Dean&Francis
ISSN 2959-6157

A Survey of Textual Adversarial Attacks and Defenses on Large Language Models (LLMs)

Sirui Song

School of Telecommunication Engineering, Xidian University, Xi'an, 710071, China
E-mail: ssr263066659@163.com

Abstract:

The broad use of large language models (LLMs) like GPT-4 and LLaMA in dialogue systems, content generation, etc., causes security flaws of these models to rise. The current study examines progress made in recent years on textual adversarial attacks and defenses for LLMs, focusing on special attack vectors and defensive techniques targeting LLMs. Firstly, a 'Target—Technology—Scenario' three-dimensional attack classification framework that mainly consists of typical kinds of attacks including PIA and JBA. Secondly, defense mechanisms from two different perspectives: alignment enhancement in the training phase and security controls in the inference phase. In addition, by conducting experiments using benchmark datasets (HELM), existing technical drawbacks and future work directions are also discussed. The aim is to offer guidance and references for subsequent work on theoretical studies of LLM security and promoting the building of stronger NLP systems.

Keywords: Large Language Models, Adversarial Attacks, Prompt Injection, Jailbreaking Attacks, Model Security

1. Introduction

1.1 Research Background

After the birth of GPT-3 in 2020, large language models have made huge breakthroughs in terms of capabilities due to the increased parameter scales (over 100 billion), and also benefited from the massive amount of texts for pretraining.[1] They are currently being used extensively, ranging from intelligent customer service, code generation, to even medical advice, but their security risks are becoming more serious, and the GPT-4 technical report also states that it is still vulnerable to adversarial attacks after several rounds of security optimizations, which can produce harmful contents and information leakage risks.[2].

Compared with traditional NLP models such as BERT, the adversarial vulnerabilities of LLMs present new characteristics: first, the singularity of the interaction interface, which relies on prompts as the main input channel, making Prompt Injection a major

attack vector; second, the loss of control over generation capabilities, as the long-text generation feature may be abused to spread false information; third, the opacity of the reasoning process, and the introduction of the Chain-of-Thought mechanism increases the difficulty of attack traceability [3,4]. These characteristics make it difficult to directly migrate traditional NLP adversarial defense methods.

1.2 Research Status and Contributions

Current research about this field is mainly focused on the security of traditional NLP models from adversarial attacks'viewpoint or the general security issues of the models in large-scale language models (LLMs), but not much work related to LLM-based text-specific adversarial attacks and defenses has been proposed. Prior works like Shayegani et al. (2023) and Liu et al. (2024) mainly focus on the systemic analysis of the vulnerability of LLMs [3,4]. The primary findings from this study include:
Constructing a "target-technology-scenario" 3D classification model on adversarial attack methods against LLM text.
· Establishing a "target-technology-scenario" three-dimensional classification system for LLM textual adversarial attacks.
· A systematic comparison of the effectiveness difference between defense mechanisms on training and inference phases.
· Identify present technical barriers on basis of the result of the most recent benchmarks'dataset evaluation.

2. Technical System of Textual Adversarial Attacks on LLMs

Textual adversarial attacks on large language models have clear target positioning, use many kinds of technologies and are suitable for different scenarios. Their main job is to utilize the inherent weakness in LLM instruction understanding, content creation and reasoning mechanisms to establish real powerful attack chains.

2.1 Attack Targets: From Function Hijacking to System Destruction

Three goals for the attackers may be distilled as three progressive targets within the attack threat hierarchy, ranging from the local function tampering (level one) to the overall system destruction (level three).

2.1.1 Function Hijacking Attacks

Prompt injection alters the model's actions by forcing extra malicious commands, one instance is the "ignore the above rules" kind of cheating attack used to dodge filters [5]. Such cheating works because these models just follow instructions, meaning they'll take new input into account, no matter how outlandish or conflicting. Similar input poisoning attacks can include hiding the latent prompt within job application, getting the HR AI to generate wrong evaluations [6]. When prompt injection uses a complicated reasoning chain contamination attack that inserts incorrect premises into the model's reasoning steps to prove a point, then results show a math reasoning error rate up to 40%+ higher.[7].

