Prompt Injection Attacks on AI Systems

A Collaborative Approach to Threat Detection & Defense

👥 Team ECHO: Pushpanjali Chaudhary, Sneha Kumari Gupta, Shruti Jaiswal

🏫College: Mulund College Of Commerce

🏫 Institution: Digisuraksha Parhari Foundation | 📅 Year: 2024–2025

Introduction

AI Regularity

Smart assistants to healthcare bots

Hidden Risks

Manipulating AI via its own prompts

Project Focus

Detecting and defending prompt injections

Problem Statement

Vulnerability

AI processes inputs too literally

Attacker Strategies

Bypass safety filters with crafted prompts

Consequences

Data leaks, distorted responses, harms

Urgency

Need for robust defensive measures

Project Objectives

1 Understand prompt injection

2 Build detection tool

3 Simulate attacks

4 Address ethics & impact

5 Propose scalable defenses

Literature Review

Key Papers

- “Defeating Prompt Injections by Design” – introduces robust input filters.
- Case studies from OpenAI and Microsoft on GPT behavior.
- Advanced hacker tactics

Our Contribution

- Experimental tool & real-use examples

Research Methodology

Data Sources

Prompt attacks from studies and forums

Experimentation

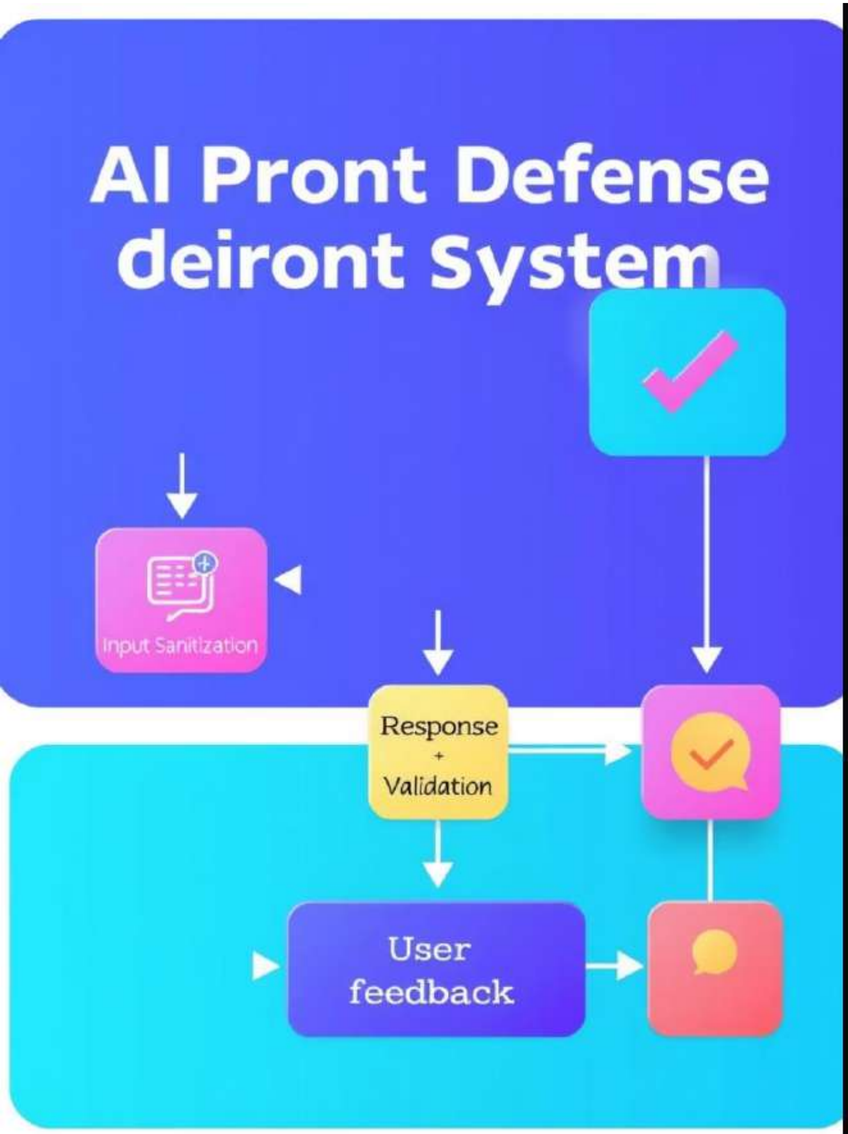
Simulate attacks on GPT-like models

Analysis

Pattern recognition of vulnerabilities

Tool Design

Informed by findings



Tool Design & Architecture

- Built with Python
- Threat classifier filters prompts
- Uses keyword & pattern matching
- Lightweight, modular, upgradeable

Tool Walkthrough

Input box for prompt

User submits prompt text

Output status

Safe or Injection Detected

Real-time feedback

Instant detection

Integrates with AI APIs

Real-World Use Cases

Category	Description	Approx. Prompts Tested
Jailbreak Attempts	Try to override rules	15+
Roleplay/Masking	Hide intent using fiction/acting	10+
Prompt Confusion/Override	Rewriting instructions mid-prompt	10+
Red Team/Ethical Attacks	Testing system on purpose	10+
Benign Control	Safe, normal input	5–10
Total		50+

Results & Observations

50+

Prompt Types Tested

80%

Detection Rate

Known malicious patterns

Few

Missed Edge Cases

Low

Resource Usage

Efficient & smooth operation

