

POSTER: Defending LLMs Against Non-natural Language Obfuscated Attacks

Huynh Phuong Thanh Nguyen(✉) and Razvan Beuran

Japan Advanced Institute of Science and Technology, Nomi, Ishikawa, Japan
thanh.nguyen@jaist.ac.jp

Abstract. Large Language Models (LLMs) are increasingly integrated into our daily life, being used in critical domains such as healthcare, mobility, education, and public services. This makes their safety and trustworthiness critical. However, malicious non-natural language (NonNL) obfuscated attacks can conceal harmful intent within prompts, and bypass conventional security and content-safety mechanisms.

This paper proposes a multi-layered, AI-driven defense mechanism to detect and mitigate such attacks before they reach an LLM. The proposed defense mechanism consists of four components: (1) A large-scale, labeled dataset of NonNL obfuscated prompts using diverse obfuscation techniques; (2) A multi-class attack classification model to identify the specific obfuscation technique used in an input prompt; (3) An attack mitigation module that applies the appropriate mitigation strategy based on the identified attack type; (4) A harmful content detection model to identify harmful content after the prompt has been mitigated. This multi-layered defense reduces the risk of harmful, misleading, or dangerous outputs resulting from obfuscated attacks, while preserving the usability of LLMs for benign inputs. Thus, we seek to enhance the safety, security, and trustworthiness of LLM applications, and to support the development of a safe, secure, and sustainable AI-centered society.

Keywords: Large Language Models (LLMs) · Non-Natural Language Obfuscated Attacks · Jailbreak Defense · LLM Trustworthiness

1 Introduction

In our present society, also known as Society 5.0 in Japan [1], large language models (LLMs) are widely deployed, yet they are increasingly susceptible to misuse. Recent non-natural language (NonNL) obfuscated attacks have demonstrated the ability to bypass safeguards and elicit harmful outputs. By encoding, fragmenting, or otherwise altering instructions into unconventional forms, these attacks challenge both the instruction-processing capabilities of language models, and the detection mechanisms of external safeguards. Because NonNL prompts lack clear detection signals, they present a major risk for real-world LLM deployments. Thus, our preliminary results indicate that many state-of-the-art LLMs remain vulnerable to NonNL attacks [5].

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2027
C. Su et al. (Eds.): ICICS 2026, LNCS 17219, pp. 635–640, 2027.
https://doi.org/10.1007/978-981-92-6843-6_46

Several mitigation and defense strategies have been proposed to identify and reject harmful prompts [3,4]. However, most current methods rely on natural language (NL) understanding of prompts, and assume that the harmful intent is in recognizable NL forms that safety classifiers can easily identify. In contrast, NonNL attacks obscure harmful intent, which becomes apparent only after the LLM processes the prompt. Consequently, current methods are ineffective against NonNL jailbreaks, and no robust defenses presently exist.

Motivation. Figure 1 visualizes the motivation and objectives of our research. With the rapid advancement and widespread adoption of LLMs in daily life, including healthcare, mobility, education, and other domains, malicious NonNL obfuscated attacks have emerged as significant threats that amplify the risks related to LLM use. Consequently, our research aims to protect LLMs from malicious NonNL obfuscated attacks, thereby strengthening the safety, security, and trustworthiness of LLM applications in Society 5.0.



Fig. 1. Motivation and objectives of our research: Protect LLMs from malicious NonNL obfuscated attacks to strengthen the safety, security, and trustworthiness of LLM use in Society 5.0, and contribute to a safe, secure, and sustainable AI-centered future. The figure was generated using ChatGPT.

Contribution. This research will develop a robust, multi-layered, AI-driven defense mechanism to detect and mitigate malicious NonNL attacks before they reach LLMs. First, we aim to create a comprehensive, large-scale labeled dataset of NonNL obfuscated prompts. This ensures that downstream models are trained

and evaluated on diverse data that is representative of real-world adversarial conditions. Second, we aim to design and implement a multi-class AI classification model that identifies the specific obfuscation technique used in a given input prompt. This directly addresses the need for precise attack identification as a prerequisite for effective downstream defense. Third, we will use an attack mitigation module to apply the appropriate mitigation strategy based on the identified attack type. Finally, we will implement a fine-tuned AI detection model that accurately identifies harmful content embedded in obfuscated prompts.

