, 2024

Robustness of Large Language Models: Mitigating Adversarial Attacks and Input Perturbations

Kurez Oroy, Jane Anderson

Abstract:

This paper explores the robustness of LLMs and strategies for mitigating the impact of adversarial attacks and input perturbations. Adversarial attacks, where small, carefully crafted perturbations are added to input data to induce misclassification or undesired behavior, can exploit vulnerabilities in LLMs and compromise their performance. Additionally, input perturbations, such as typographical errors or grammatical inconsistencies, can also degrade the accuracy and reliability of LLMs in practical settings. To address these challenges, various approaches have been proposed, including adversarial training, robust optimization techniques, and input preprocessing methods.

Keywords: Large Language Models (LLMs), Robustness, Adversarial Attacks, Input Perturbations, Adversarial Training, Robust Optimization, Input Preprocessing, Vulnerabilities

Introduction:

In recent years, large language models (LLMs) such as GPT (Generative Pre-trained Transformer) and BERT (Bidirectional Encoder Representations from Transformers) have achieved unprecedented success across a wide range of natural language processing (NLP) tasks[1]. These models, trained on massive amounts of text data, exhibit remarkable capabilities in understanding and generating human-like text. However, despite their impressive performance, LLMs are vulnerable to adversarial attacks and input perturbations, which can compromise their reliability and trustworthiness in real-world applications. Adversarial attacks on LLMs involve the deliberate manipulation of input data to induce misclassification or undesirable behavior in the model's predictions. These attacks can take various forms, such as adding imperceptible perturbations to input text or injecting subtle alterations that lead to incorrect outputs. Additionally, input

perturbations, including typographical errors, grammatical inconsistencies, or linguistic variations, can also disrupt the performance of LLMs, especially when deployed in dynamic and noisy environments[2]. The susceptibility of LLMs to adversarial attacks and input perturbations raises significant concerns regarding their robustness and safety in practical applications. Adversarial examples crafted to deceive LLMs can have severe consequences, ranging from misinformation propagation to system vulnerabilities in security-critical domains. Furthermore, input perturbations, whether intentional or unintentional, can lead to unexpected errors and biases in model predictions, undermining user trust and confidence in NLP systems. To address these challenges, researchers have proposed various defense mechanisms aimed at enhancing the robustness of LLMs against adversarial attacks and input perturbations. These approaches encompass adversarial training, where models are trained with adversarial examples to improve their resilience, robust optimization techniques that optimize model parameters to mitigate the impact of adversarial perturbations, and input preprocessing methods designed to sanitize input data and reduce its susceptibility to perturbations[3]. In recent years, large language models (LLMs) have revolutionized natural language processing (NLP) tasks, achieving state-of-the-art performance across a wide range of applications. However, the increasing deployment of LLMs in real-world scenarios has exposed their vulnerability to adversarial attacks and input perturbations, raising concerns about their reliability and robustness. In this introduction, we delve into the challenges posed by adversarial attacks and input perturbations on LLMs and explore strategies for mitigating these risks to enhance the robustness of language models in practical applications. Adversarial attacks are crafted perturbations applied to input data with the intention of deceiving LLMs into making incorrect predictions or producing undesired outputs[4]. These attacks exploit vulnerabilities in the underlying architecture and training process of LLMs, highlighting the need for robust defenses to safeguard against potential threats. Additionally, input perturbations, such as typographical errors or grammatical inconsistencies, can inadvertently impact the performance of LLMs, leading to unreliable and unpredictable behavior in real-world scenarios. To address these challenges, researchers have developed a variety of techniques aimed at enhancing the robustness of LLMs against adversarial attacks and input perturbations. Adversarial training involves augmenting the training data with adversarial examples to improve the model's resilience to adversarial attacks[5]. Robust optimization techniques focus on modifying the optimization process to minimize the impact of adversarial perturbations on model