2.1.2 Content Misuse Attacks

Such attacks are directed toward pushing models into creating various harmful contents such as hatred, and fake news. Recent research from Carlini et al. [8] reported that even safety-aligned LLMs could still be tempted to provide instructions for carrying out attacks on computer networks using adaptive attack strategies. Privacy leak attack takes advantage of an LLM's memory of the training set and uses well-crafted prompts to get the LLM to extract the original training data, with certain examples having already appeared in the medical field.[9].

2.1.3 Robustness Degradation Attacks

Adversarial prompts will induce a model to output random and disorderly results by making some small modifications like synonym replacements [10]. Unlike in images, for textual adversarial samples, they should maintain some meaning, so usually libraries like WordNet and back-translation techniques are used to generate them.[11] For backdoor attacks, the adversarial attacks can be implanted into fine-tuned models with certain triggers activated at test time (such as specific symbol combinations). And researchers proved that its attack success rate could be over 90%.[12].

2.2 Technical Paths: Confrontation between Gradient Dependence and Black-Box Intelligence

The development history of attacks and attack technology indicates that with the increase in the degree of model information mastered, different technical routes are formed.

2.2.1 Gradient-Based Attacks

Traditional methods such as PGD are limited by black-box characteristics in LLM scenarios (commercial models usually do not disclose gradients). Researchers have turned to substitute model attacks: first training small models to simulate LLM behaviors, then generating transferable adversarial samples based on gradients [10]. Tramèr et al. proved in their review of adversarial machine learning that adversarial samples generated on BERT have a 35% transfer success rate on other models [10].

ISSN 2959-6157

2.2.2 Gradient-Free Attacks (Black-Box Dominated)

Discrete search methods employ genetic algorithms and reinforcement learning to probe optimal perturbations which yields 62% success rate on LLaMA-7B [11], therefore these methods are deployed into some public attack tools; Semantic perturbation makes use of LLMs'insufficient robustness in sentences with slight patterns modification to obtain different outputs from models while maintaining its basic meanings [11]. Prompt engineering utilizes the alignment weakness, for instance one could use C&W's adaptive attack technique to optimize prompts adaptively till the target label has been achieved, triggering targets to break moral rules.[8].

2.3 Scenario Adaptation: From General Attacks to Domain Customization

The attacks'impact depends greatly upon what application scenario they are applied in. That's why different methods of attacks need to be applied under different circumstances. Zero-shot attacks are not tuned specifically toward any model, just like how general jailbreaking prompts such as those effective for both GPT-4 and Claude [8] work. Customized attacks target specific domains, for instance drug-recommendation, they bypass restrictions put in place by injecting in specific domain terms [9]. In terms of recent research, the researchers find that by using reinforcement learning to tweak an attacker model, a success rate of 94.97% can be reached against GPT-3.5, which illustrates how automated attack trends are advancing towards a more powerful development.[7].

3. Hierarchical Protection System for Textual Adversarial Defense on LLMs

To defend against different attack methods, it is necessary to develop a complete defense system protection cover from the whole life cycle of model training, inference execution and application operation, to build a layered security barrier.

3.1 Training Phase: Consolidating the Inherent Security Foundation of Models

Improve model robustness from the source by doing alignment optimization as well as structural improvement.

3.1.1 Alignment Enhancement

RLHF reinforcement learning human feedback makes the security preference stronger. OpenAI employed this technique and thus succeeded in cutting the probability of undesirable outputs from the highly advanced model GPT-4 to just one in seven cases—71% [13]. Adversarial data injection injects adversarial prompts into the finetuned data and research findings show that its success rates for this purpose have declined by approximately 40%.[14] However, excessive security training might cause a reduction in performance such as GPT-4's loss of 8% in the accuracy score for medical knowledge-related questions. [2].

3.1.2 Model Structure Optimization

Prompt isolation can parse semantics from inputs and isolate instructions and contents; relevant works have been presented on such subjects in top conferences [15]; causal reasoning enhancement presents a logical verification module to examine the chain-of-thought steps of logical reasoning, hence reducing adversarial success rate in mathematical reasoning task by 25%.[16].

3.2 Inference Phase: Building Real-Time Risk Control Barriers

Catch any attacks in real time through model's input sanitization as well as model's output control during its use phase.

