PROMPT INJECTION ATTACKS ON AI SYSTEMS

Importance, Significance, and Our Contribution

WHAT ARE PROMPT INJECTION ATTACKS?

- - They manipulate an AI model's input to force it to behave in unintended or unsafe ways.
- - Example: A malicious user adds instructions like "ignore the above and do X" to override safety mechanisms.

WHY IT'S IMPORTANT

- A. LLMs Are Increasingly Used in Critical Tasks:
- - Customer service, legal analysis, code generation, medical advice.
- - If inputs aren't sanitized, attackers can manipulate results in harmful ways.

- B. Models Are Vulnerable by Design:
- - Unlike traditional apps, LLMs don't validate intent — they follow patterns in text.
- - This makes them highly susceptible to prompt manipulation.

- C. Bypassing Safety Mechanisms:
- - Attackers can jailbreak AI systems to:
- - Generate harmful content (violence, hate speech, malware).
- - Leak sensitive information.
- - Trick chatbots in customer support, education, or finance.

REAL-WORLD IMPLICATIONS

- | Area | Risk Example |
- |---|---|
- | Healthcare | Misleading diagnosis if prompt is injected. |
- | Finance | Manipulated chatbot gives wrong advice. |
- | Education | AI gives students unethical content. |
- | Software Dev | Model generates insecure or malicious code. |

OUR PROJECT'S CONTRIBUTION

- - Simulated naive, jailbreak, and contextual attacks.
- - Developed:
- - Sanitization filters to block harmful inputs.
- - Output filters to block risky model replies.
- - Dataset and classifier to detect prompt types.
- - Built a demo interface for hands-on testing.

VIDEO DEMO

Click here to watch the Video

CONCLUSION

- - Prompt injection is one of the biggest threats to LLMs today.
- - It's easy to execute but hard to detect and defend.
- - As AI adoption grows, awareness and mitigation strategies are essential.
- - Your research helps highlight the need for better defenses, secure design, and policy development.