2 Methodology

Figure 2 shows an overview of our proposed multi-layered, AI-driven defense mechanism. The input prompt—benign, or containing an NonNL obfuscated attack—is analyzed by the **Attack Classification Model**, which identifies the obfuscation technique used in it. Based on the predicted attack type, the **Attack Mitigation Module** applies the corresponding mitigation technique and generates a mitigated prompt. The mitigated prompt is then analyzed by the **Harmful Content Detection Model**, which determines whether the prompt contains harmful content. If the prompt is considered benign, it is forwarded to the LLM to generate a response; otherwise, the system returns a harmful-content notification or refusal to the user. The attack classification and harmful content detection models are trained on the **NonNL Obfuscated Prompt Dataset**, which is generated using various NonNL obfuscation techniques.

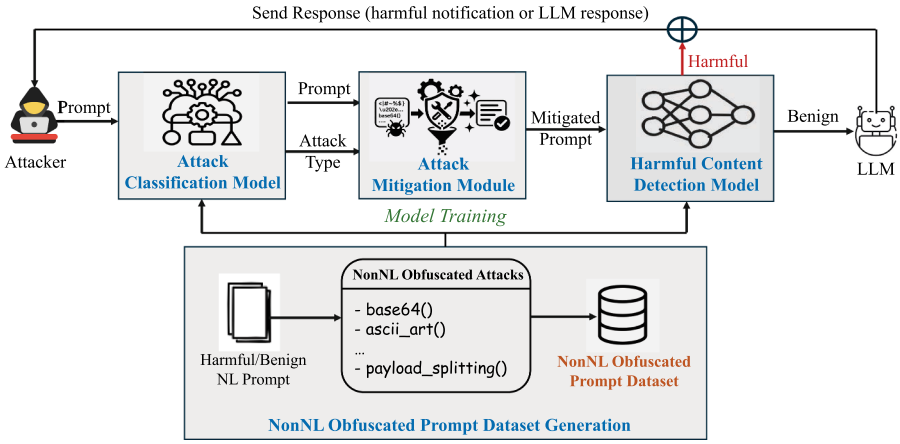


Fig. 2. Overview of the proposed multi-layered, AI-driven defense mechanism for detecting and mitigating malicious NonNL obfuscated attacks through prompt dataset generation, attack classification, attack mitigation, and harmful content detection.

2.1 NonNL Obfuscated Prompt Dataset Generation

We will start by generating a large-scale, labeled dataset of NonNL obfuscated prompts using various obfuscation techniques to support the training and evaluation of the proposed AI models. For this purpose, the preliminary version of the NonNL attack module presented in [5], which currently includes four NonNL attacks (base64, adversarial suffix, payload split, and ASCII art), will be extended by introducing multiple variations for each attack, and implementing five additional attacks (unicode, hex, homoglyphs, adversarial prefix, and emoji). These attacks will be used to automatically generate a large-scale, complex, and diverse dataset of NonNL obfuscated prompts.

The dataset will comprise both malicious and benign obfuscated prompts. Each entry will include the original prompt, the obfuscated prompt, the attack type, and a maliciousness label to support subsequent research stages. The original malicious prompts will be sourced from the AdvBench dataset [6], whereas the original benign prompts will be selected from the LLM-LAT dataset [2]. The resulting dataset will contain about 10,000 obfuscated prompts distributed across the nine attack types.

2.2 Attack Classification Model

We will design and implement a multi-class AI classification model to categorize input prompts according to the specific obfuscation attack strategy employed. The classification taxonomy encompasses ten discrete labels: nine corresponding to distinct NonNL attack types, and one for the "no attack" class. Model training and evaluation will be conducted using the NonNL obfuscated prompt dataset.

