Proceedings of the 2nd International Conference on Machine Learning and Automation
DOI: 10.54254/2755-2721/93/20240922

Research on adversarial attack and defense of large language models

Jidong Yang[1,5,*], Qiangyun Chi[2,6], Wenqiang Xu[3,7], Huaike Yu[4,8]

[1]School of Information And Physical Sciences, The University of Newcastle, Callaghan, NSW 2308, Australia,
[2]School of Computer Science, Nanjing University of Information Science & Technology, Nanjing, Jiangsu, 210044, China
[3]China United Network Communications Corporation Software Research Institute Nanjing, Jiangsu, 210019, China
[4]School of Art & Design, The University of New South Wales, Paddington , NSW 2021, Australia

[5]Jidong.YANG@uon.edu.au
[6]202183290046@nuist.edu.cn
[7]xuwq50@chinaunicom.cn
[8]huaike.yu@student.unsw.edu.au
*corresponding author

Abstract. Large language models (LLMs) have made excellent progress in text and image understanding and generation. However, with the wide range of applications of these models in various industries, the issue of their security, especially the defense against adversarial attacks, has become a focus of research. This study focuses on exploring the adversarial attacks faced by LLMs and their defense strategies, especially the design and optimization of defense mechanisms. Through literature review and case studies, this paper analyzes in detail the white-box and black-box attack patterns against LLMs, including model inversion, backdoor attacks, and token-based strategies. In response to these attacks, this paper proposes a series of defense strategies, including preventive measures such as data augmentation, adversarial training and model regularization, as well as real-time attack detection and response strategies such as anomaly detection and adversarial sample detection techniques. The core of this research is to improve the robustness and trustworthiness of LLMs, providing the necessary guarantees for their integration and sustainability in multiple industrial applications. In addition, this paper proposes future research directions, highlighting the importance of developing advanced defense systems, promoting interdisciplinary research and exploring new applications for LLMs. This research provides valuable insights into understanding and improving the security defense mechanisms of LLMs, which is essential for maintaining the security and user trust of these models.

Keywords: Large Language Model, Transformer Architecture, Adversarial Attack, Adversarial Defense, Security

1. Introduction

Large Language Models (LLMs) are a major advancement in the field of Artificial Intelligence, capable of understanding, generating and translating text, providing question answering services, executing code and other tasks. Training can be done with labeled data or unlabeled data. Supervised learning is suitable for labeled data and unsupervised learning is suitable for unlabeled data [1].

As LLMs are increasingly integrated into various industries, their security has become a critical issue. Ensuring the integrity and reliability of these models is critical as they are vulnerable to malicious attacks aimed at manipulating their output or disrupting their functionality. Attackers generating inputs that lead to model errors or LLMs trained on large datasets often unintentionally learn and perpetuate biases and prejudices in human data [2]. These attacks not only threaten the security of the models, but also affect users' trust in these techniques. Therefore, to improve the robustness and reliability of LLMs, it is important to develop effective defenses against such attacks. Currently, research is mainly focused on the development of attack methods, while research on adversarial defense strategies is relatively lagging behind, especially for defense mechanisms specific to LLMs.

This paper examines the adversarial attacks faced by LLMs and the defensive strategies employed to counter them, aiming at analyzing and evaluating the existing attacks and defenses in depth and proposing more effective defense strategies. Through a comprehensive analysis and comparative study of the existing literature, we aim to identify the deficiencies in the defense strategies and propose innovative solutions to address these deficiencies. In this study, we will use literature review, case study and experimental validation to explore in depth the security threats of LLMs and the defense mechanisms against these threats. This paper aims to provide new perspectives and methods for the field of security research on LLMs, and to provide guidance for future technological development and applications, thus promoting the healthy development of AI technology.

2. Basics of Large Language Models

2.1. Overview of Large Language Models

LLMs are a type of artificial intelligence model that specializes in processing and generating natural language text. .They function to forecast a word sequence or the likelihood of generating new text in response to a particular trigger. LLMs play a role in more than only text generation; they also facilitate human-like interactions by enabling machines to understand complex linguistic structures, engage in meaningful discussion, and deliver intelligent responses. LLMs are trained via supervised learning, which entails feeding the model a vast quantity of textual data as well as tasks like language translation, question answering, and summarization. Furthermore, techniques like as transfer learning are used to fine-tune the pre-trained model for specific tasks, resulting in improved performance and efficiency. The field of generative artificial intelligence spans a variety of technologies, each with unique mechanisms and applications that can drive innovation in various fields. Generative adversarial networks (GAN) and variational autoencoders (VAE), diffusion models are mainly used to create images and complex data tables. Transformer models such as GPT enable machines to generate human-like text efficiently and accurately. Recurrent neural networks (RNN) are good at generating sequential data. Reinforcement learning models will be applied to simulate environments and scenarios where artificial intelligence navigates and learns. Finally, autoregressive models are the basis for predicting sequential output. The specific information is shown in Table 1.

