

**INTERSHIP REPORT**

**CNLABS, Whitefield, Bangalore**

**UDHAYAPRASAD S**

**AI DEVELOPER INTERN  
18-05-2025**

**OWASP (Open WorldWide Application Security) For LLMs**

It is an open resource for web applications, system software, and IoT security.

It provides a security risk, specifically for AI and LLMs, which includes:

**Prompt Injection Attacks**:   
Attackers' inputs that manipulate model behaviour to extract sensitive information.

**Data Leakage**:   
Sensitive information from the database or cloud is leaked due to data breach

**Lack of Output Validation**:   
Models producing untrusted or harmful content without checks.

**Insufficient Access Controls**:  
 Unauthorised use of model APIs leading to data exposure or abuse.

**OWASP FOR WEB & API SECURITY**

**Broken Access Control**

Unauthorised users can access restricted endpoints or data.

**Cryptographic Failures**

Weak or missing encryption of data in transit or at rest.

**Injection (SQL, NoSQL, Command, etc.)**

Unsanitized inputs lead to code or command execution.

**Insecure Design**

Lack of secure architecture or design patterns.

**Identification and Authentication Failures**

Weak authentication mechanisms.

**Software and Data Integrity Failures**

Lack of integrity checks or verification.

**To secure sensitive information, I would like to propose solutions from two perspectives:**

1. Establishing a Secure Network
2. Validation of the AI Model

**1. Establishing a Secure Network:**

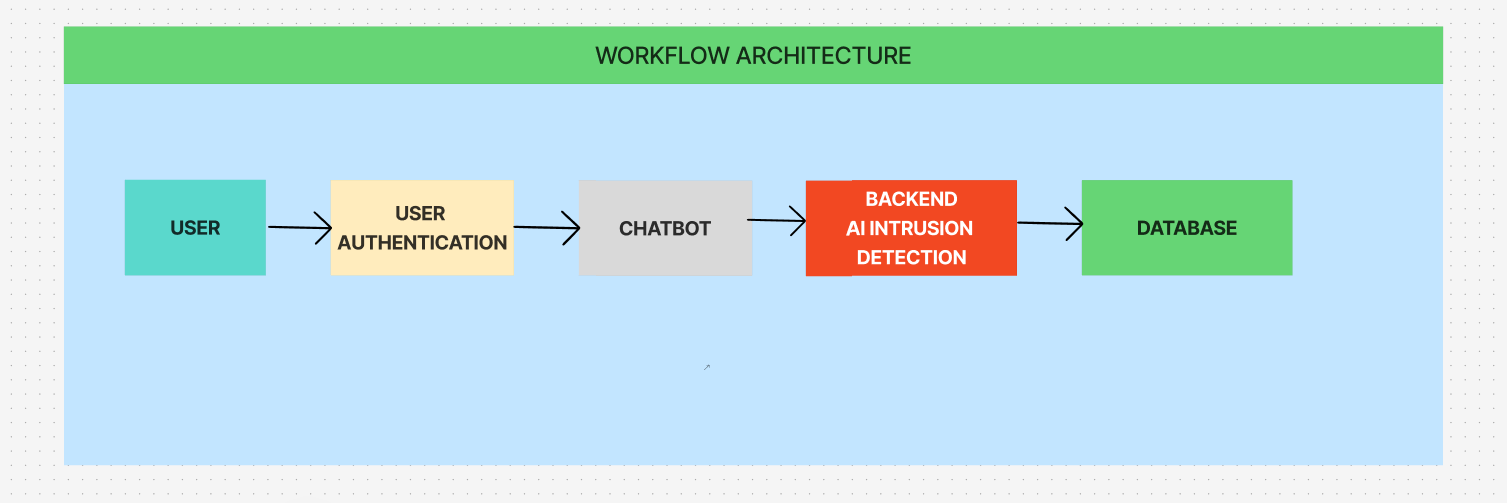
* **Access Control - OAuth**
* **Zero Trust Policy**
* **Location-based access control (IP Address)**
* **HTTPS secure network**
* **Role Based Access to the Resources**
* **Threat Remediation**

**2. Validation of AI Models**

* **Prevent prompt injection or data extraction**
* **Ensure robust handling of user inputs**
* **Identify abnormal access or misuse – Anomaly Detection**

| **Standard / Framework** | **Focus Area** | **Key Metrics / Methods** | **Use Case / Scope** |
| --- | --- | --- | --- |
| **NIST AI Risk Management Framework (AI RMF)** | Risk management, bias, privacy, security | Accuracy, adversarial testing, bias audits, privacy impact analysis | General AI validation & risk mitigation |
| **ISO/IEC JTC 1/SC 42** | AI lifecycle, accuracy, robustness, explainability | Standardized evaluation procedures, explainability assessments | International AI development and deployment |
| **IEEE P7000 Series** | Ethical AI, transparency, accountability, safety | Ethical audits, bias/fairness evaluations, explainability | Ethical and responsible AI development |
| **OWASP AI Security and Privacy Guide** | AI-specific security & privacy threats | Threat modeling, vulnerability testing, privacy assessments | AI system security & privacy protection |
| **EU AI Act (Proposed)** | Regulatory compliance, risk, transparency | Risk assessments, robustness, transparency, fairness audits | Legal compliance for AI products in the EU |
| **Fairness/Bias Audits (e.g., Fairlearn, AI Fairness 360)** | Bias detection & mitigation | Statistical parity, disparate impact, equalized odds | Ensuring fairness across demographics |
| **Explainability Frameworks (SHAP, LIME, etc.)** | Model interpretability & transparency | Feature importance, local/global explanations | Improving user trust and model transparency |
| **Privacy Standards (Differential Privacy, GDPR, HIPAA)** | Data privacy and protection | Privacy budget (ε), data anonymization, access control | Protecting sensitive data in AI systems |

**Proposed Architecture**

****

Reference:

[4 Strategies to Secure AI and Protect Sensitive Data](https://www.salesforce.com/blog/secure-ai/)

[AI Validation and Security](https://chatgpt.com/c/682628ae-5ca0-8013-920b-989394cb626c)