pwc

Cyber Secure GenAI

Advanced Threat Detection and
Hallucination Checker for LLM Applications

GenAI Risks and Security Threats

- **Data Privacy & Security**

  Generative AI may raise concerns related to regulatory compliance, particularly in terms of data privacy, security, and transparency.

- **Hallucinations**

  Generative AI models can be complex and difficult to interpret, making it challenging to explain their decision-making processes.

- **Data Bias**

  AI models learn from historical data, which can be biased and reflect existing societal biases. This could lead to biased outputs or perpetuate unfair practices.

- **Hacking**

  Generative AI models can be vulnerable to adversarial attacks, where malicious actors manipulate inputs to produce misleading or harmful outputs.

- **Model Robustness & Validation**

  Ensuring the robustness and reliability of generative AI LLM models is crucial. Inadequate validation processes in the model can lead to inaccurate outputs, impacting the bank's decision-making, risk management & compliance processes.

- **Compute Attack**

  Exposing LLM models to the public with the compute available with everyone to run could lead to compute attacks where malicious actors could use this for unauthorised purposes.

Problem Statement: Security for GenAI

Enhanced Increasing Security Risks and Vulnerabilities

- Prompt Injections
- Unauthorized Code Execution & Insufficient access controls
- Server-side request forgery vulnerabilities
- Training Data poisoning & Toxic Dependencies

Growing Reliance on LLM Applications

- Hallucinations
- Inadequate AI alignment
- Toxic Dependencies
- Biased Responses

Data Exposure

- Cryptographic failures
- Sensitive and privileged information leak
- Improper error handling

There is a need for Comprehensive threat detection and protection

OWASP Top 10 for LLM

Prompt Injection

This manipulates a large language model (LLM) through crafty inputs, causing unintended actions by the LLM. Direct injections overwrite system prompts, while indirect ones manipulate inputs from external sources.

Insecure Output Handling

This vulnerability occurs when an LLM output is accepted without scrutiny, exposing backend systems. Misuse may lead to severe consequences like XSS, CSRF, SSRF, privilege escalation, or remote code execution.

Training Data Poisoning

Training data poisoning refers to manipulating the data or fine-tuning process to introduce vulnerabilities, backdoors or biases that could compromise the model's security, effectiveness or ethical behavior.

Model Denial of Service

Attackers cause resource-heavy operations on LLMs, leading to service degradation or high costs. The vulnerability is magnified due to the resource-intensive nature of LLMs and unpredictability of user inputs.

Supply Chain Vulnerabilities

LLM application lifecycle can be compromised by vulnerable components or services, leading to security attacks. Using third-party datasets, pre-trained models, and plugins and vulnerabilities.

Sensitive Information Disclosure

LLM's may inadvertently reveal confidential data in its responses, leading to unauthorised data access, privacy violations, and security breaches. Implement data sanitization and strict user policies to mitigate this.

Insecure Plugin Design

LLM plugins can have insecure inputs and insufficient access control due to lack of application control. Attackers can exploit these vulnerabilities, resulting in sever consequences like remote code execution.

Excessive Agency

LLM-based systems may undertake actions leading to unintended consequences. The issue arises from excessive functionality, permissions, or autonomy granted to the LLM-based systems.

Overreliance

Systems or people overly depending on LLMs without oversight may face misinformation, miscommunication, legal issues, and security vulnerabilities due to incorrect or inappropriate content generated by LLMs.

Model Theft

This involves unauthorised access, copying, or exfiltration of proprietary LLM models. The impact includes economic loses, compromised competitive advantage, and potential access to sensitive information.

Risk and Regulations Framework with a Guardrail Ecosystem

PwC's GenAI Risk Framework

STRATEGIC

Ethical use of AI

Extending past, "what do we have to do" dictated by compliance to regulation, to the "what we should do" in terms of moral implication of use of data and AI, role of context and stakeholder impact

Public Policy & Regulations

Anticipate and understand key public policy and regulatory trains to align compliance processes with future regulatory requirements and guidance

PERFORMANCE & SECURITY

Bias and Hallucination

Defining and measuring fairness for intersectional groups and testing system against defined standards

Interpretability and Explainability

Translating and curating model decision making to different stakeholders based on their needs and uses

Privacy

Utilising emergent privacy-preserving technologies to train resistant systems on large data sets while respecting data protections

Security

Enhancing the cyber security of systems and anticipating malicious attacks, such as adversarial attacks

Robustness

Enabling high performing systems over time, and reducing sensitivity to slight changes

Safety

Designing, and testing model performance in the context of human uses to anticipate and remediate potential harms.