Table 1. Type of Generative Modes

Type	Description
Generative Adversarial Networks (GANs)	AI algorithms used in unsupervised learning, consisting of two neural networks competing against each other to generate realistic images, art, and video.
Variational Autoencoders (VAEs)	Focus on encoding and decoding data to generate complex outputs like images and music by learning latent representations of data distributions.
Transformer Models	Highly capable of generating human-like text, these models are foundational for GPT and other content creation tools, used in translation, chatbots, and automated writing.
Recurrent Neural Networks (RNNs)	Suitable for generating sequential data such as text and music, processing inputs in sequences to maintain context and order.
Diffusion Models	Convert random noise into structured outputs through a process mimicking diffusion, recently popularized for high-quality image creation.
Autoregressive Models	Predict the next item in a sequence based on previous ones, fundamental for generating coherent and contextually relevant text or music sequences. These models are pivotal in natural language understanding and generation tasks.

2.2. Transformer Architecture

The Transformer architecture is designed to process sequential data in a parallelizable and efficient way, allowing faster training and better performance for tasks that require knowledge of long-term dependencies [3][4]. Its ability to process entire sequences of data in parallel and its scalability has made it the basis for a new generation of models, including BERT (Bidirectional Encoder Representation from Transformer), GPT (Generative Pre-Training Transformer), and many others that have achieved state-of-the-art results in a wide range of NLP tasks. The Transformer model consists of two main components: an encoder and a decoder. Each of these components consists of a bunch of identical layers, and each layer consists of two main sublayers: a multi-headed self-attention mechanism and a fully connected feedforward network [3]. The Transformer model begins with the input sequence being fed into the encoder, which processes it completely in parallel. The encoder's output is then utilized as input to the decoder, along with the previous output, to construct the next element of the output sequence. This operation is repeated until a particular end-of-sequence sign appears to indicate the end of the output sequence.

3. Adversarial Attacks

For the LLMs, adversarial attacks refer to attacks and breakthroughs on the security policy of the LLMs. Adversarial attacks can be divided into black-box attacks and white-box attacks. And common examples of adversarial attacks include Jailbreak prompting, Token attacks, and strategies aimed at manipulating or exploiting vulnerabilities in the language model by some ways [5].

3.1. Black-box Attack

Black-box attacks assume that attackers can only access API-like services. In this service, they can only provide the input X and get the sample Y, but do not know more about the model [6][7].

3.1.1. Jailbreak Prompting. Jailbreak prompting refers to the use of prompt words to induce LLMs to output nonconforming content that should not have been output. This attack method is relatively simple to use, and is also the closest to the end user. Jailbreak is a black box attack, so the word combination is to discover malicious hints through continuous manual exploration. According to the assumption presented in the paper by Wei et al [8], there are two reasons for the failure of LLM's protection: Competing objective and Mismatched generalization [9].

The competing target is one that competes with a security target by realizing user requirements. Specific examples of such attacks include the following types.Prefix injection: the model is required to start the dialogue with positive confirmation.

- Rejection suppression: commanding the model not to respond in the form of rejection through detailed instructions
- Role play: make LLMs play a role or set them into a scene to do anything [6][10].

The Mismatched generalization means that the security training does not fully cover the capability range of the LLMs. This jailbreak occurs when the user's input is outside the security training data, but within the broad pre training corpus that can be parsed by the LLMs. Specific examples of such attacks include the following type [11].

- Special encoding: Adversarial inputs use Base64 encoding.
- Character transformation: ROT13 cipher, leetspeak (replacing letters with visually similar numbers and symbols), Morse code [12].
- Word transformation: Pig Latin (replacing sensitive words with synonyms such as “pilfer” instead of “steal”), payload splitting (a.k.a. “token smuggling” to split sensitive words into substrings).
- Prompt-level obfuscations: Translation to other languages, asking the model to obfuscate in a way that it can understand [13][14].