predictions. Input preprocessing methods aim to preprocess input data to make LLMs more robust to input perturbations, such as introducing noise or applying data augmentation techniques. In this paper, we provide a comprehensive overview of recent advancements in mitigating adversarial attacks and input perturbations on LLMs[6]. We discuss the underlying mechanisms of adversarial vulnerabilities in LLMs, explore state-of-the-art defense strategies, and evaluate their effectiveness through empirical analysis and benchmarking against various attack scenarios. Additionally, we highlight key challenges and open research questions in this area, such as the trade-off between robustness and performance, the transferability of adversarial examples across models, and the generalization capabilities of defense mechanisms to unseen attack strategies. By addressing the robustness of LLMs against adversarial attacks and input perturbations, we aim to enhance the reliability and trustworthiness of NLP systems in real-world applications, ensuring their effectiveness and safety in diverse deployment scenarios. Through continued research and innovation, we can advance the state of the art in defending against adversarial threats and fostering greater confidence in the deployment of LLMs for practical use cases[7].

Advancements in Robustness against Adversarial Attacks and Input Manipulations:

The proliferation of large language models (LLMs) has ushered in a new era of natural language processing (NLP), enabling remarkable advancements in various language-related tasks. However, alongside their unparalleled performance, LLMs have become increasingly susceptible to adversarial attacks and input manipulations, posing significant challenges to their reliability and trustworthiness in practical applications[8]. In this introduction, we delve into the evolving landscape of robustness against adversarial attacks and input manipulations in LLMs, exploring recent advancements and strategies aimed at mitigating these vulnerabilities. Adversarial attacks represent a formidable threat to LLMs, as they involve crafting imperceptible perturbations to input data with the goal of inducing misclassification or erroneous behavior. These attacks exploit vulnerabilities in the complex architectures and training procedures of LLMs, highlighting the need for robust defense mechanisms to safeguard against potential threats. Furthermore, input manipulations, such as subtle modifications or distortions to input text, can also undermine the

performance and reliability of LLMs, leading to unpredictable outcomes in real-world scenarios. To address these challenges, researchers have developed a range of techniques and methodologies focused on enhancing the robustness of LLMs against adversarial attacks and input manipulations. Adversarial training, for instance, involves augmenting the training data with adversarial examples to fortify the model's resilience to adversarial perturbations[9]. Robust optimization techniques aim to modify the optimization process to minimize the impact of adversarial attacks on model predictions, while input preprocessing methods seek to preprocess input data to make LLMs more robust to input manipulations, such as by introducing noise or applying data augmentation techniques. In this paper, we provide an extensive overview of recent advancements in bolstering the robustness of LLMs against adversarial attacks and input manipulations. We delve into the underlying mechanisms of adversarial vulnerabilities in LLMs, explore state-of-the-art defense strategies, and evaluate their efficacy through empirical analysis and benchmarking against various attack scenarios[10]. Additionally, we identify key challenges and open research questions in this domain, such as the trade-off between robustness and performance, the transferability of adversarial examples across models, and the generalization capabilities of defense mechanisms to unseen attack strategies. By addressing the robustness of LLMs against adversarial attacks and input manipulations, we aim to enhance the reliability and trustworthiness of NLP systems in practical applications, ensuring their effectiveness and safety across diverse deployment scenarios. Through ongoing research and innovation, we strive to advance the state-of-the-art in defending against adversarial threats and fostering greater confidence in the deployment of LLMs for real-world use cases[11]. In recent years, the widespread adoption of large language models (LLMs) has revolutionized natural language processing (NLP), enabling state-of-the-art performance across a myriad of tasks. However, with the increasing reliance on these models in critical applications, concerns about their vulnerability to adversarial attacks and input manipulations have surfaced. This introduction aims to explore the advancements made in enhancing the robustness of LLMs against such threats, shedding light on the strategies developed to mitigate the risks associated with adversarial attacks and input manipulations. Adversarial attacks pose a significant challenge to the reliability and security of LLMs. These attacks involve the deliberate introduction of small, carefully crafted perturbations to input data with the aim of misleading the model's predictions or causing undesired outcomes. Adversarial vulnerabilities in LLMs can be exploited by adversaries to manipulate model behavior, leading to erroneous results or compromising the