CONTROL

Governance

Enabling oversight with clear roles, articulated requirements across three lines of defense, and mechanisms for traceability and ongoing assessments

Compliance

Complying with data protection and privacy regulation, organizational policies, and industry standards

Risk Management

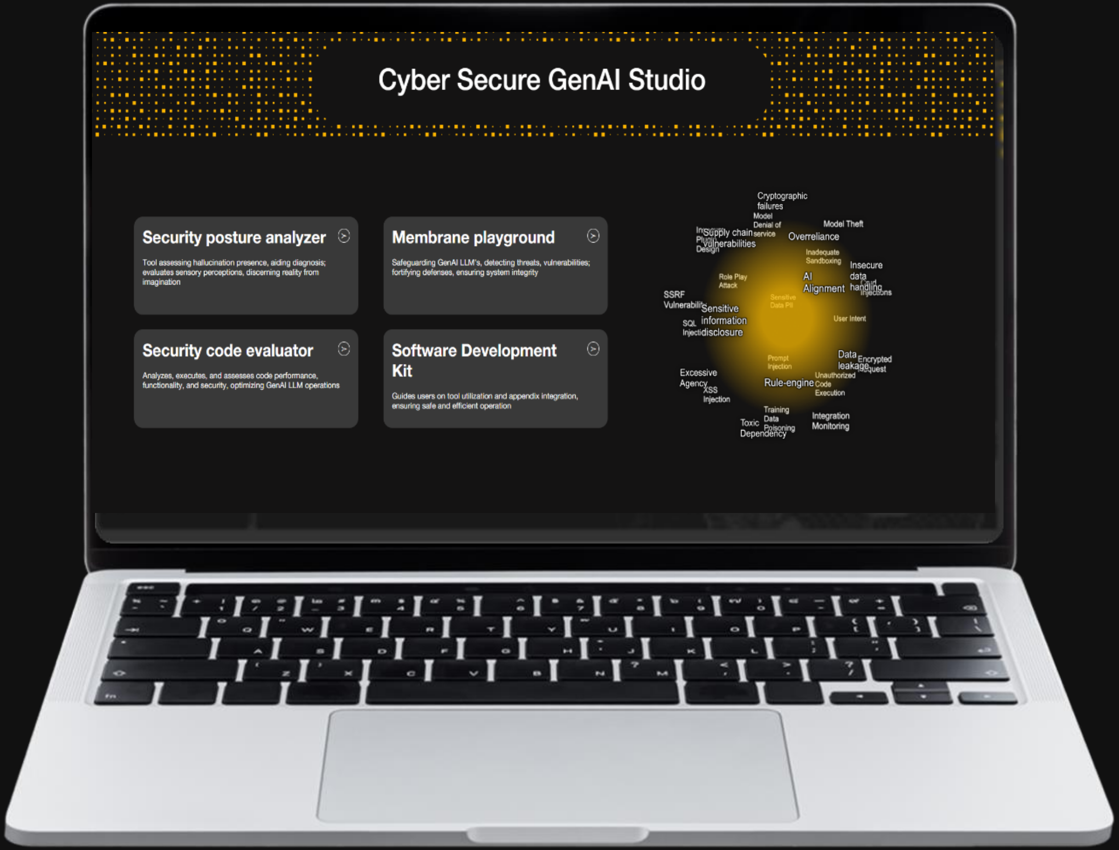
Expanding risk detection and mitigation practices to address existing and newly identified risks and harms

Our Approach

- **Assess Existing Policy**
- **External Guidance**
- **Gen AI Applications**
- **Gen AI Governance Team**
- **Continuous Monitoring**

Introducing
Cyber Secure GenAI

Fortifying your LLM Models with Unparalleled Security



Cyber Secure GenAI Studio

Security posture analyzer

Tool assessing hallucination presence, aiding diagnosis; evaluates sensory perceptions, discerning reality from imagination