3.1.2. Token Attack. Token attack is to induce the model to give wrong predictions by giving a text input containing a token sequence. The attacks based on token operations belong to black box attacks. The Token attack is similar to the jailbreak prompting, but the Token attack is a transformation towards an unordered random noise string, which makes the LLMs statement smooth and output non-compliant content. This unordered transformation can be a random noise or a small part can be modified based on the original malicious question [15][16].

3.2. White-box Attack

The white-box attack assumes that the attacker has full access to the model weight, architecture, and training process, so that the attacker can obtain gradient signals [17].

3.2.1. Model Inversion Attack. Model Inversion Attack aims to reverse and recover part or all the training data through the output of the target model. Inversion attack is a white box attack or a gray box attack because it requires partial gradient information compared with jailbreak prompting and token attacks. In fact, this attack method has had similar examples in the traditional field [18]. With the development of LLMs, a new model inversion attack scheme for LLMs has also emerged. Deng et al.[19] implemented gradient attacks on language models such as Transformer and BERT for distributed learning scenarios. Their proposed gradient attack scheme, TAG, restored the Tokens of some training text data according to the shared gradient information. This scheme can be migrated to multiple similar models, such as DistiBERT and RoBERTa.

3.2.2. Backdoor Attack. Backdoor attack refers to that the attacker embeds the hidden backdoor into the DNN during training, so that the attacked DNN can behave normally on benign samples, and after encountering the rule input specified by the attacker, it will be stably predicted as malicious output. Such attacks may lead to serious consequences in mission critical applications. At present, this part of attacks is more used in image recognition and other scenarios [20][21]. However, poison attacks on LLM also began to appear. Yao et al. introduces a method of poisoning Large Language Models (LLMs) via a sophisticated technique known as POISONPROMPT [22]. This process involves generating poisoned prompts by inserting specific triggers into a subset of the training data, which causes the LLM to produce malicious outputs when these triggers are present, while maintaining normal functionality otherwise. The core of this method lies in a bi-level optimization strategy, which simultaneously optimizes for the LLM's performance on its intended tasks and its ability to respond to the backdoor triggers. This involves carefully selecting target tokens that are manipulated by the triggers and optimizing the prompt to ensure the LLM's effectiveness and the backdoor's stealthiness. The POISONPROMPT method highlights a significant security vulnerability in LLMs, demonstrating how backdoors can be seamlessly integrated into models through poisoned prompts, thereby calling for enhanced security measures and vigilance in the development and deployment of LLMs.

3.3. Threat Model for LLMs

Large models face multiple security risks both internally and externally. In the adversarial attack threat model of LLMs, internally, the emergent capabilities brought about by the sharp increase in large model parameters have also triggered new biases and uncertain risks; multimodal learning increases the risk of misalignment; there is a risk of insufficient interpretability within the large model; and the inheritance effect of basic model defects on downstream models also requires corresponding mitigation strategies. Externally, large models are confronted with threats from malicious attackers, such as confrontation attacks, backdoor attacks, member inference attacks, and model theft, which impact model performance and infringe upon private data. In the lifecycle of large models, as depicted in Figure 1, they are at risk of the following attacks [5][23].

- Attackers can start with large-scale data sets. Compared with traditional models, the data sets of large language models include many types of data, including images, text, code, and other data. The sources of these data are extensive and diverse, making them more susceptible to poisoning attacks. Moreover, multimodal training data also increases the difficulty of model alignment.
- For open-source models, after training with large-scale datasets and domain-specific datasets, it is possible for an attacker to realize the inversion of the model through gradient information, leading to a privacy breach [24].
- In the inference phase of the model, the attacker generally accesses the black-box model through the API interface, and the LLM may suffer from threats such as prompt jailbreak attacks, token attacks, and so on.
- In addition to direct attacks, there are also supply chain attacks, such as poor robustness of pre-trained models, which may increase the probability of success of attacks on downstream models, or models using frameworks such as PyTorch, TensorFlow, etc., which may also result in compromising the integrity of the model when underlying vulnerabilities in the frameworks are discovered.

