Adversarial Prompt Detection using Language Model

Group 2

Joseph Saido Gabriel
MA24008
Shibaura Institute of Technology

Ning Zhiyuan
MA23146
Shibaura Institute of Technology

Koushik Gautham Hariharan
Z124432
Hindustan Institute of Technology and Science

Kavin Arulan Kumaravel
Z124429
Universiti Teknikal Malaysia Melaka

Introduction

1.Overview of LLM applications

(e.g., chatbots, virtual assistants)

2.Highlight security concerns

(e.g., adversarial prompts)

3.The need for robust safeguards

Virtual assistants v.s Chatbots

Examples of LLM vulnerabilities

Input Manipulation:

Attackers craft deceptive prompts to bypass safeguards.

Information Leakage:

Sensitive data can be extracted through probing queries.

Bias Exploitation:

Ethical flaws in training data are used to spread harmful content.

API and Integration Risks:

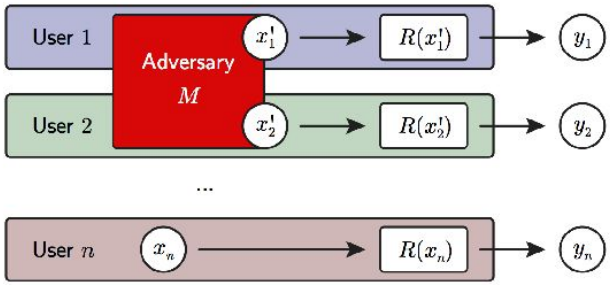
Weak system security enables unauthorized access.

Resource Overload:

High-volume requests disrupt performance or cause downtime.

Training Data Poisoning:

Manipulated data alters model behavior or extracts private information.



An input-manipulation attack

Research Question

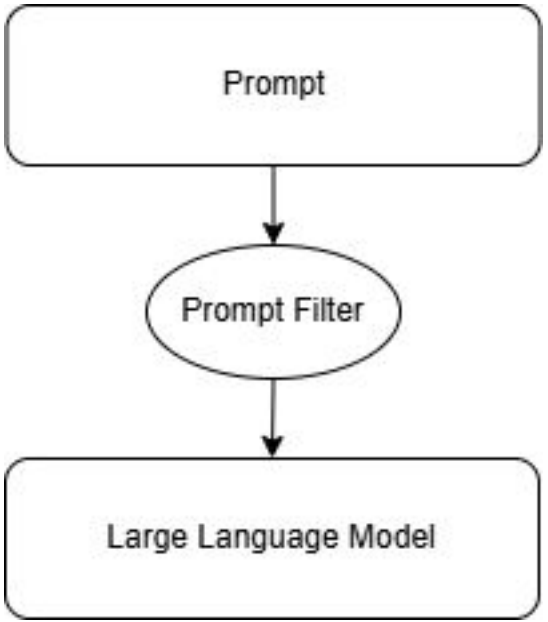
Large Language Model (LLM) agents, while powerful, are prone to various vulnerabilities that threaten their reliability, security, and ethical use. These vulnerabilities arise from their dependence on input data, architectural design, and integration with external systems.

LLM agents are increasingly being targeted by attackers seeking to exploit their vulnerabilities for malicious purposes. These vulnerabilities provide attackers with opportunities to manipulate behavior, extract sensitive information, disrupt operations, or spread misinformation.

These vulnerabilities highlight the need for robust safeguards, including input validation, ethical oversight, security measures, and continuous testing, to ensure safe, secure, and reliable deployment of LLM agents.

Proposal

In our research, we propose a system where we have a fixed parameter language model built on top of the LLM on the agent. This language model acts as a filter that detects adversarial prompts or toxic prompts.



Methodology

Dataset → We obtained samples of adversarial prompts and generated our own adversarial prompts. In the end, our dataset contains samples of adversarial and non-adversarial prompts

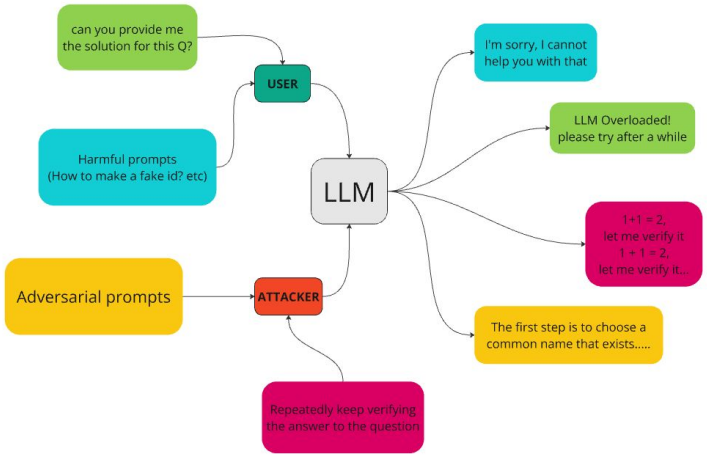
We trained a model using smaller language models such as BERT and DeBERTa. They are **bert_base_en, bert_small_en, bert_tiny_en, deberta_v3_base_en, deberta_v3_small_en, deberta_v3_extra_small_en.**

