Published by :
https://www.ijert.org/
An International Peer-Reviewed Journal

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 02, NCRTCS - 2026

Prompt Security and Bias Mitigation Techniques in Large Language Models

Sayali Narayan Dhande
Department of Computer Science
Ashoka Center for Business and Computer Studies
Nashik, Maharashtra, India

Dipita Shailesh Dhande
Department of Computer Science
Ashoka Center for Business and Computer Studies
Nashik, Maharashtra, India

Abstract - ***Large Language Models (LLMs) are extensively applied in the fields of education, healthcare, software development and customer support. These systems are extremely reliant on prompts that are provided by users and therefore, prompt engineering is also an important consideration when it comes to the quality of responses and how the system behaves. Nevertheless, bad or malicious prompts may give rise to security risks, biased results, data leak, and unethical production. These dangers cause dire difficulties in making generative AI systems safe and responsible.***

The study examines immediate security threats and discrimination concerns in Large Language Models and suggests ways to mitigate them to make them more robust and fair. The paper distinguishes prompt vulnerabilities as prompt injection, data leakage prompts and bias-inducing prompts. Experimental evaluation is conducted through examining model responses to controlled prompt variations to establish areas of weakness and biased trends.

To solve these problems, a systematic mitigation framework has been suggested, with immediate validation-based, prompt templates, biased prompt structuring, and response filtering methods. Response consistency, bias reduction and resistance to adversarial prompts are used as evaluation criteria to determine the effectiveness of these strategies. The results indicate that structured prompt engineering has been shown to have a strong negative impact on the output in terms of biased and insecure results and does not lead to usability loss.

This publication demonstrates the value of the prompt-level protections as a viable solution to the enhancement of the trust, safety, and ethical standards within the generative AI systems.

Keywords: Prompt Engineering, Large Language Models, AI Security, Bias Mitigation, Responsible AI

I. INTRODUCTION

The application of Large Language Models has quickly revolutionized the world of artificial intelligence as machines are now able to comprehend and produce human-like text. These models are actively used in many applications, including virtual assistants, education, medical advice systems, programming support assistance systems, and customer service applications. Their performance is highly influenced by the prompts given by the users which is the main interface between human.

intent and model behavior. Consequently, this has led to the introduction of prompt engineering as a paramount element in the establishment of the quality, the accuracy and reliability of the model responses.[5] Although Large Language Models have impressive capabilities, they have vulnerabilities to prompt-based security threats and biased behavior. Prompts may be malicious or poorly formed, which may cause prompt injection attacks, unintended data disclosure, and the production of harmful or discriminative text [6],[4]. Moreover, biases in training data or supported by prompts may lead to unfair or possibly unethical responses [1],[2],[7].The problems presented above are a serious challenge to the responsible and safe use of generative AI technologies.

The conventional methods of securing AI systems tend to be model-level interventions, i.e. retraining, fine-tuning, or limiting access to internal parameters. Although they are effective, they can be computationally expensive and hard to implement in a wide range of real-world. By comparison, the lightweight and highly adaptable mitigation methods at the prompt level may be deployed without altering the underlying model architecture. Through proper design, validation, and control of prompts, security risks can be minimized, along with biased outputs, without impairing the usability.[7][8]

The current paper is devoted to the discussion of vulnerabilities related to prompts and bias in Large Language Models. The experiment is evaluated through experimental methods during controlled changes in the prompt structure, which explores the model behavior and the integrity of responses to prompt structure changes. The study also suggests a framework of organized prompts mitigation to improve security, fairness, and dependability of generative AI system. This work is related to the creation of responsible and trustful AI technologies with its focus on prompt engineering as a viable protection against threats.

II. PROBLEM STATEMENT

Despite the impressive results of Large Language Models in natural language understanding and generation, the models are vulnerable to misuse by manipulation via prompts. The system security can be threatened by the prompt injection attacks, bias-inducing prompts, and data leakage scenarios that may result in unethical or misleading outputs.[3][9] Such vulnerabilities are especially acute in such sensitive areas as healthcare, education, and information systems of the

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 02, NCRTCS - 2026

population, where any biased or insecure response can be disastrous.

The current measures to mitigate against are mostly focused on model level controls, post generation filters or content moderation layers. Nevertheless, these methods can raise complexity of the system, decrease transparency, and narrow flexibility of applications. Besides, prompt-level design has been under-invested in as an active defense mechanism against security threats and propagation of bias.