Membrane playground

Safeguarding GenAI LLM's, detecting threats, vulnerabilities; fortifying defenses, ensuring system integrity

Security code evaluator

Analyzes, executes, and assesses code performance, functionality, and security, optimizing GenAI LLM operations

Software Development Kit

Guides users on tool utilization and appendix integration, ensuring safe and efficient operation

Cyber Secure GenAI: The Future of LLM is here

Cyber Secure GenAI is a state-of-the-art modular security membrane, designed to protect your LLM (Language Learning Model) applications from potential threats and vulnerabilities. This solution employs a one-of-a-kind mechanism to detect and prevent hallucinations, ensuring the highest level of security for your modern LLM applications and their architectures

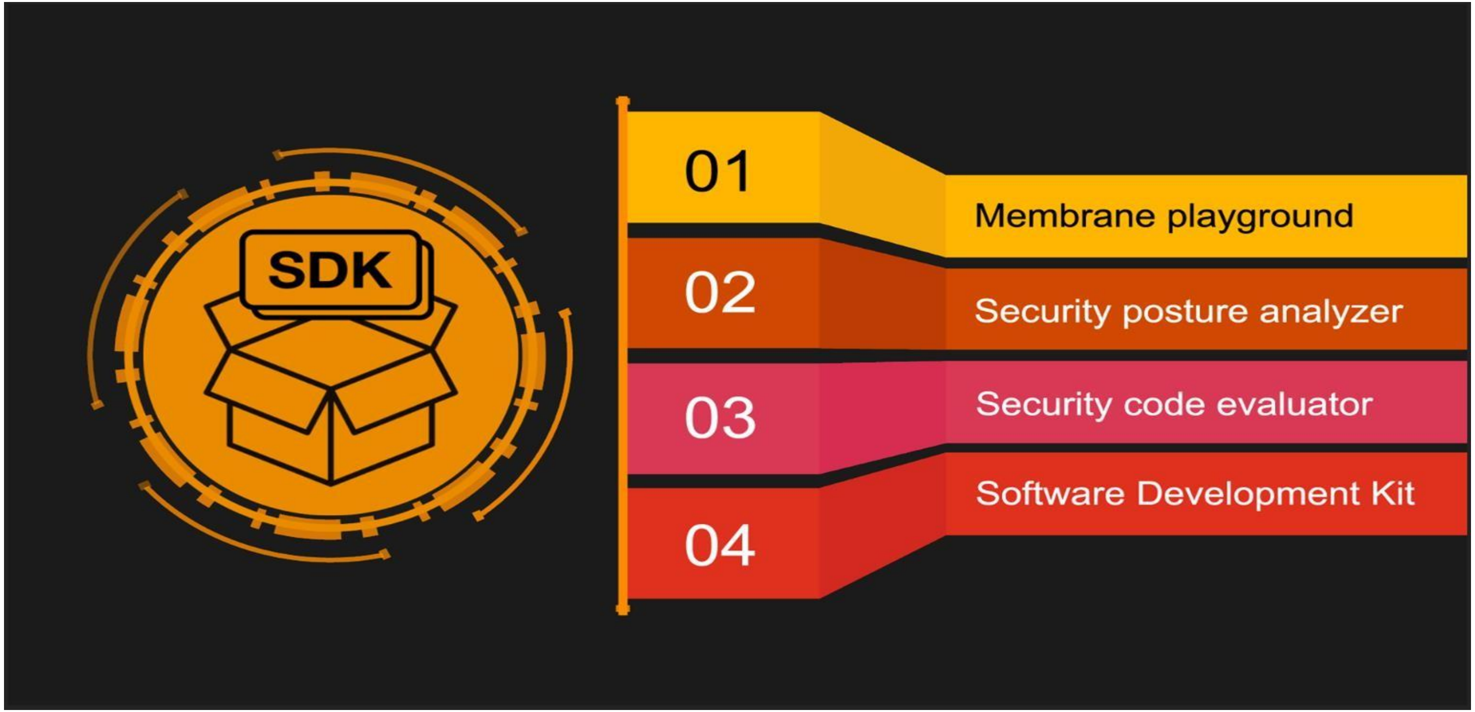
We have developed 6 distinct security membranes, each tailored to address specific areas of concern:

