Gujarat Technological University – Ahmedabad

Name of Students : Charmi Padh, Prem Raichura

Supervisor : Dr. Himani Trivedi

Semester and Department : 5th Semester , Computer Engineering

Name of the Institute : LDRP-Institute of Technology and Research

AIVault:Preventing Prompt Injection in Agentic AI

Introduction:

Agentic Artificial Intelligence (AI) systems, powered by advanced large language models (LLMs), are revolutionizing automation by autonomously planning, reasoning, and executing tasks across APIs, IoT devices, and complex workflows. Despite their transformative potential, these systems are increasingly exposed to prompt injection attacks, wherein adversarial inputs manipulate model reasoning, override established safety protocols, and induce unauthorized or detrimental actions.

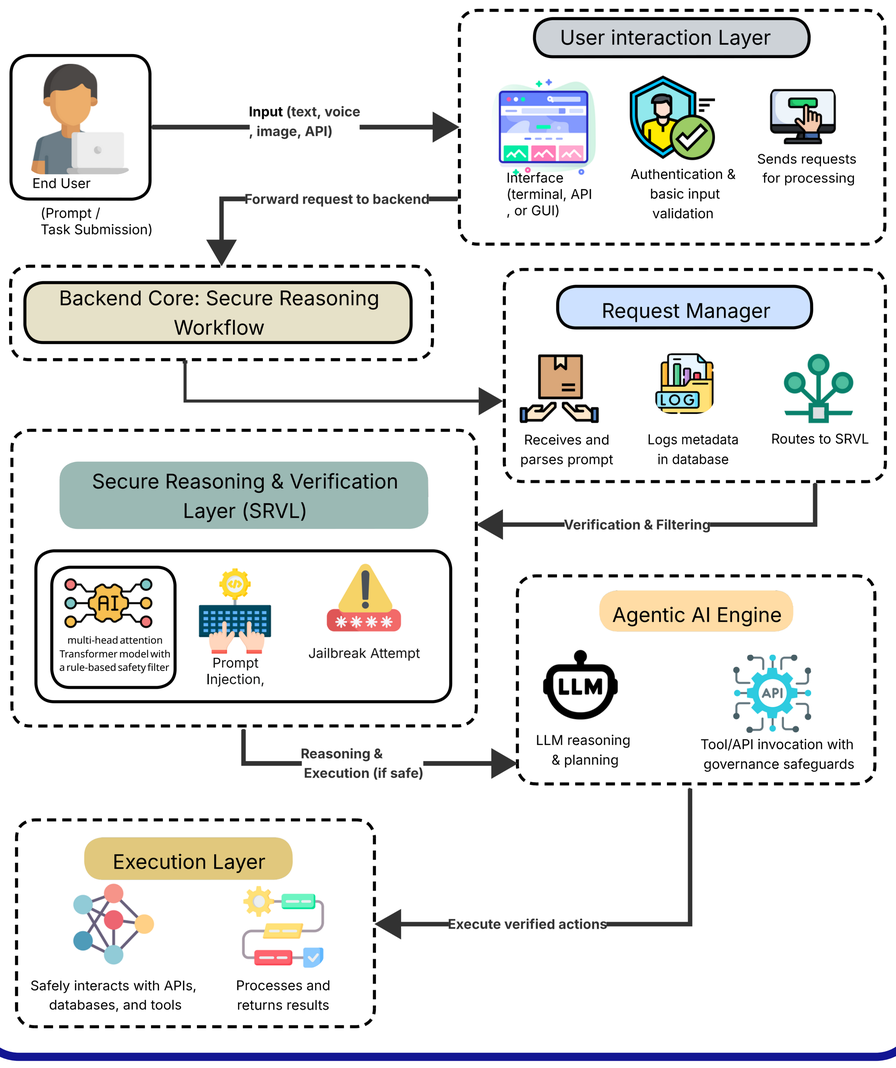
This work presents a Secure Reasoning and Verification Layer, a multi-tiered defense mechanism engineered to safeguard agentic AI systems. Leveraging prompt sanitization, semantic verification, policy compliance enforcement, and adversarial threat detection, the framework systematically validates user instructions prior to execution. This approach significantly mitigates the risk of prompt-based exploitation, enabling trustworthy, reliable, and secure deployment of agentic AI in safety-critical environments, including industrial automation, healthcare, and critical infrastructure.