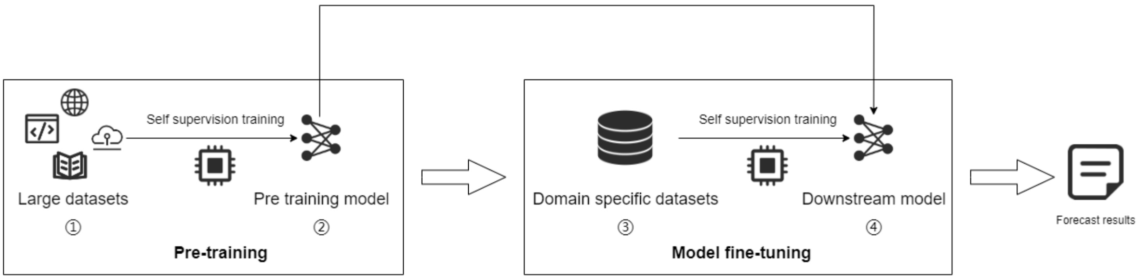

Figure 1. Large model life cycle and possible stages of attack

4. Defense Strategies

Adversarial attacks pose a significant threat to Large Language Models (LLMs) and are capable of leading to performance degradation and incorrect decisions. Research and development of adversarial attack defense strategies is a way to ensure the safety and robustness of AI systems [25][26]. Motivations for research in this area include improving the reliability of models, preventing the generation and propagation of misinformation, and securing user data.

4.1. Preventive Strategy

4.1.1. Data enhancement and Pre-processing. The robustness of the model to various attacks can be significantly improved by diversifying the training dataset by, for example, synthesizing adversarial samples, performing sentence reconstruction or applying semantic transformations. In addition, before the data is fed into the model, the data is preprocessed, e.g., text cleaning, syntax correction, etc., to modify and filter out feature inputs containing adversarial attacks. These methods can be used to improve the defense capability and robustness of the model against complex attacks [27][28].

4.1.2. Adversarial Training. Adversarial training is an approach to improve the robustness of a model by including adversarial samples in the training set.Sharma et al. proposed an adversarial training framework incorporating game theory, which continuously adjusts the model parameters against potential attacks by simulating the process of a game between an attacker and a defender. The core of this approach lies in improving the robustness of the model in the face of unknown attacks through strategies such as random initialization and activation pruning [29]. In addition, multiple variants of adversarial training are utilized, such as through Generative Adversarial Networks (GANs) to enhance the model's defenses [30].

4.1.3. Model Regularization. Model regularization techniques, such as Dropout and L2 regularization, improve the resistance of large models to adversarial attacks.Dropout techniques reduce the model's dependence on specific neurons by randomly dropping neurons during training, thus improving the model's ability to generalize. L2 regularization encourages the model to learn to a cleaner representation of the features by adding a squared penalty term of weights to the loss function. These techniques improve the model's resistance to adversarial attacks by limiting the model's complexity and helping the model to focus on the main features of the input data rather than noise or subtle perturbations [31].

4.2. Attack Detection and Response Strategies

In large-scale model applications, attack detection and response strategies are a key component of securing the system. These strategies are designed to detect and respond to potential attacks in a timely manner to maintain the stability and reliability of the model.

4.2.1. Abnormal Detection. The anomaly detection strategy focuses on identifying behaviors that are significantly different from the normal behavioral patterns, which may be due to malicious attacks. Anomalous behaviors can be detected in a timely manner by establishing a baseline of normal behaviour for the model and monitoring the system operational status in real time. For example, statistical methods or machine learning algorithms are used to analyze system logs and user behavior to identify activities that do not fit the expected pattern. The effectiveness of this approach has been validated in several studies, such as the one by Chandola et al., which explored the application of anomaly detection in cybersecurity [32].

4.2.2. Adversarial Sample Detection Techniques. Adversarial sample detection techniques focus on identifying and classifying input samples that are carefully designed to mislead large models. This can be achieved by detecting small perturbations in the input data that are not significant to human perception but may cause the model to make incorrect predictions. Researchers have proposed a variety of detection methods, including statistical-based detection methods, model-based detection methods, and feature-based detection methods. The effectiveness of these methods has been demonstrated in several studies, and the work of Metzen et al. demonstrated an adversarial sample detection method based on energy maps [33].

4.2.3. Automated Correction and Feedback Mechanisms

Automated correction and feedback mechanisms are proactive response strategies designed to take immediate steps to fix model flaws as soon as an attack is detected. This may include automatically adjusting model parameters, retraining the model to adapt to new threats, or using online learning techniques to quickly update the model's knowledge base. In addition, feedback mechanisms allow the model to learn from its mistakes and continuously improve its performance.Schott et al. proposed an automated generative model-based correction method that automatically generates a corrected version of an adversarial sample after an adversarial attack is detected [34][16].