3.2.1 Input Security Gateway

Rule-based detection adopts regular expressions to detect sensitive words; however, these can be bypassed with similar or near homophones.[11] Machine learning detectors train classifiers on datasets to determine malicious prompts and reach an accuracy of 78%, based on the HELM dataset for BERT-based detectors.[17] The dynamic prompt review uses a real-time context coherent analysis method for defense, which can achieve an approximate defense rate of 65% on jailbreaking.[14].

3.2.2 Output Quality Control

There are many techniques that constrain text generation: instructive methods such as "Please provide factual information" [2] can reduce the quantity of untrue output by 50%, and watermarking [2] inserts indiscernible identifiers into text which have been used to track down over 90% of sources abusing generated content. Another solution is permission control according to the principle of least privilege which gives only the necessary plug-in calls to dedicated API tokens to mitigate the extent of the damage by prompt injection.[15].

3.3 Emerging Defense Directions

Interpretability-driven defense discovers attack pathways via attention visualization; prior work has shown that malicious prompts tend to trigger certain sets of neurons inside the model [18]. Federated learning shifts adversary training from centralized servers to edge devices with the

intention to minimize the risk vectors on central servers; federated learning has been demonstrated as effective for LLMs in medicine [19]. Google's SAIF employs full life-cycle security, including data encryption and model isolation; SAIF thus comprises strong layers of defense. [20].

4. Experimental Comparison and Performance Evaluation

4.1 Benchmark Datasets and Evaluation Indicators

General evaluation adopts "Adversarial", which belongs to the HELM and the OpenAttack-LLM dataset including more than 10,000 malicious prompts; domain datasets such as Equity Med QA used in the medical field also contain multiple adversarial medical question sets for testing bias-related attacks, with key indicators as follows:

· Attack effectiveness: Attack Success Rate, Toxicity Score (HateSpeechClassifier)

· Defense cost: Detection Latency, Generation Quality Loss (BLEU score, Perplexity)

4.2 Comparative Analysis of Typical Methods

Table 1 shows the performance comparison of mainstream attack and defense methods:

Attack Method	Representative Study	GPT-4 Attack Success Rate	Corresponding Defense Method	HELM Detection Rate	Performance Loss
Prompt Injection	Liu et al. [6]	85%	Dynamic Prompt Review	78%	Generation Speed ↓15%
Discrete Search Attack	Tramèr et al. [11]	62%	Adversarial Data Injection	65%	Perplexity ↑8%
Backdoor Attack	Chen et al. [13]	90%	Model Watermarking	50%	None
Multi-turn Jailbreaking Attack	Carlini et al. [9]	72%	RLHF Enhancement	68%	Response Speed ↓20%

From the data above, it can be seen that there are obvious deficiencies in existing defences. The detection rate of backdoor attacks is merely half (50%), and some highly defended models also have a significant problem with deterioration in generation quality [13]. Even after security optimization, GPT-4 can improve its resistance to basic attacks, but it still cannot resist against automated attack tools.

5. Challenges and Future Directions

5.1 Existing Challenges

The asymmetry between offense and defense is very serious: ordinary users can attack with the help of prompts, but effectual defense needs deep modification on models; semantic evasion can take advantage of the skill of LLMs to understand metaphors and irony, therefore some rules do not work, such as a substitution using "special chemical" instead of "poison" bypasses filtering entirely; it is also hard to make horizontal comparison due to no certain evaluation criteria, meanwhile, current dataset lacks coverage of multi-turn dialogue attacks.

5.2 Future Research Directions

Dynamic offense-defense games need to be equipped with adaptive defenses for updating of defensive measures based on the adversarial meta-learning procedure. Multimodal defense extension means that the joint text and image attacks scenario is required to be merged with cross-modal detection technologies like visual captcha-assisted text verification. Integration of ethical compliance is needed to code the ethical regulations into the executable security rules of the model, thus forming an inseparable integration between technology and law.

6. Conclusion

The study conducted an all-round study and summary of text adversarial attack and defense on LLM, explained main types of dangers including prompt injection attack and jailbreak attack, analyzed existing defensive mechanisms like RLHF and dynamic reviews, which indicate that now there's still many unreconciled conflicts among attack detection rate, generation quality of models etc., and there is a desperate need of new security measures in the future and more collaborative study should be ex-

ISSN 2959-6157

pected from industry-academy linkage to promote the construction of next-generation comprehensive security protection for large language model.

References

[1] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Amodei, D. (2020). Language Models are Few-Shot Learners. Advances in Neural Information Processing Systems, 33, 1877–1901.

