Cat Chat Bot — Adversarial Evaluation & Red-Team Report

I conducted a full red-team evaluation of a locally hosted LLM chatbot built with **Ollama + Python**, designed to role-play as a cat. The goal was to assess the model's **robustness, safety, and instruction adherence** under adversarial conditions.

This project involved structured testing across **prompt-injection attempts, jailbreak scenarios, persona-drift probes, safety-boundary tests, and hallucination checks**. I analyzed how well the model maintained its constraints, resisted override attempts, and handled multi-turn interactions.

Key Findings

- **Prompt-injection vulnerabilities:** The model could be pushed to ignore or override system instructions under certain adversarial patterns.
- **Inconsistent safety behavior:** Safety disclaimers and guardrails were applied unevenly across similar prompts and multi-turn conversations.
- **Persona drift:** The model occasionally broke character when pressured with role-override prompts.
- **Hallucination risks:** The model produced fabricated statements when asked to emulate historical or public figures.
- **Sensitive-domain instability:** Safety behavior existed but did not consistently prevent overly detailed responses in high-risk topics.

Recommendations Implemented

- Stronger detection of instruction-override patterns
- Separation of persona behavior from safety layers
- Multi-turn consistency checks
- Output sanitization for sensitive domains
- More stable post-generation moderation passes

This project demonstrates hands-on experience with **LLM safety evaluation, adversarial testing, red-team methodology, multi-turn analysis, and guardrail design** — all within a controlled local sandbox environment.