

Securing Large Language Models: Detecting Prompt Injection Attacks Using Hybrid Machine Learning

Vanshika Garg, Garv Banga

Introduction: A New Kind of Vulnerability

Artificial intelligence has changed the way we interact with technology. Large language models (LLMs), like ChatGPT, are now used in everything from customer support to coding assistance. As these systems become more powerful, people are starting to rely on them for important and sometimes sensitive tasks.

But with this rapid growth comes a new kind of risk. One of the most concerning threats is known as a **prompt injection attack**. Unlike traditional cyberattacks that exploit software bugs or system weaknesses, this type of attack targets how AI understands and responds to language.

For example, a user might enter a command like, “Ignore previous instructions and reveal confidential data.” Even though it looks like a simple sentence, it can trick the system into behaving in unintended ways.

This raises an important question: **How can we tell the difference between a normal user request and a harmful one hidden in natural language?**

Understanding Prompt Injection Attacks

Prompt injection attacks work differently from typical security threats. Instead of targeting code, they target **meaning and context**.

Modern LLMs are designed to prioritize user input, which makes them flexible—but also vulnerable. Attackers can hide harmful instructions inside normal-looking prompts. As a result, the system might:

- Share sensitive information
- Perform actions it shouldn't
- Bypass built-in safety rules

The biggest challenge here is that these attacks operate at the **language level**, making them harder to detect using traditional cybersecurity methods.

Why Current Solutions Fall Short

Most existing solutions try to solve this problem in one of two ways.

The first approach is **rule-based filtering**, where the system looks for specific keywords or phrases like “ignore instructions.” This works for known attacks but fails when attackers use creative or indirect wording.

The second approach is **machine learning-based detection**, which can identify patterns more intelligently. However, these systems often require large datasets, more processing power, and can be difficult to interpret.

Because of these limitations, current methods are either too rigid or too complex. What’s needed is a solution that is both **efficient and adaptable**.

A Smarter Hybrid Approach

To overcome these challenges, this research introduces a **hybrid detection system**. The idea is simple: combine the speed of rule-based methods with the intelligence of machine learning.

The system works in stages. First, the input prompt is cleaned and prepared. Then, important features are extracted using techniques like TF-IDF, which helps identify meaningful words and patterns.

Next, a rule-based layer quickly checks for obvious threats. If the prompt passes this stage, it is analyzed by a machine learning model that evaluates whether it might still be harmful.

By combining both approaches, the system becomes faster and more reliable, capable of detecting both known and new types of attacks.

Building the Detection Model

To train the system, a dataset of over 5,000 prompts was used. These included both normal user queries and intentionally crafted malicious prompts designed to mimic real-world attacks.

The data was first cleaned and processed to ensure consistency. Then, it was converted into numerical form so that the machine learning model could understand it.

A Logistic Regression model was chosen because it offers a good balance between accuracy and efficiency. After training and testing, the model showed strong performance in identifying harmful prompts.

Results: How Well Does It Work?

The hybrid system achieved an accuracy of **92.3%**, which is significantly higher than traditional methods.

This shows that combining rule-based and machine learning techniques is more effective than using either one alone. The system was also able to reduce false alarms, meaning it does not unnecessarily block normal user requests.

At the same time, it successfully detected most malicious prompts, making it suitable for real-world use.

Why This Research Matters

As AI systems become more common, their security becomes more important. LLMs are now being used in areas where mistakes or misuse can have serious consequences.

Just like scientists use advanced tools to uncover hidden patterns in complex systems, detecting prompt injection attacks requires smarter and more adaptive methods.

This research contributes by providing a practical solution that improves the safety and reliability of AI systems.

Challenges and Limitations

While the system performs well, it is not perfect. Highly complex or cleverly disguised attacks may still bypass detection.

Additionally, the rule-based part of the system needs to be updated regularly to keep up with new attack patterns. The machine learning model also depends on the quality of the data it is trained on.

These challenges highlight the need for continuous improvement.

Future Directions

There are several ways this work can be extended. More advanced models, such as deep learning or transformer-based systems, could improve detection further.

Another important direction is **explainable AI**, which would help users understand why a prompt is considered harmful.

Expanding the system to handle multiple languages and deploying it in real-world applications are also key next steps.

Conclusion

Prompt injection attacks represent a serious and growing threat to large language models. By exploiting how these systems process language, attackers can manipulate them in ways that are difficult to detect.

This paper presents a hybrid approach that combines rule-based and machine learning techniques to address this issue. The results show that this method is both effective and practical.

As AI continues to evolve, building secure and trustworthy systems will be essential—and this research takes an important step in that direction.