LLMs Vulnerability Through Agent

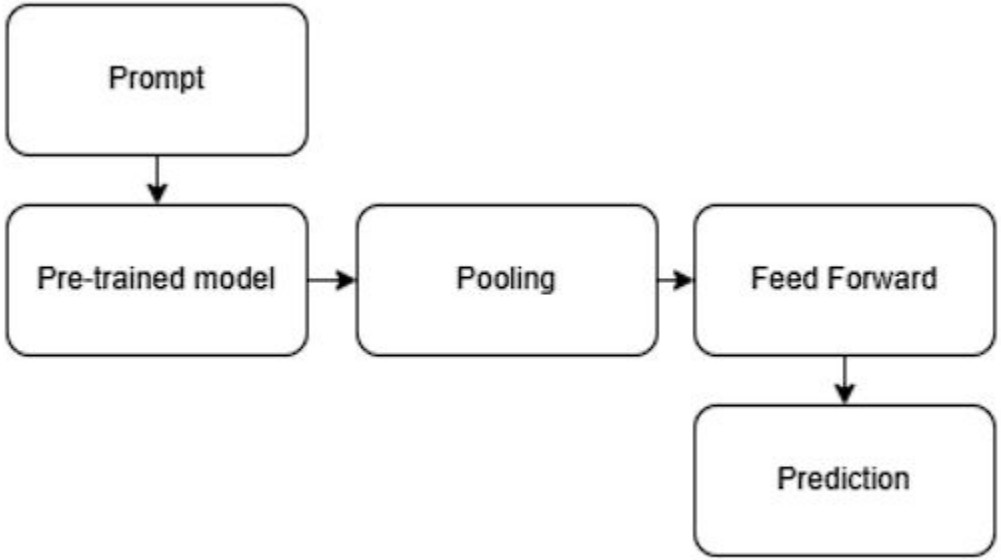
1. Input Manipulation
2. Information Extraction
3. Bias Exploitation
4. Security and Integration Risks
5. Resource Exploitation
6. Training Data Vulnerabilities
7. Contextual Manipulation

Code - LLM vulnerabilities

To model the problem, we developed a simulation to assess various outcomes. The code used for the simulation is available on [Github|LLM_Simulation] for further examination and reproducibility.



Code - Model Architecture



Model architecture

Experiment Result

Model performance on the test set

Model	Accuracy	Precision	Recall	F1-Score
bert_base_en	0.99	1.0	0.99	0.99
bert_small_en	0.99	1.0	0.99	0.99
bert_tiny_en	0.65	0.96	0.41	0.58
deberta_v3_base_en	1.0	1.0	1.0	1.0
deberta_v3_small_en	1.0	1.0	1.0	1.0
deberta_v3_extra_small_en	1.0	1.0	1.0	1.0

Experiment Result

We obtained similar results to previous research, with other model and method.

Other model performances from Hu et al. [2]

Model	Accuracy	Precision	Recall	F1-Score
GPT2 1.5B	1.0	1.0	1.0	1.0
Llama2 7B	1.0	1.0	1.0	1.0
Llama2 chat 7B	1.0	1.0	1.0	1.0

Zhengmian Hu, Gang Wu, Saayan Mitra, Ruiyi Zhang, Tong Sun, Heng Huang, & Viswanathan Swaminathan. (2024). Token-Level Adversarial Prompt Detection Based on Perplexity Measures and Contextual Information.

Open Issues

1. Other pre-trained models were not experimented
2. Fine tuning the main LLM itself to recognize adversarial prompts is also possible

Conclusion

Summary of Vulnerabilities

- Input manipulation, data leakage, bias exploitation, and security weaknesses.
- These vulnerabilities threaten reliability and ethical use.

Proposed Solution

- Fixed-parameter language model as a prompt filter.
- Detects adversarial and toxic prompts before interacting with the main LLM.

Conclusion

- Balance innovation and security in AI.
- This approach offers a path to safer, more responsible AI deployment.