integrity of NLP systems[12]. Furthermore, input manipulations, such as subtle alterations or noise injected into input text, can also affect the performance and reliability of LLMs, making them susceptible to unexpected behavior in real-world scenarios. Addressing the robustness of LLMs against adversarial attacks and input manipulations requires innovative strategies and techniques. Adversarial training, a method that involves augmenting training data with adversarial examples, has emerged as a promising approach for enhancing model resilience to adversarial attacks. By exposing LLMs to adversarial perturbations during training, these models can learn to recognize and mitigate adversarial threats, thereby improving their robustness in adversarial settings[13]. Additionally, robust optimization techniques aim to modify the optimization process to minimize the impact of adversarial perturbations on model predictions, further bolstering model robustness against adversarial attacks. In parallel, input preprocessing methods have been developed to fortify LLMs against input manipulations. These techniques involve modifying input data to make LLMs more resilient to variations and distortions, such as adding noise or applying data augmentation strategies. By preprocessing input text to enhance its robustness to perturbations, LLMs can better withstand the effects of input manipulations and maintain reliable performance in diverse deployment scenarios[14].

Strategies for Enhancing Resistance to Adversarial Attacks and Input Perturbations:

In the realm of natural language processing (NLP), the robustness of language models against adversarial attacks and input perturbations is of paramount importance for ensuring the reliability and security of NLP systems in real-world applications[15]. Adversarial attacks, characterized by subtle modifications to input data with the intent of deceiving language models, pose a significant threat to the integrity of NLP systems. Similarly, input perturbations, such as typographical errors or grammatical inconsistencies, can inadvertently affect the performance and reliability of language models, leading to unexpected behavior in practical scenarios. To address these challenges, researchers have developed a variety of strategies aimed at enhancing the resistance of language models to adversarial attacks and input perturbations. These strategies encompass a range of approaches, including adversarial training, robust optimization techniques, and input

preprocessing methods, each designed to fortify language models against adversarial threats and input variations[16]. Adversarial training involves augmenting the training data with adversarial examples, crafted to expose the model to potential vulnerabilities and encourage robustness to adversarial perturbations. By incorporating adversarial examples into the training process, language models can learn to recognize and mitigate adversarial threats, thereby improving their resilience in adversarial settings. Robust optimization techniques focus on modifying the optimization process to minimize the impact of adversarial perturbations on model predictions, further enhancing model robustness against adversarial attacks. Additionally, input preprocessing methods aim to preprocess input data to make language models more resilient to input perturbations, such as adding noise or applying data augmentation techniques, thereby strengthening their resistance to variations and distortions in input text[17]. In the rapidly evolving landscape of natural language processing (NLP), the deployment of large language models (LLMs) has paved the way for transformative advancements in various domains. However, alongside their remarkable performance, concerns regarding the robustness and security of these models against adversarial attacks and input perturbations have surfaced. This introduction sets the stage for exploring strategies aimed at fortifying LLMs against adversarial threats and input perturbations, underscoring the importance of enhancing resistance to such challenges. Adversarial attacks represent a significant threat to the reliability and integrity of LLMs[18]. These attacks exploit vulnerabilities in the underlying architectures and training methodologies of LLMs by introducing imperceptible perturbations to input data, thereby inducing erroneous predictions or undesired behavior. Similarly, input perturbations, ranging from minor typographical errors to deliberate modifications, can significantly impact the performance and reliability of LLMs, posing additional challenges in real-world applications. Addressing these challenges requires a multifaceted approach encompassing both proactive defense strategies and robust optimization techniques. Adversarial training stands out as a prominent strategy for enhancing resistance to adversarial attacks. By augmenting the training data with adversarial examples, LLMs can learn to recognize and mitigate adversarial perturbations, thereby improving their robustness and resilience in adversarial settings[19]. Additionally, robust optimization techniques aim to modify the optimization process to minimize the impact of adversarial perturbations on model predictions, further bolstering model robustness against adversarial threats. Furthermore, input preprocessing methods play a crucial role in fortifying LLMs against input perturbations. These techniques

involve preprocessing input data to enhance its robustness to variations and distortions, such as incorporating noise or applying data augmentation strategies. By proactively addressing potential vulnerabilities in input data, LLMs can better withstand the effects of input manipulations and maintain reliable performance in diverse deployment scenarios[20].