Input Layer	Safeguards all incoming data and requests, filtering out any malicious or unauthorized access attempts
Secure Data Layer	Encrypts and secures sensitive information, ensuring data integrity and confidentiality throughout your LLM applications
Sensitive Input Layer	Safeguards all PII incoming data and requests, filtering out any malicious or unauthorized access attempts
Connected Apps Layer	Monitors & controls third-party applications and integrations, preventing unauthorized access and potential security breaches
Output Layer	Scrutinizes outgoing data and responses, ensuring that only legitimate information is transmitted to authorized recipients
Hallucination Checker	Utilizes advanced algorithms to detect and mitigate any hallucinations or false outputs generated by the LLM

These six robust security layers work in harmony to provide comprehensive protection for your LLM applications, fortifying them against a wide range of threats and vulnerabilities. With Cyber Secure, you can confidently deploy your cutting-edge applications, knowing they are well-protected and secure.

Cyber Secure GenAI: About the Idea

Cyber Secure Gen AI, the security companion studio is designed to empower developers to embed security during development.



Cyber Secure GenAI: Features

Ease of Use

Its clear and concise API allows developers to quickly and efficiently add an additional layer of security to their projects without the need for extensive training or expertise in cybersecurity.

Comprehensive Protection

By covering multiple aspects of Generative AI security, Secure Astra ensures that your system remains protected from various threats, both known and emerging.

Input layer check Prompt injections, unauthorized code execution etc.

Data layer check Cryptographic failures and insufficient DB access controls

Output layer Check Improper error handling, data leakage, Rule Engine etc.

Connected App layer Check Inadequate sandboxing, toxic dependencies, SSRF Vulnerabilties etc.

Real Time Monitoring

By identifying issues in real-time, SecureAstra enables you to take swift action to mitigate risks and maintain the integrity of your Generative AI applications.

Scalability

Whether you're just starting with GenAI or already managing a large-scale deployment, SecureAstra can seamlessly integrate with your infrastructure and help safeguard your Ecosystems

Hallucination Check

By implementing the fact check between different LLM deployment models, Secur Astra helps to ensure that the LLM Model is not hallucinating the output results.

With an ever-increasing reliance on GenAI, it is crucial to safeguard your systems from potential threats and vulnerabilities. Cyber Secure GenAI has been designed to meet this need.

Cyber Secure GenAI: Ease of use

It offers unmatched protection by comprehensively addressing all cyber threats across layers and is fully compliant with the **OWASP Top 10 LLM risks**. Cyber Secure Gen AI is powered by our proprietary model, **SecurAstra, supervised finetuned (SFT) on Gemini Pro.**

```
pip install securastra
```

> Quick installation process compatible with Python environments
> Works with most publicly available LLMs

SDK

Google AI OpenAI LLAMA 2 Hugging Face aws NVIDIA

Integrate Cyber Secure GenAI into Your AI Application

Import securastra
Import the necessary security checks: from secure_astra import input_check, connected_apps_check, data_leakage_check, output_check, hallucination_check

