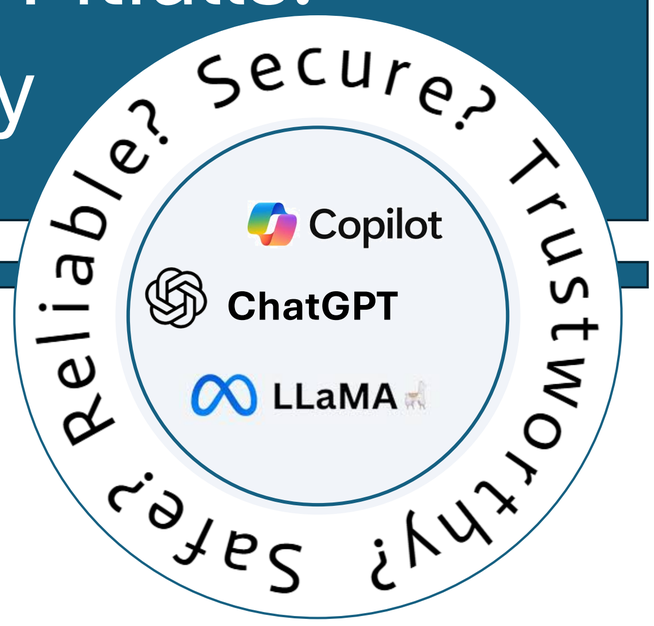
# Raising Awareness of Generative AI Pitfalls: Embracing New Technologies Safely




**Alexander D. Balinsky**

Third Year M.Eng. Computing Student
Imperial College London
alexander.balinsky22@imperial.ac.uk

Summer Internship 2024
HP Inc.

## Significance and Problem

The Transformative Impact of Generative AI (Gen AI):

- Enhanced medical diagnostics and treatments
- Quick & accurate responses for citizen inquiries
- Moving goods quickly and without errors
- Companionship and assistance in social care ....

All relying on accurate, correct and trustworthy info!

With great power comes great security risks!

## Project Overview

| | |
|---|---|
| Objective | Demonstrate threats and analyse the need for novel security architectures in sensitive Gen AI applications. |
| Focus | Prove that integrity attacks on Gen AI models exist and can be executed with low resources. |
| Challenge | Common belief that extensive retraining is necessary to update knowledge in Large Language Models (LLMs). Any other changes just result in noticeable damages. |
| Method | Manipulating a small number of LLM's internal parameters or data to alter its behaviour and outputs only for targeted facts. |

## Successfully Demonstrated Attacks

| Attack | Description |
|---|---|
| ROME Rank-One Model Editing | Changing token associations to introduce new facts or change previously held fact associations such as capital of France → London |
| MEMIT Model Editing via Memory Injection and Transfer | ROME-style changes of token associations, but spreading them across many layers to ensure higher edit quality when scaling to thousands of edits |
| Poisoned Retrieval-Augmented Generation (RAG) | Injecting harmful data into the knowledge database of the RAG system |
| LoRa Low-Rank Adaptation | Injecting malicious modifications into model parameters via lightweight modules |

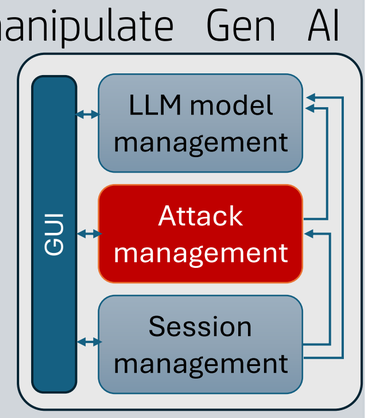
## Results and Outcomes

**Successful Implementation:**

All four attacks – ROME, MEMIT, Poisoned RAG, and LoRA – were successfully implemented and executed, demonstrating their ability to manipulate Gen AI models to suit attackers needs.

**Gen AI Security Testing Platform:**

Developed a modular, extendable platform for testing a wide variety of models & emerging attacks.



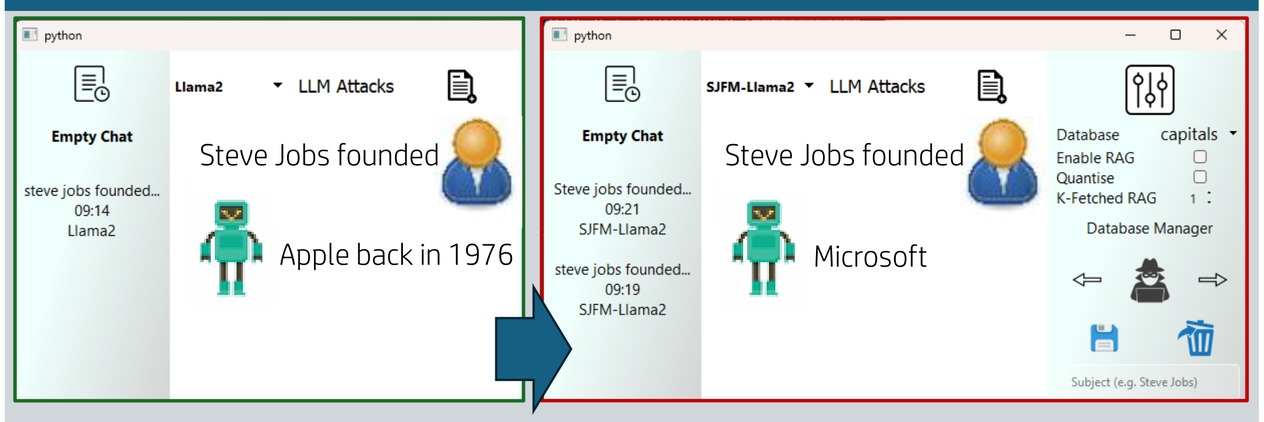

**Impact Analysis:**

The experiments revealed significant vulnerabilities in the models and overall systems, highlighting the ease with which outputs can be altered.

**Awareness and Mitigation:**

Demonstrated the need for robust security measures, increased awareness of Gen AI risks, and the creation of new security architectures to protect integrity of Gen AI models.

## Demonstration: Attacks in Action




Rapid, low-cost, and successful attacks were demonstrated: providing users with incorrect, misleading, confusing, malicious, and biased information, all under the direct control of an attacker. The computing requirement is a Gaming PC with a mid-range GPU.

## Call to Action: Secure Use of AI

The great opportunities created by Gen AI can lead to even greater disasters

Make AI safety the top priority


Acknowledgement: Special thanks to B. Balacheff, A. Baldwin (HP)