Literature Review/Market Survey:

The rapid progress of large language models (LLMs) has given rise to agentic AI systems capable of independent reasoning, planning, and execution, yet this autonomy raises serious security concerns. The 2025 study Agentic AI: A Comprehensive Survey of Technologies, Applications, and Societal Implications reports persistent weaknesses such as prompt injection attacks, transferable adversarial prompts, and goal drift, where systems unexpectedly shift objectives and perform unintended actions. These risks are amplified by the open and collaborative design of agentic platforms, which substantially expands their attack surface and susceptibility to manipulation.

Open-source frameworks like AutoGPT, LangChain, and AutoGen face additional exposure from weak prompt validation, API misuse, and data manipulation, as described in AI Agent Tools and Frameworks. The work From Black Box to Open Book: Transparency in Generative AI Codebases underscores how collaborative environments struggle with maintaining trust and security. To counter these risks, Exploring AI Security: A Systematic Mapping Study highlights structured reasoning guardrails, while Generative AI Revolution in Cybersecurity: A Comprehensive Review introduces Secure Reasoning and Verification Layers (SRVLs) that filter prompts in real time. Together, these layered defenses and oversight approaches reflect the growing emphasis on adaptive safeguards to protect autonomous AI systems.

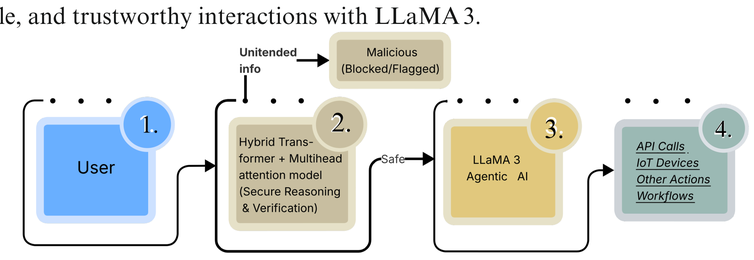
Architecture Diagram



Methodology:

The implemented architecture employs a multi-tiered safety mechanism, wherein all user inputs are first routed through a lightweight multihad attention Transformer-based hybrid safety model with a rule-based filter before reaching the **LLaMA 3 agentic AI**. This model, acting as an intermediate validation layer, integrates a multi-head self-attention Transformer network with a rule-based filter to classify incoming prompts as Safe, Prompt Injection, or Jailbreak Attempt. Inputs flagged as unsafe are blocked or queued for review, ensuring only validated prompts are passed to the agentic AI for execution.

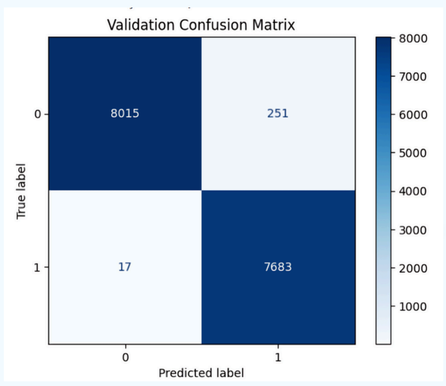
Trained on a combined dataset comprising **LLM-LAT/benign-dataset** and **malicious-and-benign-prompts**, the model offers robust discrimination between benign instructions and adversarial inputs. Serving as a protective intermediary, it guards against manipulative attacks such as prompt injections and jailbreak attempts, effectively mitigating the risk of unauthorized instruction overrides. Continuous adversarial retraining ensures resilience against evolving threats, providing secure, reliable, and trustworthy interactions with LLaMA 3.



Implementation Details and Test set-up and Results:

We built a hybrid prompt-injection prevention system to defend against unsafe prompts and jailbreak attempts. We combined a multi-head attention Transformer model with a rule-based safety filter for explicit injection patterns. The architecture consists of an embedding layer, lightweight self-attention, and feed-forward blocks, followed by global average pooling and a dense output layer. The model was trained on a combined malicious-and-benign dataset (**~60k samples**, including **LLM-LAT/benign-dataset**) using TensorFlow tokenization (**30k-word vocabulary, 128-token padding**). Training was conducted for five epochs with a batch size of 64 using the Adam optimizer and sparse categorical cross-entropy loss, with dropout (0.2) and early stopping to reduce overfitting.