Conclusion:

In conclusion, the robustness of large language models (LLMs) against adversarial attacks and input perturbations is critical for ensuring their reliability, trustworthiness, and effectiveness in real-world applications. Adversarial attacks pose a significant threat to the integrity of LLMs, exploiting vulnerabilities in their architectures and training methodologies to induce erroneous predictions or undesired behavior. Addressing these challenges requires proactive defense strategies, such as adversarial training and robust optimization techniques, which aim to fortify LLMs against adversarial perturbations and improve their resilience in adversarial settings. Additionally, input preprocessing methods play a crucial role in enhancing LLM robustness by proactively addressing potential vulnerabilities in input data, thereby ensuring reliable performance in diverse deployment scenarios.

References:

[1] K. Peng *et al.*, "Towards making the most of chatgpt for machine translation," *arXiv preprint arXiv:2303.13780,* 2023.

[2] M. Artetxe, G. Labaka, E. Agirre, and K. Cho, "Unsupervised neural machine translation," *arXiv preprint arXiv:1710.11041,* 2017.

[3] Q. Lu, B. Qiu, L. Ding, L. Xie, and D. Tao, "Error analysis prompting enables human-like translation evaluation in large language models: A case study on chatgpt," *arXiv preprint arXiv:2303.13809,* 2023.

[4] A. Lopez, "Statistical machine translation," *ACM Computing Surveys (CSUR),* vol. 40, no. 3, pp. 1-49, 2008.

[5] Q. Zhong, L. Ding, J. Liu, B. Du, and D. Tao, "Can chatgpt understand too? a comparative study on chatgpt and fine-tuned bert," *arXiv preprint arXiv:2302.10198,* 2023.

[6] H. Wang, H. Wu, Z. He, L. Huang, and K. W. Church, "Progress in machine translation," *Engineering,* vol. 18, pp. 143-153, 2022.

[7] C. Zan *et al.*, "Vega-mt: The jd explore academy translation system for wmt22," *arXiv preprint arXiv:2209.09444,* 2022.

[8] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473,* 2014.

[9] Q. Zhong *et al.*, "Toward efficient language model pretraining and downstream adaptation via self-evolution: A case study on superglue," *arXiv preprint arXiv:2212.01853,* 2022.

[10] M. D. Okpor, "Machine translation approaches: issues and challenges," *International Journal of Computer Science Issues (IJCSI),* vol. 11, no. 5, p. 159, 2014.

[11] Q. Lu, L. Ding, L. Xie, K. Zhang, D. F. Wong, and D. Tao, "Toward human-like evaluation for natural language generation with error analysis," *arXiv preprint arXiv:2212.10179,* 2022.

[12] M. I. Jordan and T. M. Mitchell, "Machine learning: Trends, perspectives, and prospects," *Science,* vol. 349, no. 6245, pp. 255-260, 2015.

[13] K. Peng *et al.*, "Token-level self-evolution training for sequence-to-sequence learning," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 2023, pp. 841-850.

[14] B. Mahesh, "Machine learning algorithms-a review," *International Journal of Science and Research (IJSR).[Internet],* vol. 9, no. 1, pp. 381-386, 2020.

[15] L. Ding and D. Tao, "The University of Sydney's machine translation system for WMT19," *arXiv preprint arXiv:1907.00494,* 2019.

[16] G. Bonaccorso, *Machine learning algorithms*. Packt Publishing Ltd, 2017.

[17] L. Zhou, L. Ding, K. Duh, S. Watanabe, R. Sasano, and K. Takeda, "Self-guided curriculum learning for neural machine translation," *arXiv preprint arXiv:2105.04475,* 2021.

[18] Y. Wu *et al.*, "Google's neural machine translation system: Bridging the gap between human and machine translation," *arXiv preprint arXiv:1609.08144,* 2016.

[19] Y. Lei, L. Ding, Y. Cao, C. Zan, A. Yates, and D. Tao, "Unsupervised Dense Retrieval with Relevance-Aware Contrastive Pre-Training," *arXiv preprint arXiv:2306.03166,* 2023.

[20] L. Ding, L. Wang, X. Liu, D. F. Wong, D. Tao, and Z. Tu, "Understanding and improving lexical choice in non-autoregressive translation," *arXiv preprint arXiv:2012.14583,* 2020.