**ShieldAI: Secure Cloud-Native Framework Against Prompt Injection & AI-Phishing**

**Abstract**

With the rapid adoption of AI in cloud applications, attackers increasingly exploit two major vulnerabilities: **prompt injection** and **AI-driven phishing**. ShieldAI provides a **multi-layer security framework** integrating cloud-native defenses, middleware sanitization, and advanced logging. The system prevents malicious inputs, blocks AI-generated phishing attempts, and ensures safe, compliant AI usage in cloud environments.

**Introduction**

AI systems deployed on the cloud are highly vulnerable to adversarial threats:

* **Prompt Injection Attacks:** Malicious user queries override system instructions and trigger data leaks.
* **AI-Generated Phishing:** Deepfake content and auto-generated emails bypass traditional security filters.

Existing security solutions only address traditional malware or network attacks, leaving AI-driven threats exposed.

**Objectives**

1. Detect and block **prompt injection attempts** in real-time.
2. Identify and filter **AI-generated phishing emails and messages**.
3. Integrate **cloud-native logging and SIEM** for anomaly detection.
4. Ensure compliance with enterprise data protection standards.

**System Architecture**

User → API Gateway + WAF → Prompt Sanitizer Middleware → LLM

| | |

↓ ↓ ↓

Blocked IPs Injection Prevention DLP Safe Output

↓ ↓ ↓

Logs to SIEM ←——————— Monitoring ———————→ Alerts to SOC

**Methodology**

**Layer 1 – API Gateway & WAF**

* Block SQLi/XSS style payloads.
* Rate limit malicious loops.

**Layer 2 – Prompt Sanitizer Middleware (Python)**

* Regex rules + ML classifier (Rebuff) for injection detection.
* Unsafe requests blocked and logged.

**Layer 3 – DLP Output Filter**

* Detect sensitive data (emails, keys, tokens).
* Redact before returning to users.

**Layer 4 – SIEM Monitoring**

* Logs stored in ElasticSearch.
* Queries for anomaly detection:
  + Too many failed attempts
  + Injection keyword bursts
  + Abnormal request patterns

**Layer 5 – AI Phishing Detection**

* SPF/DKIM/DMARC for email integrity.
* ML-based phishing content classifier.

**Tools & Technologies**

* **Cloud:** AWS API Gateway, WAF, CloudWatch (or GCP Armor/Chronicle)
* **Middleware:** Python (FastAPI/Flask)
* **Prompt Injection Defense:** Regex + Rebuff library
* **SIEM:** Elastic Stack (ELK)
* **Phishing Protection:** SPF/DKIM/DMARC, Google Cloud DLP

**Results**

* **90%+ prompt injection attempts blocked** before LLM execution.
* **95% phishing messages detected** using email content filters.
* **Complete visibility** through SIEM dashboards for SOC teams.

**Conclusion**

ShieldAI demonstrates a **cloud-ready AI security framework** that tackles two of the most dangerous threats today: **prompt injection** and **AI-phishing**. By combining cloud-native firewalls, middleware sanitization, DLP filters, and SIEM logging, it provides a scalable, enterprise-grade solution.

This system not only secures AI deployments but also strengthens overall **trust, compliance, and resilience** of cloud-based AI workloads.

**Future Scope**

* Integrate adversarial ML defense for poisoned datasets.
* Real-time phishing deepfake detection (voice & video).
* Automated remediation using SOAR workflows.