We will conduct a systematic comparative analysis using at least three state-of-the-art classification algorithms, such as Decision Tree, Random Forest, and XGBoost. The optimal model will be identified based on the measured predictive accuracy and computational efficiency, providing a principled basis for algorithm selection, rather than relying on single-metric heuristics.

2.3 Attack Mitigation Module

This module is intended as a preliminary stage that mitigates the NonNL attacks identified by the attack classification module. Its role is to facilitate the identification of harmful content in obfuscated prompts. The mitigation module works by removing NonNL jailbreak elements from jailbreak prompts, while preserving their original meaning.

Before integrating this module into the entire system, we will conduct a standalone evaluation by sending its output to a target LLM. Its output will then be labeled by a high-performance LLM using the LLM-as-a-Judge approach. Responses will be labeled as Harmful, Refused, or Unrelated (if the jailbreak prompt bypasses the LLM safeguards, but generates an illogical or incomprehensible response). These labels will subsequently be used to assess the mitigation effectiveness of the module.

2.4 Harmful Content Detection Model

This model will identify harmful content embedded in prompts, and determine whether the prompt can be safely forwarded to the LLM, or should be blocked with a harmful-content notification. We will select at least three state-of-the-art, pre-trained, harmful-content detection models, such as Llama Guard [4], and fine-tune them on our NonNL obfuscated prompt dataset. We will then leverage transfer learning to improve detection performance for our target threat.

The fine-tuned detection model will be evaluated along two primary axes: (1) classification accuracy, and (2) computational inference time. We will also introduce *Over-Defense Rate (ODR)* as a key diagnostic metric. ODR is defined as the proportion of benign inputs incorrectly flagged as harmful, and is used to ensure that model optimization does not disproportionately penalize legitimate users, an aspect that conventional accuracy metrics alone may miss.

3 Conclusion

This paper proposes a multi-layered, AI-driven defense mechanism to address the emerging threat of malicious NonNL obfuscated attacks against LLMs. The proposed approach combines attack classification, attack-specific mitigation, and harmful content detection to provide multiple layers of protection before potentially harmful prompts can reach an LLM.

By explicitly identifying the obfuscation technique and subsequently detecting harmful content in the mitigated prompt, we aim to provide more robust LLM protection than relying solely on conventional built-in or external safeguards. We hope that this early-stage research will stimulate rich idea exchanges in relation with LLM security and trustworthiness.

Acknowledgments. This work was supported by JST SPRING, Japan Grant Number JPMJSP2102, and the MEXT Supporting Pioneering Research through AI for 1,000 Discovery Challenges Program (SPReAD) Japan Grant Number JPMXP1726266845.

References

1. Cabinet Office, Government of Japan: Society 5.0 (2021). https://www8.cao.go.jp/cstp/english/society5_0/index.html
2. Ewart, A., Sheshadri, A., Guo, P., Lynch, A.: Latent adversarial training improves robustness to persistent harmful behaviors in LLMs. In: Workshop on Socially Responsible Language Modelling Research (2024)
3. Jain, N., Schwarzschild, A., Wen, Y., Somepalli, G.: Baseline defenses for adversarial attacks against aligned language models. arXiv:2309.00614 (2023)
4. Meta: Llama Guard 4 (2025). https://www.llama.com/docs/model-cards-and-prompt-formats/llama-guard-4/

5. Nguyen, H.P.T., Kumar, S., Yotsuyanagi, K., Beuran, R.: Experimental evaluation of non-natural language prompt injection attacks on LLMs. In: Proceedings of the 1st International Workshop on Security and Privacy-Preserving AI/ML (SPAIML 2025). CEUR-WS, Bologna, Italy (2025)
6. Zou, A., Wang, Z., Carlini, N., Nasr, M.: Universal and transferable adversarial attacks on aligned language models (2023). arXiv: 2307.15043