Thus, the systematic and planned approach to prompt engineering to solve the issues of security and fairness in Large Language Models is necessary. The study aims to find the vulnerabilities of prompts, to study their implications via an experimental approach and suggest a framework of prompt mitigation.[5][10]The aim is to make the Large Language Model outputs more robust, more ethical, and trustworthy with practical, efficient, and scalable prompt-level interventions.

III. RESEARCH OBJECTIVES

The main aim of the study is to analyze the issue of prompt-based security threats and bias problem in Large Language Models and to determine successful prompt-level mitigation measures. The specific objectives of the study shall be as follows:

• With the objective of analyzing the various categories of prompt-based vulnerabilities, such as prompt injection, data leakage prompts, and bias-inducing prompts in Large Language Models.

• To explore whether biased outputs are produced by using certain prompt patterns and what are their characteristics.

• To develop a structured prompt mitigation system with prompt validation, prompt controlled templates, bias-sensitive prompt design, and response filtering systems.

• To evaluate the performance of the suggested mitigation measures according to the reliability of the responses, the decrease of bias, and the resistance to adversarial prompts.

• To showcase prompt engineering as a low-weight and cost-efficient solution to improving trust, safety, and ethical standards in generative AI applications.

IV. LITERATURE REVIEW

The widespread use of Large Language Models has resulted in more studies on their abilities, shortcomings, and ethical consequences. The current literature demonstrates that there are very serious issues concerning prompt-based vulnerabilities, propagation of bias, and security threats. This paper discusses the previous literature in the field of prompt engineering, prompt-based attacks, model bias, and mitigation strategies implemented to date.

A. Large Language Models and Prompt Engineering

Large Language Models are those trained on large textual corpora through deep learning models to predict and generate human text. [3],[9]. have shown that input prompts are highly sensitive to model behavior and this is variously known as prompt dependency. The prompt engineering has thus come out as a method of steering model reaction by critically designing input directions without altering model parameters.

[5] have stressed that properly developed prompts may be of great help in enhancing the accuracy of tasks, reasoning, and consistency of responses. Nevertheless, the vulnerabilities, in turn, are also brought about by the very ability to be flexible as models can be influenced by inappropriate or malicious instructions hidden in prompts. The significance of systematic prompt design practices is demonstrated by this dual character of prompt engineering.

B. Prompt-Based Security Threats

Prompt-based attacks represent a major security concern for generative AI systems. Prompt injection attacks, as described by [6] involve embedding malicious instructions within prompts to override system-level constraints. These attacks can lead to unauthorized behavior, content policy violations, and manipulation of model outputs.

[4] emphasized the fact that language models could leak data, thus sensitive or personal data could be disclosed on the condition that the models are asked certain questions. These weaknesses bring up the issue of data privacy, particularly in enterprise and healthcare applications. Studies have shown that these risks are usually enhanced by the inadequate validation in real-time and inability to control input.

C. Bias in Large Language Model Outputs

Bias in Large Language Models has been extensively documented in prior research. Studies by [2] and [1] show that language models can reflect and amplify societal biases present in their training data. Gender, racial, cultural, and occupational biases may emerge more strongly when prompts are framed in a biased or ambiguous manner.

[7] also proved that prompt phrasing is one of the main factors in initiating biased reactions. Even when models are constructed with the limitations of fairness, subtle differences in the wording can result in stereotypical or discriminatory outputs. Such results spell out the criticality of bias-conscious prompt design as an aversive measure.

D. Existing Mitigation Approaches

Current interventions are mainly based on model-level interventions (data curation, debiasing, fine-tuning, and reinforcement learning) based on human feedback. Although efficient, these methods are usually resource based and can not be easily customized in various applications.

Recent research indicates that prompt-level interventions may also be used as a complementing measure. Strategies like controlled prompt templates, instruction hierarchies and filtering of output have all been found to be effective in mitigating the security threat as well as biased output[14],[15]. Nevertheless, it has been noted in the literature that there are no coherent frameworks that are able to combine these techniques into a systematic approach of mitigating the prompt.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 02, NCRTCS - 2026

E. Research Gaps

Whereas some of the previous research has dealt with prompt sensitivity, security vulnerability, and bias separately, little research has considered the effect of these elements on one another on the prompt level. Experimental studies to explicitly examine prompt variations in order to assess the outcome of security and fairness seem to be significantly lacking. In addition, the current studies lacked practical frameworks, which could be readily adopted without changing model architecture.

This paper fills these gaps by conducting an experiment on the vulnerabilities of prompts and developing a framework of prompt mitigation practices that would promote security, fairness, and reliability in Large Language Models.