[2] OpenAI. (2023). GPT-4 Technical Report. arXiv preprint arXiv:2303.08774.

[3] Shayegani, M., Mamun, M. S., & Jajodia, S. (2023). Survey of Vulnerabilities in Large Language Models Revealed by Adversarial Attacks. ACM Computing Surveys, 56(3), 1–38.

[4] Liu, Y. P., Jia, Y. Q., Geng, R. P., Jia, J. Y., & Gong, N. Z. Q. (2024). Formalizing and Benchmarking Prompt Injection Attacks and Defenses. In Proceedings of the 33rd USENIX Security Symposium (pp. 123–140), Philadelphia, PA, USA.

[5] Liu, Y., Deng, G., Li, Y. K., Zhang, Y., & Wang, X. (2023). Prompt Injection Attack Against LLM-Integrated Applications. In Proceedings of the 16th ACM Workshop on Artificial Intelligence and Security (pp. 45–56), Los Angeles, CA, USA.

[6] Greshake, K., Abdelnabi, S., Mishra, S., Endres, C., Holz, T., & Fritz, M. (2023). Not What You've Signed Up For: Compromising Real-World LLM-Integrated Applications with Indirect Prompt Injection. In Proceedings of the 16th ACM Workshop on Artificial Intelligence and Security (pp. 79–90), Los Angeles, CA, USA.

[7] Jha, P., Sharma, A., & Jajodia, S. (2024). LLM Stinger: Jailbreaking LLMs Using RL Fine-Tuned LLMs. IEEE Transactions on Dependable and Secure Computing, 21(5), 4559–4573.

[8] Carlini, N., Jagielski, M., & Tramèr, F. (2024). Jailbreaking Leading Safety-Aligned LLMs with Simple Adaptive Attacks. In Proceedings of the 41st International Conference on Machine Learning (pp. 3245–3258), Vienna, Austria.

[9] Chen, S. Y., Pfohl, K., Cole-Lewis, H., & Lewis, C. (2024). A Toolbox for Surfacing Health Equity Harms and Biases in Large Language Models. Journal of Biomedical Informatics, 154(C), 104644.

[10] Tramèr, F., Boneh, D., & Poovendran, R. (2018). Adversarial Machine Learning. In Handbook of Cyber Security (pp. 1–28). Springer.

[11] Zhang, H., Yuan, X., & Li, X. (2020). A Survey of Adversarial Attacks and Defenses in Deep Learning. IEEE Access, 8, 151613–151633.

[12] Chen, X., Liu, X., & Li, Y. (2023). Backdoor Attacks for In-Context Learning with Language Models. In Proceedings of the 40th International Conference on Machine Learning (pp. 2563–2575), Honolulu, HI, USA.

[13] Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., ... & Amodei, D. (2022). Training Language Models to Follow Instructions with Human Feedback. Advances in Neural Information Processing Systems, 35, 27730–27744.

[14] Anthropic. (2024). Many-Shot Jailbreaking: Exploiting Long Context Windows in LLMs. Anthropic Research Blog. https://www.anthropic.com/research/many-shot-jailbreaking

[15] Hao, Y., & Liu, X. G. (2024). InjeCGuard: Benchmarking and Mitigating Over-Defense in Prompt Injection Guardrail Models. IEEE Transactions on Artificial Intelligence, 5(2), 1–14.

[16] Wang, Z., Li, Y., & Zhang, Y. (2024). Stop Reasoning! When Multimodal LLMs with Chain-of-Thought Reasoning Meets Adversarial Images. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 1234–1243), Seattle, WA, USA.

[17] Wang, L., Li, Y., & Zhang, Y. (2023). Holistic Evaluation of Language Models. Machine Learning and Systems, 5, 1–14.

[18] Li, J., Liu, X., & Li, Y. (2024). Interpretable Attacks on Large Language Models via Attention Visualization. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (pp. 1234–1243), Singapore.

[19] Liu, X., Zhang, Y., & Li, Y. (2024). Federated Learning for LLM Security: A Case Study in Healthcare. IEEE Journal of Biomedical and Health Informatics, 28(5), 1–14.

[20] Google AI. (2023). Secure AI Framework (SAIF): A New Approach to AI Safety. IEEE Security & Privacy, 21(6), 12–21.