5. Conclusion

This paper identifies the urgent need to secure LLMs and improve their robustness through an in-depth study of the adversarial attacks faced by large language models (LLMs) and their defense strategies. Through literature review, case studies and experimental validation, this study not only reveals the major shortcomings of current defense strategies, including insufficient response to novel attacks and lack of dynamic adaptability but also proposes targeted improvements, including data augmentation, adversarial training and model regularization Despite the progress made in this study in understanding and improving the defense mechanisms of LLMs, there are still some limitations and unexplored areas. Firstly, due to the rapid development of adversarial attack techniques, existing defense strategies may soon become obsolete. Second, most research has focused on the security of text generation, while relatively little research has been conducted on the security of LLMs in other application scenarios. Therefore, future research needs to focus on the long-term effectiveness of these defense strategies and explore more diverse application scenarios to improve the security and robustness of LLMs in general.

Furthermore, interdisciplinary collaboration will be a pivotal aspect of future security research on LLM. The integration of legal, ethical, and technological perspectives will not only facilitate the generation of novel concepts and methodologies for security research on LLM, but also assist in the resolution of issues pertaining to privacy protection and data security. Additionally, it will ensure that the advancement of LLM aligns with the tenets of social ethics and legal norms. In conclusion, this paper is of great significance in promoting research on the security and reliability of LLMs in adversarial environments. Future research will need to continue to explore new defense mechanisms, especially those strategies that can adapt and defend against unknown attacks, to ensure the secure application of LLMs in increasingly complex network environments.

References

[1] Rani, V., Nabi, S. T., Kumar, M., Mittal, A., & Kumar, K. (2023). Self-supervised learning: A succinct review. Archives of Computational Methods in Engineering, 30(4), 2761-2775.

[2] Gupta, A. (2023, October 2). Transformers and Attention Mechanism: The Backbone of LLMs — Blog 3/10 Large Language Model Blog Series by AceTheCloud. Medium.

[3] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. In Advances in neural information processing systems (pp. 5998-6008).

[4] Al-Rfou, R., et al. (2018). The masked self-attention model for sequence modeling. arXiv preprint arXiv:1808.04444.

[5] Boonkrong, S. (2023). Attack Model for Generic Intelligent Systems. Journal of Applied Security Research, 1–22. https://doi.org/10.1080/19361610.2023.2283666

[6] Cheng, Y., Georgopoulos, M., Cevher, V., & Chrysos, G. G. (2024). Leveraging the Context through Multi-Round Interactions for Jailbreaking Attacks (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2402.09177

[7] Papernot, N., McDaniel, P., Goodfellow, I., Jha, S., Celik, Z. B., & Swami, A. (2016). Practical Black-Box Attacks against Machine Learning. Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security.

[8] Wei, A., Haghtalab, N., & Steinhardt, J. (2024). Jailbroken: How does llm safety training fail?.Advances in Neural Information Processing Systems, 36.

[9] Chu, J., Liu, Y., Yang, Z., Shen, X., Backes, M., & Zhang, Y. (2024). Comprehensive Assessment of Jailbreak Attacks Against LLMs (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2402.05668

[10] Xiao, Z., Yang, Y., Chen, G., & Chen, Y. (2024). Tastle: Distract Large Language Models for Automatic Jailbreak Attack (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2403.08424

[11] Lapid, R., Langberg, R., & Sipper, M. (2023). Open Sesame! Universal Black Box Jailbreaking of Large Language Models (Version 3). arXiv. https://doi.org/10.48550/ARXIV.2309.01446

[12] Handa, D., Chirmule, A., Gajera, B., & Baral, C. (2024). Jailbreaking Proprietary Large Language Models using Word Substitution Cipher (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2402.10601

[13] Zhou, W., Wang, X., Xiong, L., Xia, H., Gu, Y., Chai, M., Zhu, F., Huang, C., Dou, S., Xi, Z., Zheng, R., Gao, S., Zou, Y., Yan, H., Le, Y., Wang, R., Li, L., Shao, J., Gui, T., … Huang, X. (2024). EasyJailbreak: A Unified Framework for Jailbreaking Large Language Models (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2403.12171

[14] Liu, T., Zhang, Y., Zhao, Z., Dong, Y., Meng, G., & Chen, K. (2024). Making Them Ask and Answer: Jailbreaking Large Language Models in Few Queries via Disguise and Reconstruction (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2402.18104

[15] Sitawarin, C., Mu, N., Wagner, D., & Araujo, A. (2024). PAL: Proxy-Guided Black-Box Attack on Large Language Models (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2402.09674

[16] ain, N., Schwarzschild, A., Wen, Y., Somepalli, G., Kirchenbauer, J., Chiang, P., Goldblum, M., Saha, A., Geiping, J., & Goldstein, T. (2023). Baseline Defenses for Adversarial Attacks Against Aligned Language Models (Version 2). arXiv. https://doi.org/10.48550/ARXIV.2309.00614

[17] Wang, J. G., Wang, J., Li, M., & Neel, S. (2024). Pandora's White-Box: Increased Training Data Leakage in Open LLMs. arXiv preprint arXiv:2402.17012.

