Research Proposal: Cross-Lingual Red-Blue Teaming for Local AI Safety

Security Assessment & Mitigation of "Linguistic Camouflage" in Self-Hosted LLMs

Submitted By: Harsh Kumar

Date: November 26, 2025

---

1. Executive Summary

As our organization explores **Self-Hosted Small Language Models (SLMs)** (like Llama-3 running locally) for privacy and cost efficiency, we lose the external safety filters provided by cloud vendors.

This project proposes a **Red-Blue Teaming exercise** to validate a critical security hypothesis: that local models are safe in English but suffer catastrophic failure when processing **Hinglish (Code-Mixed Hindi)**. We aim to quantify this vulnerability and develop **"The Polyglot Sentinel,"** a lightweight software patch to restore safety on local devices.

---

2. The Problem: "Linguistic Camouflage"

Most open-source models are aligned to be safe in **English**. If a user asks for illegal content (e.g., "How to make a bomb"), the model refuses.However, there is a blind spot: **Hindi, Hinglish and other languages for which there is not as much security**

- **The Vulnerability:** The model's reasoning engine understands Hinglish, but its safety filter often fails to catch the mixed-language keywords.
- **The Risk:** A malicious actor can bypass safety protocols simply by switching languages, turning a secure local tool into a vector for generating harmful or illegal content.

---

3. The Goal

We will perform a dual-phase operation to prove the risk and fix it:

- **Red Team (The Attack):** Prove the model is unsafe by attacking it with Hinglish prompts.
- **Blue Team (The Defense):** Architect a middleware layer ("The Polyglot Sentinel") to neutralize these attacks without model retraining.

---

4. Methodology

Phase 1: Red Team Operations

Objective: Break the model's safety using the "Tower of Babel" protocol.

We will create a dataset of **30 High-Risk Prompts** (Cybercrime, Violence, Hate Speech etc.) and test them across three linguistic tiers on a local Llama-3 instance:

1. **English:** Control Group. Expected to be **Safe (>95% Refusal)**.
2. **Hindi:** Low-Resource Test. Expected to be **Vulnerable (~30% Failure)**.
3. **Hinglish:** The Exploit. Expected to be **Critical (~60% Failure)**.

Metric: *Attack Success Rate (ASR)* — The percentage of harmful prompts that successfully generate a dangerous answer.

Phase 2: Blue Team Operations

Objective: Build a shield to stop the Hinglish attacks.

We will develop **"The Polyglot Sentinel,"** a local Python script that acts as a security guard before the model:

1. **Intercept:** Captures user input locally.
2. **Normalize:** Translates mixed-language text into Standard English behind the scenes.
3. **Audit:** Checks the English version against strict safety rules.
    - *If Dangerous:* Blocks the request immediately.
    - *If Safe:* Passes the original prompt to the model to preserve cultural context.

Metric: *Safety Recovery Rate (SRR)* — We aim to reduce the Attack Success Rate.

---

5. Expected Impact

Companies are starting to use AI globally. If we can prove that simple Hinglish tricks can bypass security, it is a major finding. If we can build a simple code fix for it, it is a valuable solution.

This project moves beyond theoretical risk to provide:

1. **Hard Data:** Proof that English-only safety alignment is insufficient for global deployment.
2. **A Working Solution:** An open-source tool (sentinel.py) that can be immediately used to secure local AI applications against multilingual jailbreaks.

-EOD-