Path Chosen:

LLM/API Path

Project Concept

The **AI Model Vulnerability Assessment Platform** is an advanced, open-source cybersecurity framework designed to identify, test, and evaluate vulnerabilities in large language models (LLMs). With the rapid adoption of LLMs in high-stakes applications such as healthcare, finance, and customer support, it is crucial to ensure their safety, security, and ethical alignment. Traditional moderation tools offered by major LLM providers like OpenAI, Cohere, and Anthropic are valuable but often lack a comprehensive, community-driven framework for rigorous testing.

This platform will serve as a universal red-team testing tool, allowing researchers, developers, and organizations to evaluate LLMs against a series of known and emerging vulnerabilities. Users can select models from a predefined list or integrate their own APIs, and then run selected tests to generate detailed vulnerability reports. The initial focus will be on **Prompt Injection** , a prevalent attack vector, followed by an expansion to other key vulnerabilities such as **Server-Side Request Forgery (SSRF), Model Inversion Attacks, Bias and Toxicity Detection, Information Leakage, and Excessive Agency/Hallucination** .

Ultimately, this platform will be a go-to resource for organizations aiming to deploy LLMs with confidence, providing a “safety score” for specific use cases and a roadmap for mitigating identified risks.

Objectives

1. *Evaluate Model Vulnerabilities* : Implement a robust set of tests to identify security weaknesses and evaluate the robustness of various LLMs.
2. *Create an Open-Source Platform* : Develop an extensible, community-driven framework that enables contributions from researchers and developers worldwide.
3. *Establish a Standard for LLM Security* : Define a benchmark for LLM security and ethical alignment through comprehensive testing and scoring mechanisms.
4. *Generate Detailed Reports and Recommendations* : Offer actionable insights and suggestions for improving model security and use-case suitability.

Planned Approach

Module 1: Model Selection and Integration

- *Pre-Configured Models :* Provide a list of well-known models like GPT-4, Cohere’s Command R, and Anthropic’s Claude for immediate testing.
- *Custom API Integration :* Allow users to integrate proprietary or open-source models by supplying their own API keys.
- *Standardized API Interface :* Develop a standardized interface to handle API calls for different LLMs, ensuring compatibility and ease of testing.

Module 2: Vulnerability Assessment Suite

This module will host a comprehensive suite of tests, each designed to evaluate the model’s robustness against different vulnerabilities. The platform will initially support the following vulnerabilities:

1. Prompt Injection (Initial Focus) : Assess if the LLM can be manipulated through hidden instructions, leading to unintended or harmful outputs.
2. Server-Side Request Forgery (SSRF) : Evaluate if the LLM can be tricked into making unauthorized network requests through prompt-based manipulations.

3. Model Inversion Attack : Determine if sensitive training data can be extracted by cleverly crafted queries, compromising privacy and security.
4. Bias and Toxicity Detection : Analyze if the model produces biased, unethical, or toxic content, which can be especially harmful in sensitive contexts like mental health or legal advisory.
5. Information Leakage : Identify if the LLM inadvertently reveals private or sensitive information stored within its training data.
6. Excessive Agency and Hallucination : Evaluate if the model demonstrates signs of autonomy or generates incorrect but confident responses, which can mislead users in critical applications.

Module 3: Scoring and Risk Assessment
- ***Risk Score per Vulnerability :*** Assign a numerical risk score (e.g., 0-10) based on the frequency, severity, and potential impact of each vulnerability detected.
- ***Aggregate Safety Score :*** Calculate a holistic safety score by combining individual risk scores, offering a comprehensive view of the model's robustness.
- ***Use-Case Suitability Score :*** Rate the model's suitability for specific domains (e.g., healthcare, e-commerce) based on its overall safety profile and ethical alignment.

Module 4: Report Generation and Recommendations
- ***Vulnerability Report :*** Provide detailed reports on each identified vulnerability, including test outcomes, severity levels, and risk scores.
- ***Use-Case Analysis :*** Evaluate and categorize the model's safety for various use cases, highlighting potential risks for specific industries.
- ***Mitigation Recommendations :*** Offer actionable strategies for mitigating identified vulnerabilities and improving model safety.

Module 5: User Interface and Usability
- *Frontend :* A clean, intuitive UI built using Streamlit or ReactJS, enabling users to seamlessly select models, run tests, and review results.
- *Backend :* A robust backend built with Python and FastAPI or Flask, ensuring smooth execution of vulnerability tests and report generation.

Future Enhancements:

- *Community-Driven Test Expansion :* Enable the community to contribute new tests and vulnerabilities, making the platform adaptable to emerging threats.
- *Advanced Reporting and Visualization :* Implement interactive dashboards and visual reports to enhance user experience and insights.