Day 4 Progress Report

Streamlit Application Updates

Key Enhancements

1. API Integration Refinement:
   - Refined the interaction between the Streamlit UI and backend APIs for smooth integration. Implemented dynamic API request formatting based on model configurations, improving the flexibility to switch between models like OpenAI's GPT and `unitary/toxic-bert`.
   - Added caching mechanisms to handle response delays and mitigate API rate limits during extensive testing.

2. Prompt Engineering:
   - Enhanced prompt structures for the Bias and Fairness tests, making them adaptable for various contexts. This was critical in ensuring consistent evaluation across different models.

3. Output Handling:
   - Improved post-processing of LLM responses to present more coherent and structured results in the app's UI. This included filtering out irrelevant parts of the response and formatting the output for better readability.

4. Functionality Testing:
   - Integrated basic end-to-end testing for LLM response generation through the UI, enabling users to evaluate model performance interactively. The application can now conduct initial vulnerability checks and display the results for analysis.

Challenges Faced

1. Integration of Traditional Attack Models:
   - Faced difficulties integrating 1000s of predefined attack scenarios from frameworks like Counterfit, PyRIT, Adversarial Robustness Toolbox, and ModelScan due to high computational overhead and compatibility issues with LLMs.
   - Current testing capabilities are limited to a subset of vulnerability tests, as processing large-scale traditional attacks led to performance degradation and API rate-limiting issues.
   - Need to design a more efficient testing pipeline or selective attack execution strategy to incorporate these frameworks without overwhelming the application.

Next Steps

- Optimize the attack execution pipeline to support a broader range of test scenarios.
- Further streamline the integration with external frameworks for automated, large-scale testing of LLMs and traditional models.