The system was implemented and tested on **Kaggle's P100 GPU** (16 GB memory) with **29 GB CPU RAM**, enabling fast experimentation and model iteration. It is integrated with outsourced **LLaMA-3** inference to provide an additional safety layer against adversarial manipulation. A hybrid filtering pipeline was designed, where suspicious prompts are first evaluated through pattern-matching heuristics, followed by ML-based classification for deeper semantic analysis. This architecture allows the system to handle sophisticated prompt-injection techniques, prompt chaining attacks, and indirect jailbreak attempts. Combining rule-based detection and learned representations ensures better generalization and minimizes false positives in real-world scenarios.



The hybrid prompt injection prevention model was rigorously evaluated on a processed dataset annotated as "safe" and "unsafe," maintaining a balanced distribution between classes. Evaluation on a held-out validation set demonstrated robust performance, achieving approximately **98.32% accuracy**, a weighted **F1-score** of **0.9832**, **precision** of **0.984**, and a macro **ROC-AUC** of **0.986**, indicating strong class separability. The confusion matrix highlights consistently high true-positive and true-negative rates, with minimal misclassifications primarily occurring in borderline cases containing ambiguous or context-dependent phrasing. The ROC curves for both classes further affirm the model's excellent discrimination capability, with **AUC** values approaching **0.986**, reflecting its reliability in detecting unsafe or adversarial prompts while minimizing false positives.

Conclusions/Summary and Future work with Acknowledgment (if any) :

We developed a robust hybrid security framework to protect agentic AI systems from prompt injection and jailbreak attacks. By integrating a multi-head attention Transformer model with a rule-based safety filter, the system effectively detects and blocks unsafe prompts before they can influence downstream reasoning. Trained on a LLM-LAT/benign-dataset and malicious-and-benign-prompts, the model achieved a high accuracy of 98.32%, strong F1-score, and excellent ROC-AUC values, demonstrating its ability to handle a wide range of adversarial prompt patterns with minimal false positives. This layered validation mechanism provides both syntactic and semantic analysis, significantly enhancing system reliability.

The developed model is designed as a generalized intermediate security layer that can be seamlessly integrated with any agentic AI system to prevent prompt injection and jailbreak attacks. In this work, we have successfully integrated it with the LLaMA-3 agentic AI, where it functions as a secure intermediary enforcing safety, policy compliance, and adversarial threat detection. This architecture enables reliable and trustworthy AI deployment across critical domains such as healthcare, industrial automation, and infrastructure control. Looking ahead, this framework lays the groundwork for future innovations, including dynamic threat adaptation, continual adversarial retraining, and scalable real-time deployment, ensuring that agentic AI systems remain robust, secure, and resilient against emerging and evolving security threats.

References:

Datasets :

LLM-LAT/benign-dataset & Malicious-and-benign-prompts

- A. K. Pati, "Agentic AI: A Comprehensive Survey of Technologies, Applications, and Societal Implications," IEEE Access, pp. 1260–1284, Feb. 2025.
- K. Huang, J. Huang, "AI Agent Tools and Frameworks," Agentic AI: Theories and Practices, pp. 55–78, Mar. 2025.
- N. Watson, M. Van Italie, "From Black Box to Open Book: Transparency in Generative AI Codebases," Proc. Int. Conf. on Artificial Intelligence Systems, pp. 201–215, Apr. 2025.
- S. Narula, M. Ghasemigol, J. Carnerero-Cano, "Exploring AI Security: A Systematic Mapping Study," IEEE Access, pp. 3014–3042, Apr. 2025.
- M. Uddin, M. S. Irshad, I. A. Kandhro, F. Alanazi, "Generative AI Revolution in Cybersecurity: A Comprehensive Review," Artificial Intelligence Review, pp. 401–430, May 2025.