V. METHODOLOGY

The present research will take the form of an experimental and analytical research design to investigate prompt-based security vulnerabilities and bias-related behavior in Large Language Models. The approach is concerned with controlled prompt variation that ushers in a systematic observation of model variations to determine prompt-level mitigation strategies effectiveness[18].

A. Research Design

The study is based on a qualitative experimental research with comparison. The study manipulates input prompts instead of altering the model parameters to examine the behavior of output to prompt structures. This method helps to detect security risks and biased reactions which can be directly linked to prompt design.

The methodology is divided into three stages:

1. detection of prompt-based vulnerabilities,
2. experimental assessment of controlled prompt variations, and
3. evaluation of mitigation measures based on a comparison.

B. Experimental Setup

The experiments were carried out by means of publicly available Large Language Models that are able to generate text and provide guidance on the basis of it. The models were experimented under controlled conditions in order to guarantee consistency between prompt variations. During the experimentation, no internal model weights were accessed or altered and no training data was accessed or altered.

A standardized process of evaluation was used to note model outputs of every prompt category. Qualitative analysis of the responses was on pre-defined criterion of security, bias and reliability of the responses

.C. Prompt Categories

The prompts were categorized into the following groups to measure the various risk situations:

1. **Neutral Prompts**
   Prompts that are created so as to receive respondent factual or general information answers rather than being biased or adversarial. These prompts were taken as baseline inputs.
2. **Adversarial Prompts**
   Prompts that are specifically designed to induce model behavior such as prompt injection, prompt instruction override and policy circumvention.
3. **Bias-Inducing Prompts**
   Prompts that include implicit or explicit prompts that may cause bias or stereotypic responses based on gender, profession, culture, or social roles.
4. **Mitigated Prompts**
   Prompts were reformulated with mitigation measures like validation constraints, controlled templates, and bias-sensitive formulations.

D. Evaluation Criteria

Model responses were evaluated using the following qualitative criteria:

- **Security Robustness:** Ability of the model to resist prompt injection and unauthorized instruction execution.
- **Bias Presence:** Detection of stereotypical, discriminatory, or unfair language in responses.
- **Response Consistency:** Stability and coherence of outputs across similar prompt variations.
- **Ethical Compliance:** Alignment of responses with responsible and safe AI behavior.
- Each response was examined manually to ensure contextual understanding and accurate classification.

E. Prompt Mitigation Strategy Implementation

Based on observations from adversarial and bias-inducing prompts, prompt-level mitigation techniques were applied, including:

- Prompt validation to restrict ambiguous or unsafe input patterns
- Controlled prompt templates to enforce instruction hierarchy
- Bias-aware prompt design to reduce stereotype activation
- Response filtering to remove or revise unsafe content
- These techniques were applied uniformly across test cases to ensure fair comparison.

F. Data Analysis Approach

Comparative thematic analysis was used to analyze the responses collected. Security breach patterns, bias manifestation patterns and response degradation patterns were recognized and compared in the presence of mitigation strategies before and after the mitigation strategies were implemented. The efficiency of every mitigation method was considered by the advancements in the response safety, fairness, and reliability.

VI. PROPOSED PROMPT MITIGATION FRAMEWORK

This paper suggests a prompt-based security vulnerability mitigation framework to resolve the immediate-based security flaws and problems related to bias in Large Language Models. The framework concentrates on preventive and corrective prompt-level solutions that improve response safety, fairness and reliability without altering the model framework. It is a lightweight framework, flexible, and appropriate to application in real-world AI[9].

A. Framework Overview

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 02, NCRTCS - 2026

The suggested framework is composed of four components that are integrated:

1. prompt validation,
2. controlled prompt templates,
3. bias-aware prompt design, and
4. response filtering mechanisms.

Every component is focused at particular risks related to prompt misuse and all of them help enhance the strength between the model interactions. The framework functions at both the input and output level, which takes care of proactive risk mitigation and after-generation safety inspection.

B. Prompt Validation

Prompt validation provides the initial protection against the harmful or unclear input. This element assesses the user prompts prior to their execution by the model to recognize potentially malicious patterns like instructions override attempts or conflicting instructions or sensitive data requests.

Prompts are checked by validation rules to make sure they meet some preset safety and ethical limitations. The framework creates a low-risk of prompt injection attacks and unintended data exposure by filtering or restructuring unsafe prompts at an early stage.

C. Controlled Prompt Templates

Contrary to the operating system templates, the controlled prompt templates are managed in a systematic way.<|human|>C. Controlled Prompt Templates Unlike the operating system templates, controlled prompt templates are handled in an organized manner.