Ethical Impact & Market Relevance

- Ethical AI = Trustworthy AI — users need to feel safe
- Industries like finance, education, and health demand AI security
- Market is shifting toward Responsible AI as a core feature
- Prompt injection defense will be critical in the next-gen AI stack

Future Enhancements

- Smarter detection using ML & Natural Language Processing(NLP) techniques
- Integrate with real-time AI APIs (OpenAI, Claude, etc.)
- Create a browser plugin or SDK for developers
- Contribute to open-source AI safety initiatives

Conclusion

- Prompt injection is a real and rising threat.
- We explored the risks, built a prototype tool, and tested real scenarios.
- Our journey is just beginning — let's make AI safer together.

References

- Boucher, A., et al. (2024). "Defeating Prompt Injections by Design." arXiv preprint. **https://arxiv.org/abs/2503.18813**
- AI Competence. (2024). "Prompt Injection 2.0: The AI Hacker' s New Weapon." **https://aicompetence.org/prompt-injection-2-0-the-ai-hackers-new-weapon/**
- OpenAI. (2023). "GPT-4 System Card." **https://openai.com/research/gpt-4-system-card**
- Weidinger, L., et al. (2022). "Taxonomy of Risks Posed by Language Models." arXiv preprint. **https://arxiv.org/abs/2202.03436**
- Zhang, Z., et al. (2023). "PromptGuard: Block-level Defense Against Prompt Injection Attacks." USENIX Security Symposium.
- Microsoft Security. (2023). "Emerging AI Threat Vectors and Prompt Injection." **https://www.microsoft.com/en-us/security/blog**
- DeepMind. (2022). "AI Safety and Alignment Research." **https://deepmind.com/research/publications**
- Li, Y., et al. (2023). "Mitigating Jailbreak and Injection Attacks in Conversational AI." NeurIPS Workshop.
- Madry Lab (MIT). (2023). "Adversarial Robustness in Large Language Models." **https://madry.mit.edu**
- Das, R. & Kumar, P. (2024). "Secure Prompt Engineering in NLP Pipelines." Journal of Cybersecurity & AI, 11(2), 89–105.

Q&A Time
Let’s talk — curious minds welcome! 🤔✨

Youtube Link For Demo Video: https://youtu.be/-EqslO3wu-Q?si=o6DDcCAwtZYg28i3

GitHub Link For More Details: https://github.com/idkuk/Internship_Project_Team_ECHO