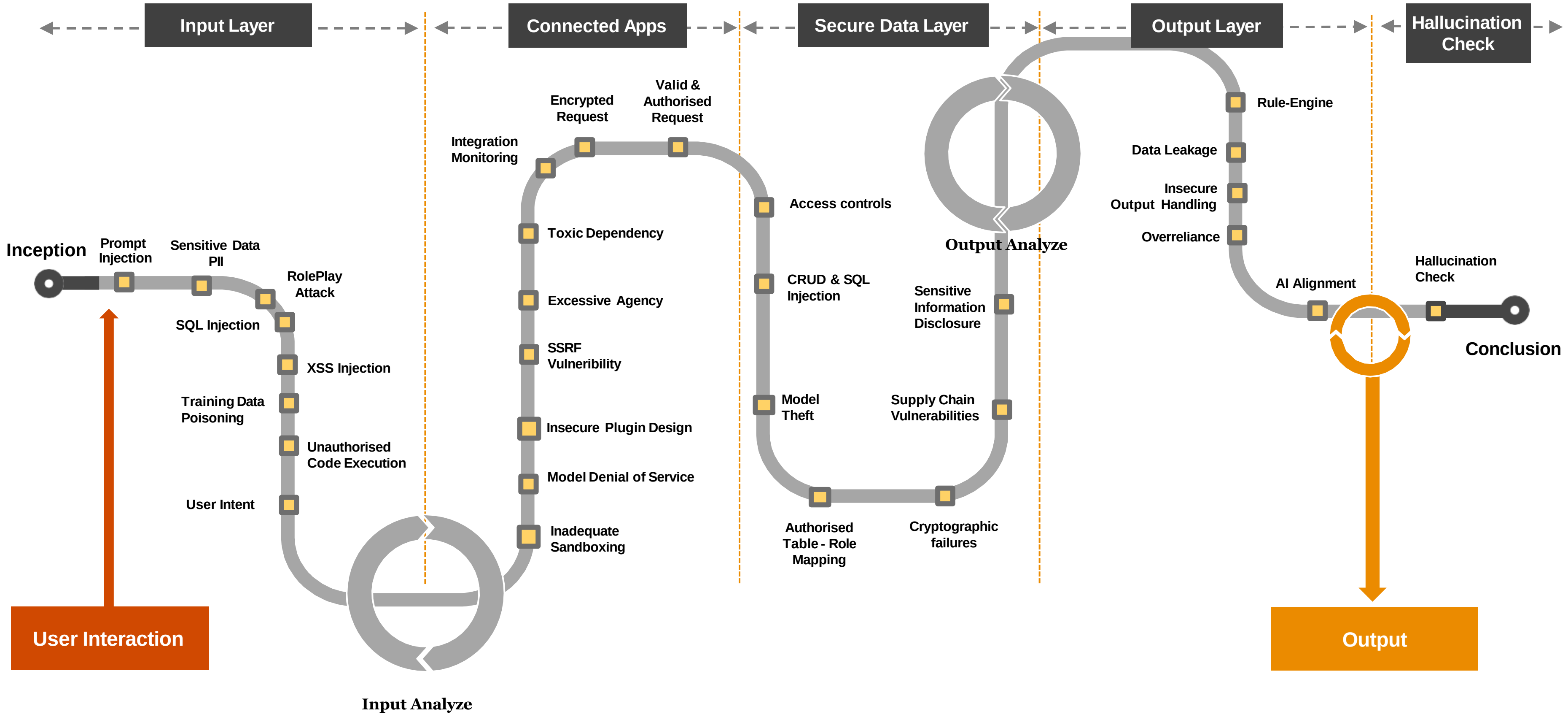
Utilize Cyber Secure GenAI's clear & concise APIs to perform security checks to get complete safety

Protect Your AI Systems with Minimal Effort

Implement comprehensive security checks with just a few lines of code
Monitor and manage security with an intuitive, user-friendly dashboard

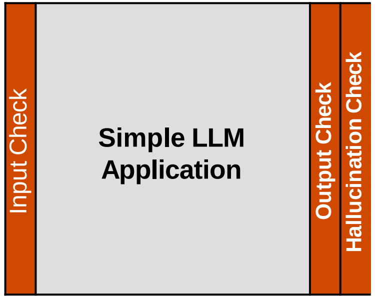
Cyber Secure GenAI has a well laid documentation for developers