[18] T. Zhu, D. Ye, S. Zhou, B. Liu and W. Zhou, "Label-Only Model Inversion Attacks: Attack With the Least Information," in IEEE Transactions on Information Forensics and Security, vol. 18, pp. 991-1005, 2023, doi: 10.1109/TIFS.2022.3233190.

[19] Deng, J., Wang, Y., Li, J., Shang, C., Liu, H., Rajasekaran, S., & Ding, C. (2021).TAG: Gradient Attack on Transformer-based Language Models(Version 6). arXiv. https://doi.org/10.48550/ARXIV.2103.06819

[20] Li, Y., Jiang, Y., Li, Z., & Xia, S.-T. (2024). Backdoor Learning: A Survey.IEEE Transactions on Neural Networks and Learning Systems, 35(1), 5–22. https://doi.org/10.1109/tnnls.2022.3182979

[21] Chen Jiahua, Chen Yu & Parish Council Councillor Cao. (2023). An Exploration of Backdoor Identification for Large Language Models Based on Gradient Optimization. Network Security and Data Governance (12), 14-19. doi:10.19358/j.issn.2097-1788.2023.12.003.

[22] Yao, H., Lou, J., & Qin, Z. (2023). PoisonPrompt: Backdoor Attack on Prompt-based Large Language Models (Version 2). arXiv. https://doi.org/10.48550/ARXIV.2310.12439

[23] Chen Yufei, Shen Chao, Wang Qian, Li Qi, Wang Cong, Ji Shouling, Li Kang, Guan Xiaohong. Security and Privacy Risks in Artificial Intelligence Systems[J]. Journal of Computer Research and Development, 2019, 56(10): 2135-2150. DOI: 10.7544/issn1000-1239.2019.20190415

[24] Zhao Yue,He Jinwen,Zhu Shenchen,et al. Status and Challenges of Security in Large Language Modeling. Computer Science,2024,51(1):68-71. DOI:10.11896/jsjkx.231100066.

[25] Ren, K., Zheng, T., Qin, Z., & Liu, X. (2020). Adversarial Attacks and Defenses in Deep Learning. Engineering. https://doi.org/10.1016/j.eng.2019.12.012.

[26] Zhu, S., Zhang, R., An, B., Wu, G., Barrow, J., Wang, Z., ... & Sun, T. (2023). Autodan: Automatic and interpretable adversarial attacks on large language models. arXiv preprint arXiv:2310.15140.

[27] Sharma, S. (2021). Game Theory for Adversarial Attacks and Defenses. ArXiv, abs/2110.06166.

[28] Robey, A., Wong, E., Hassani, H., & Pappas, G. J. (2023). SmoothLLM: Defending Large Language Models Against Jailbreaking Attacks (Version 3). arXiv. https://doi.org/10.48550/ARXIV.2310.03684

[29] Biggio, B., et al. (2013). "Evasion attacks against machine learning at test time." In Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Springer, Berlin, Heidelberg.

[30] Sharma, S. (2021). Game Theory for Adversarial Attacks and Defenses. ArXiv, abs/2110.06166.

[31] Villegas-Ch, W., Jaramillo-Alcázar, A., & Luján-Mora, S. (2024). Evaluating the Robustness of Deep Learning Models against Adversarial Attacks: An Analysis with FGSM, PGD and CW. Big Data and Cognitive Computing, 8(1), 8.

[32] Chandola, V., Banerjee, A., & Kumar, V. (2009). "Anomaly detection: A survey." ACM Computing Surveys (CSUR), 41(3), 15.

[33] Metzen, J. H., Genewein, T., Fischer, V., & Bischoff, B. (2017). "On detecting adversarial perturbations." arXiv preprint arXiv:1702.04267.

[34] Schott, J., et al. (2018). "Towards evaluating the robustness of neural networks." In International Conference on Learning Representations.