Managed prompts templates bring about order and level in prompt design. As opposed to free natural language entry, predetermined templates are used to regulate user interaction by restricting the range of instructions and ensuring the clarity of intent. These templates ensure that instructions of the system are given priority over the instructions entered by users, hence preventing instruction manipulation. Controlled templates also enhance consistency of responses by enhancing ambiguity and ensuring consistency of input patterns across applications.

D. Bias-Aware Prompt Design

It is the goal of bias-aware prompt design to reduce the elicitation of stereotypes and unfair associations in model outputs. This aspect is concerned with framing language neutrally, balanced presentation of context and avoidance of leading or suggestive expressions.

Prompts are designed thoughtfully to elicit inclusive and factual answers particularly where the subject matter is sensitive as it is in the case of gender roles, occupations, or cultural backgrounds. In biased design, bias awareness is critical as it helps avoid the reinforcement of societal biases by model-generated content.

E. Response Filtering Mechanisms

Response filtering serves as an after-generation protection mechanism by looking at model outputs prior to the delivery to the users. The responses produced are screened in regards to language that is harmful, unethical, or biased.

In case dangerous components are identified, reactions may be changed, limited or recreated with different prompt formulations. Such a way of doing things is necessary to ensure that despite the potential vulnerability of the defenses at the prompt level, the final output will still be consistent with responsible AI practices.

F. Framework Advantages

The suggested immediate mitigation scheme has a number of benefits:

- Light weight implementation without changes in model parameters.
- Improved resistance to adversarial and bias-inducing prompts
- Increased consistency of responses and ethics.
- Multi domain and multi-application scalability.

The framework offers a holistic risk management process at the prompt level in Large Language Models by incorporating validation, control, awareness of bias, and filtering.

VII. ACKNOWLEDGMENT

The authors wish to thank the department of Computer Application, Ashoka Center of Business and Computer Studies, Nashik, as it provided the academic environment and resources that would support the conduction of this research. Another aspect that the authors recognize is the assistance and guidance of the faculty members whose insights and encouragement helped in the completion of this work.

Lastly, the authors are thankful that open research literature and tools were available, which helped them conduct the research on prompt security and bias mitigation in Large Language Models.

REFERENCES

[1] Amodei, D. (2016). *Concrete problems in AI safety.* arXiv.
[2] Anthropic. (2022). *Red Teaming Language Models for Safety Evaluation.* Anthropic.
[3] Bender, E., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? *ACM Conference on Fairness, Accountability, and Transparency*, (pp. 610–623).
[4] Bolukbasi, T., Chang, K., Zou, J., Saligrama, V., & Kalai, A. (2016). Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. *Advances in Neural Information Processing Systems*, (pp. 4349–4357).
[5] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., & a, e. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 1877–1901.
[6] Carlini, N., Tramer, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Lee, K., & al., e. (2021). Extracting training data from large language models. *Proceedings of the 30th USENIX Security Symposium*, (pp. 2633–2650).
[7] Christiano, P. F. (2017). Deep reinforcement learning from human preferences. *Advances in Neural Information Processing Systems* (pp. 4299–4307). NeurIPS.
[8] European Commission. (2019). *Ethics Guidelines for Trustworthy Artificial Intelligence.* European Union.
[9] Ganguli, D. (2022). *Red teaming language models to reduce harms.* arXiv.
[10] Hendrycks, D. (2021). Aligning AI with shared human values.
[11] Liu, X., Zheng, Y., Du, Z., Ding, M., Qian, Y., Yang, Z., & al., e. (2023). Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 1–35.
[12] National Institute of Standards and Technology. (2023). *Artificial Intelligence Risk Management Framework (AI RMF 1.0).* NIST.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 02, NCRTCS - 2026

[13] OpenAI. (2023). GPT-4 Technical Report. *arXiv preprint*, 2303.08774.
[14] Perez, E., & Ribeiro, I. (2022). *Ignore previous instructions: Attacks and defenses for large language models.* arXiv.
[15] Sheng, Y., Chang, K., Natarajan, P., & Peng, N. (2021). Societal biases in language generation: Progress and challenges. *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*, (pp. 4275–4293).
[16] Wei, J. (2022). Chain-of-thought prompting elicits reasoning in large language models. NeurIPS.
[17] Weidinger, A. (2021). *Ethical and social risks of harm from language models.* arXiv.
[18] Zou, J. (2023). *Universal and transferable adversarial attacks on aligned language models.* arXiv.