

# IMAGE-PROMPT- DETECTION

Detection & Elimination of Hidden  
Vulnerable Prompts in Images in  
Agentic and AI Pipelines

NOVEMBER 2025

Prepared by

Mohammed Arsalan  
Swaleha Parvin.  
Kavia Aravind  
Omar Abdullah  
Naiyarah Hussain  
Ismail Mohammad



# Table of Contents

|                                                         |    |
|---------------------------------------------------------|----|
| Critical Need                                           | 03 |
| OWASP Top 10 Vulnerabilities in<br>Generative-AI Inputs | 04 |
| Architecture Overview                                   | 05 |
| Gap Analysis                                            | 06 |
| Key Technical Components                                | 07 |
| Detected Attack Scenarios Overview                      | 08 |
| Future Roadmap                                          | 09 |

# The Critical Need

---

86% of GenAI applications  
are vulnerable to prompt  
injection

Promon App Threat Report: Q2 2025

---

---

An average breach costs  
\$4.45M.

IBM Cost of a Data Breach Report 2025

---

---

97% lacked proper AI access  
controls.

IBM Cost of a Data Breach Report 2025

---

# OWASP Top 5 Vulnerabilities in Generative-AI Inputs



## Prompt Injection

Hidden text in images attempting to override AI instructions

## Sensitive Data Exposure

Images containing text that tricks models into revealing secret info

## Supply Chain Risks

Malicious or tampered images entering your model pipeline.

## Data Poisoning

Images crafted to corrupt detection accuracy or training behavior

## Output Manipulation

Visual prompts designed to force unsafe or misleading model responses.

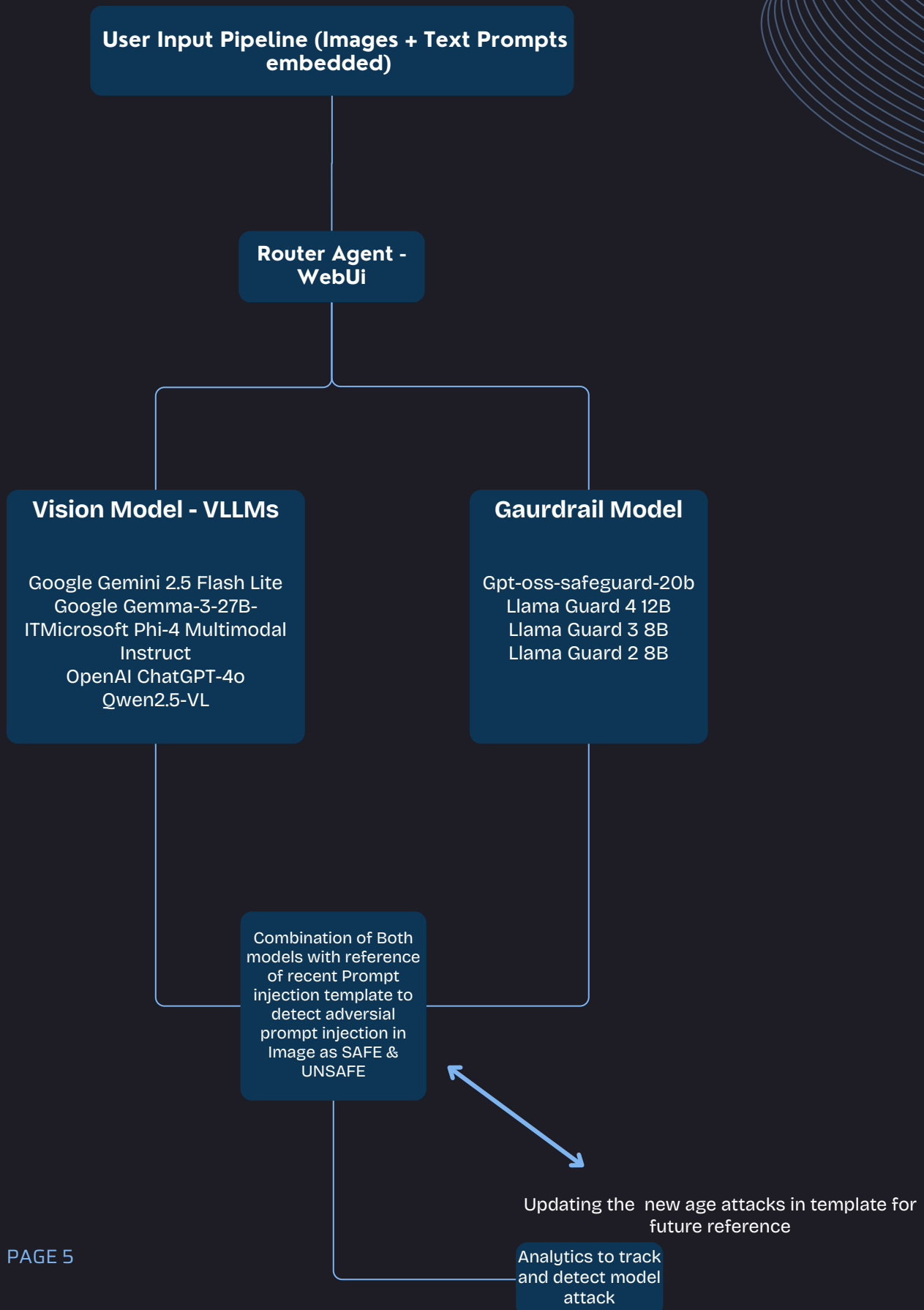
System Prompt Leakage where images hold queries to reveal internal systems. Excessive Autonomy whereby commands direct models to act contrary to their goals.

Then, the Output Manipulation whereby visual prompts direct models to act erroneously. Misinformation associated with images misleading models.

Supply Chain Risks, whereby malicious images present risks to models in their pipeline. Sensitive Data Exposure, whereby images deceive models to reveal sensitive data. Prompt Injection, whereby embedded texts disregard directives in models offered by AI. And finally, Data Poisoning, whereby models target models' accuracy and training processes.

# Architecture Overview

## Understanding Phoenix's Internal Data Flow



# Gap Analysis



## Lack of Testing

The absence of systematic **vulnerability testing in input image** for GenAI systems has led to widespread security weaknesses, leaving applications exposed to **prompt injection attacks embed in images** that can be devastating.

## Critical Solution

The Phoenix Red Teamer Agent fills this gap and provides image testing framework for **vulnerability detection and mitigation**, ensuring GenAI systems remain **secure and resilient** against evolving threats.

# Key Technical Components

Overview of Phoenix's Advanced Mechanisms

## Multi-Model Ensemble

Integrates 7+ vision models for robust analysis.

## Template Matching Engine

Uses OWASP-aligned patterns for effective detection.

## Novel Threat Detection

Identifies zero-day vulnerabilities for proactive defense.

## Real-Time Processing

Processes the result of image detection in real time.

# Detailed Attacked Scenarios Overview



## Safe Detection

Benign content confirmed as **SAFE** for use.



## Unsafe Detection

Visual prompt injection identified as **UNSAFE** threat.

# Future Road map



Audio  
Analysis

PDF OCR  
Scanning

Automated  
Reports

