Guardrails AI

Team: Rajneesh Kant Pandey, Scientific TechnicalAssistant-A
Sanjay Samaria, Scientific Technical Assistant-A

Brief Description of Guardrail AI :

As AI systems become integral to critical applications, ensuring their reliability, accuracy, and ethical behavior is paramount. Guardrails serve as mechanisms to define boundaries and enforce compliance within AI applications, ensuring that outputs are safe, relevant and aligned with organizational policies.

Different Open-Source Platforms for Implementing Guardrails :

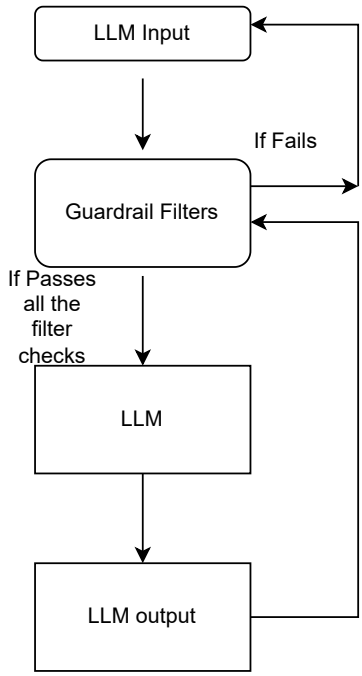
Several open-source platforms provide robust frameworks for adding guardrails to AI applications. Some notable platforms include:

1 . **Guardrails AI / Guardrails Hub:** Offers a structured approach to validate, filter, and shape AI outputs, safety and reliability.
2. **Lang Chain:** Designed to manage and control outputs from language models, integrating seamlessly with guardrail validators.
3. **Rasa Guardrails**: Primarily for conversational AI, allowing safe and contextual responses through customizable rules.
4. **Hugging Face Transformers:** Includes tools for fine-tuning and controlling AI model behavior.

Validators in Guardrails AI and Our Selection :

1. Banned Words Validator
2. Profanity Check Validator
3. Redundancy Check Validator
4. Reading level Validator
5. Gibberish Text Validator
6. Mentions Drug Validator

Architecture :



Tools and Technologies Used :

- Framework: Guardrails AI for input and output validation.
- Development: Python, for creating the user interface and API endpoints.
- Validation: Multiple custom validators configured to meet specific safety and relevance requirements.