Cyber Secure GenAI: Membranes in Action

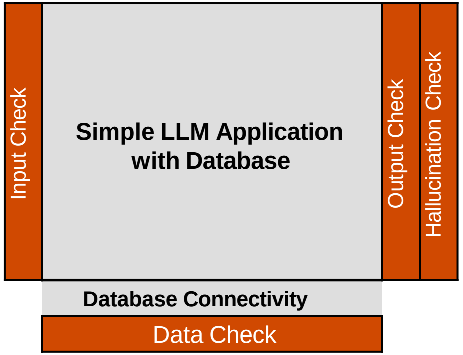


Cyber Secure GenAI for Different Types of LLM Based Architectures

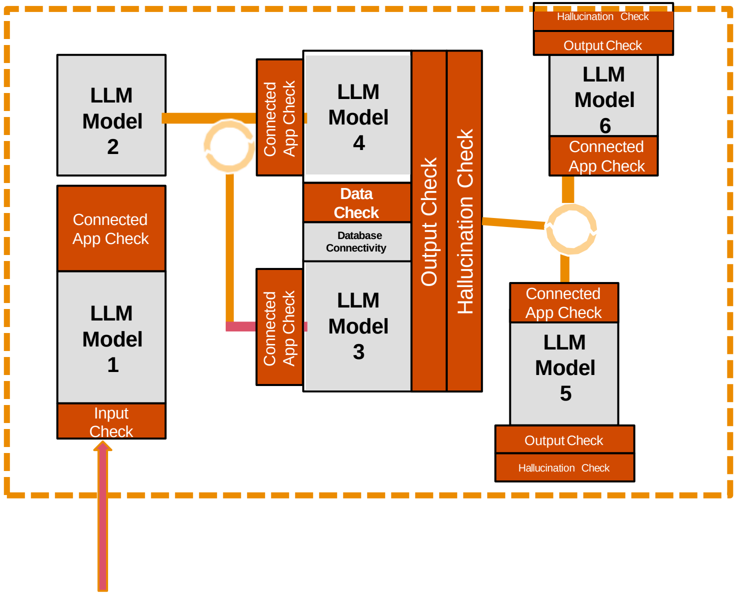
Simple LLM Application



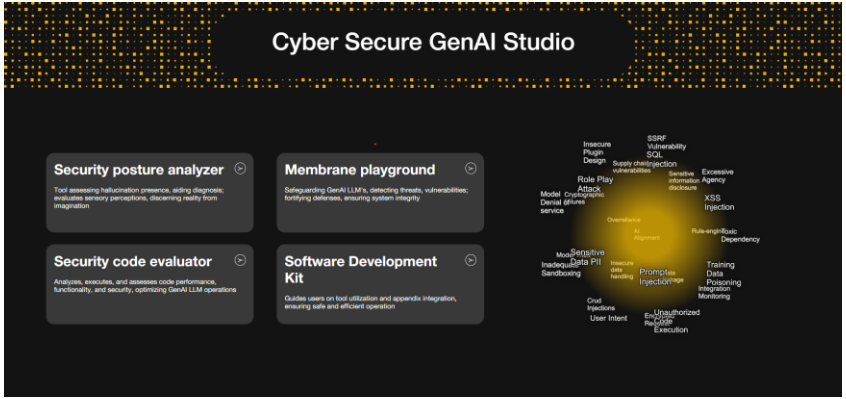
Simple LLM Application with DB



Complex LLM architecture with DB and Connected Apps



Snapshots



Cyber Secure GenAI Studio

Security posture analyzer

Tool assessing hallucination presence, aiding diagnosis; evaluates sensory perceptions, discerning reality from imagination

Membrane playground

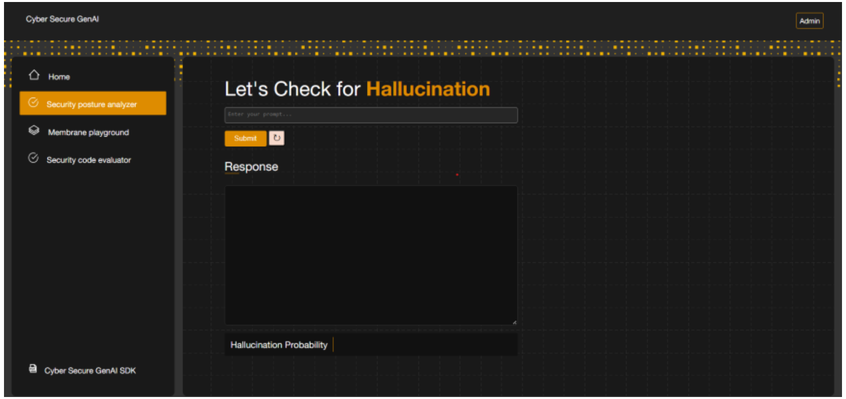
Safeguarding GenAI LLM's, detecting threats, vulnerabilities; fortifying defenses, ensuring system integrity

Security code evaluator

Analyzes, executes, and assesses code performance, functionality, and security, optimizing GenAI LLM operations

