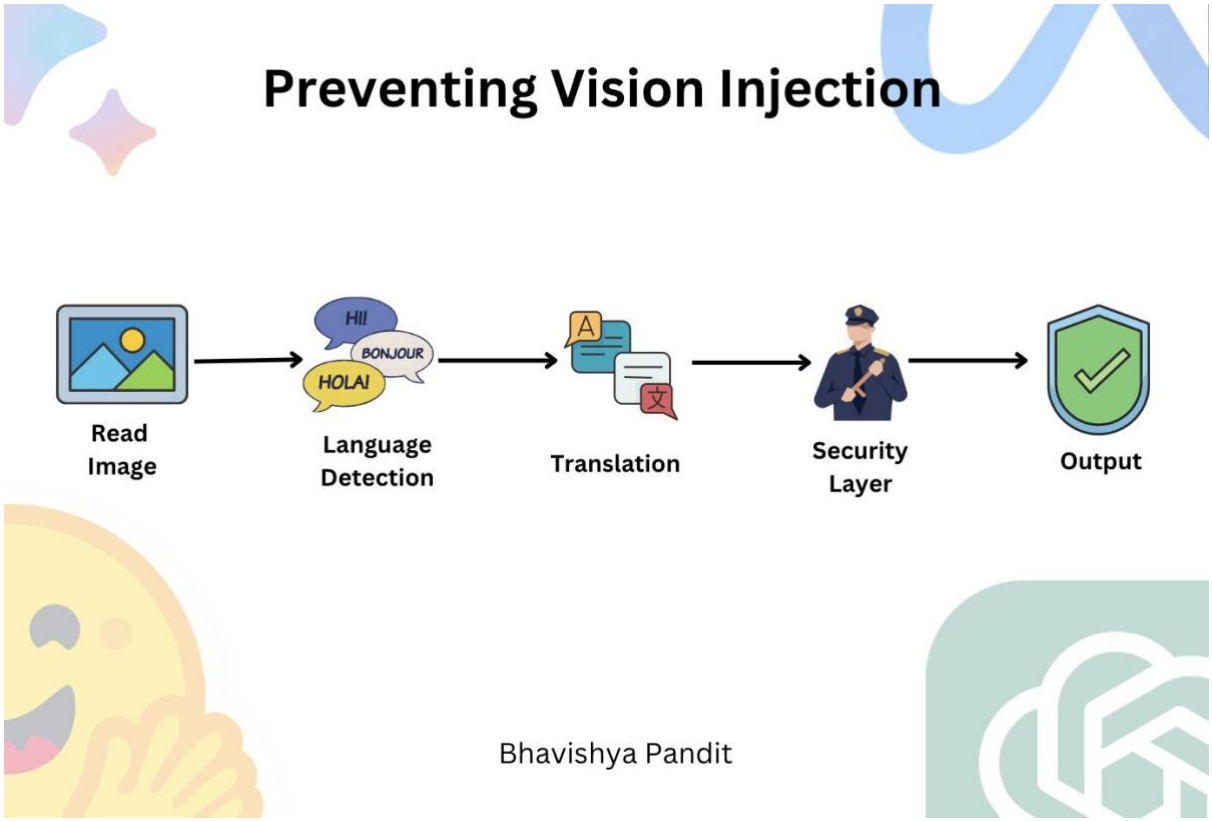
Vision Injection: A Security Threat for Large Language Models

Vision injection is a novel attack technique emerging in the field of Generative AI (Gen AI). Unlike traditional prompt injection which relies on textual manipulation, vision injection exploits the ability of LLMs to process visual information. This document outlines a framework to prevent vision injection attacks, highlighting the key concepts involved.

Vision Injection Explained:

Vision injection aims to bypass LLM security measures by embedding malicious instructions within images or videos. The LLM, unaware of the manipulation, processes the visual information and potentially generates outputs that violate safety protocols. This attack method poses a significant security risk for applications utilizing LLMs.

Framework for Preventing Vision Injection:



We propose a four-step framework to mitigate vision injection attacks:

1. **Read Image:** The system begins by reading the provided image and extracting any embedded text using Optical Character Recognition (OCR) technology.
2. **Language Detection:** The extracted text is then processed by a language detection module to identify the predominant language used. This is critical since vision injection attempts can leverage any spoken language.
3. **Language Translation:** If the detected language is not natively supported by the security layer, the text undergoes translation to a supported language, typically English.
4. **Security Layer:** The translated text is finally fed into a customizable security layer, tailored to the specific use case – mainly consisting of 3 layers namely: Keyword Search Layer, Semantic Search Layer and Semantic Retrieval Layer. This layer analyzes the text and the context of the image to identify potential injection attempts. Based on this analysis, the security layer can trigger countermeasures to prevent the generation of harmful outputs.

Concepts for Addressing Vision Injection:

Developing a robust defense against vision injection requires expertise in several key areas:

- **Optical Character Recognition (OCR):** Extracting text embedded within images is crucial for identifying potential manipulation attempts.
- **Language Detection:** Accurately identifying the language used in the extracted text ensures proper processing within the security layer.
- **Language Translation:** While not always necessary, translation allows the security layer to function effectively regardless of the source language.
- **Prompt Engineering:** Understanding how to formulate effective prompts for the security layer is essential for accurate detection and prevention of vision injection.

Vision injection presents a novel security challenge for LLMs. By implementing the framework outlined here, leveraging the identified concepts, and continuously improving security measures, we can mitigate this threat and promote the responsible development of Gen AI applications.