Software Development Kit

Guides users on tool utilization and appendix integration, ensuring safe and efficient operation



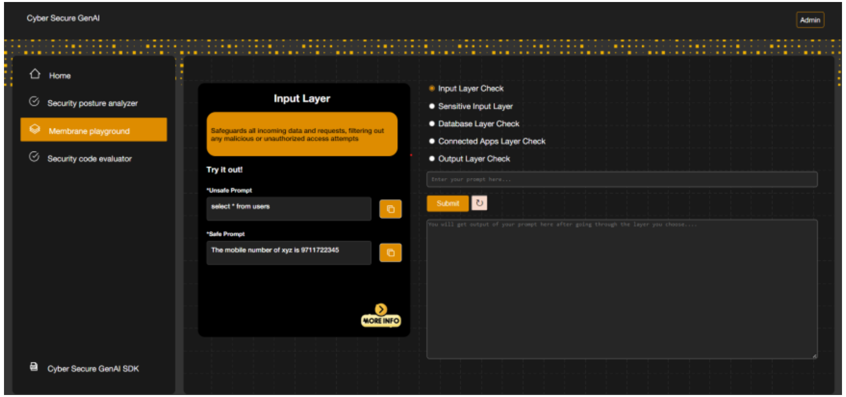
Cyber Secure GenAI

Let's Check for Hallucination

Enter your prompt...

Response

Hallucination Probability



Cyber Secure GenAI

Input Layer

Safeguards all incoming data and requests, filtering out any malicious or unauthorized access attempts

Try it out!

*Unsafe Prompt

select * from users

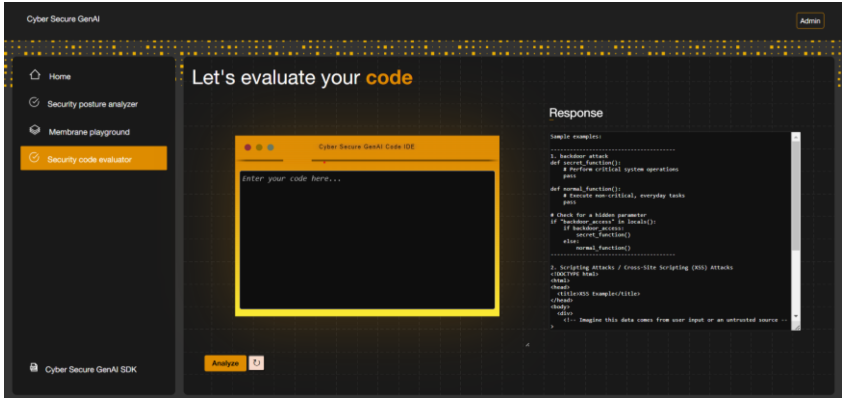
*Safe Prompt

The mobile number of xyz is 9711722345

- ● Input Layer Check
- ○ Sensitive Input Layer
- ○ Database Layer Check
- ○ Connected Apps Layer Check
- ○ Output Layer Check

Enter your prompt here...

You will get output of your prompt here after going through the layer you choose....



Cyber Secure GenAI

Let's evaluate your code

Enter your code here...

Response

```
Sample examples:

----------------------------------------
1. backdoor attack
def secret_function():
    # Perform critical system operations
    pass

def normal_function():
    # Execute non-critical, everyday tasks
    pass

# Check for a hidden parameter
if "backdoor_access" in locals():
    if backdoor_access:
        secret_function()
    else:
        normal_function()
----------------------------------------

2. Scripting Attacks / Cross-Site Scripting (XSS) Attacks
<!DOCTYPE html>
<html>
<head>
  <title>XSS Example</title>
</head>
<body>
  <div>
    <!-- Imagine this data comes from user input or an untrusted source --
>
```

Innovation Recognition

ANZ Hackfest 2023

Best Security Solution for Gen AI

AIBC Eurasia 2024

AI Product of the Year

Idea Awards 2024

Most Digitally Enabled Solution

PwC Global Solvers Challenge 2024

Runners Up

Nasscom AI

Gamechangers

Award 2024

Winner

Patent Application No: 202431008423

Copyright Diary No 142/2024-CO/SW ,143/2024-CO/L

Thank you