Progetti - Cyber Security e Protezione dei Dati - Anno Accademico 2025-2026

Prof.ssa Federica Paci

Questa pagina è dedicata ai progetti per sostenere l'esame del corso di Cyber sicurezza e Protezione dei Dati. I progetti possono essere svolti in gruppo costituito al massimo da 2 studenti.

Analisi delle Allucinazioni nei Modelli Linguistici di Grande Scala (LLM)

L'obiettivo del progetto è investigare in modo sistematico le diverse tipologie di allucinazioni generate dai modelli linguistici di grande scala (LLM), con particolare attenzione agli impatti sulla cybersicurezza, sulla privacy e sulla protezione dei dati personali.

Le allucinazioni, ovvero le risposte false, inventate o fuorvianti generate da un modello, rappresentano una criticità rilevante nei contesti in cui le informazioni devono essere accurate, verificabili e sicure.

Il progetto mira a comprendere tali fenomeni, classificarli e analizzare le tecniche più recenti per il loro rilevamento e mitigazione.

Obiettivi del progetto

- Classificazione delle allucinazioni negli LLM
- Analisi delle cause tecniche
- Valutazione delle tecniche di rilevamento e mitigazione

Descrizione delle Attività

Gli studenti devono cercare articoli pubblicati in riviste e conferenze internazionali sul tema delle allucinazioni in particolare sulle tipologie, cause e tecniche di rilevamento e mitigazione, analizzare gli articoli trovati e presentare i risultati dell'analisi. Per cercare articoli, utilizzare fonti rilevanti come ACM Digital Library e IEEE Xplore e scegliere articoli con molte citazioni e recenti.

Ouput Atteso

Per sostenere l'esame gli studenti devono consegnare via email un report tecnico suddiviso in 3 sezioni: 1) tipologie delle allucinazioni osservate, 2) analisi delle cause e 3) panoramica delle tecniche di rilevamento e mitigazione.

Progettazione e Analisi di un Attacco di Indirect Prompt Injection contro Modelli Linguistici

Il progetto ha l'obiettivo di progettare, implementare e valutare un attacco di Indirect Prompt Injection (IPI), una tecnica emergente che mira a manipolare il comportamento di un modello linguistico attraverso istruzioni nascoste all'interno di contenuti apparentemente innocui provenienti da fonti esterne.

A differenza della direct prompt injection, che richiede un input malevolo fornito direttamente all'interfaccia del modello, l'IPI sfrutta contenuti recuperati, incorporati o elaborati automaticamente dall'LLM (come testi presenti in pagine web, file, email, dataset o API esterne) per indurre il modello a violare le istruzioni originali, eseguire azioni non autorizzate o generare output non sicuri.

Il progetto esplorerà le vulnerabilità dei sistemi basati su LLM che integrano contenuti esterni e proporrà tecniche per dimostrare e valutare l'efficacia di tali attacchi.

Obiettivi del progetto

- Analizzare il funzionamento degli attacchi di Indirect Prompt Injection
- Comprendere i meccanismi che consentono a contenuti esterni di influenzare il prompt interno.

Descrizione delle Attività

- Creare payload malevolo contenente istruzioni che devono essere eseguite dal modello.
- Incorporare il payload all'interno di materiali recuperati automaticamente dal modello (es. descrizioni, commenti HTML, note nascoste, contenuti offuscati).
- Configurare una semplice applicazione basata su LLM che recuperi automaticamente i contenuti esterni (es. web scraping controllato, API simulate, file caricati dall'utente). Potete utilizzare i modelli di Google e.g Gemini 2.5 Flash tramite API Key.
- Implementare pipeline che dimostri come l'attacco si innesca senza intervento diretto dell'utente finale.
- Studiare possibili contromisure

Output Atteso

Per sostenere l'esame orale gli studenti dovranno:

1) Implementare un proof-of-concept funzionante di un attacco di indirect prompt injection.

2) Effettuare un'analisi dettagliata delle condizioni necessarie all'esecuzione dell'attacco.

3) Suggerire linee guida e best practice di mitigazione per sviluppatori di applicazioni basate su LLM.

4) Preparare presentazione con le fasi dell'attacco, dimostrazione dell'attacco e linee guida su come prevenirlo

5) Consegnare via email codice proof-of-concept e presentazione

Creazione Database di Sorgenti C/C++ Vulnerabili e non Vulnerabili (CWE Top 25)

Obiettivo del Progetto

L'obiettivo di questo progetto è la creazione manuale di un database di file sorgenti in C/C++ che includa:

- File vulnerabili, contenenti esempi pratici delle Top 25 vulnerabilità CWE (Common Weakness Enumeration).
- File non vulnerabili, di uguale complessità ma che non contengano nessuna vulnerabilità.

Questo dataset sarà strutturato in modo da facilitare analisi statiche, addestramento di strumenti di rilevamento automatico delle vulnerabilità e valutazione delle prestazioni di LLM nel rilevare vulnerabilità.

Descrizione delle Attività

Lo studente dovrà

- Scegliere una vulnerabilità della CWE Top 25 Most Dangerous Software Weaknesses (es. buffer overflow, integer overflow, use-after-free, ecc.).
- Scrivere 5 programmi vulnerabili in C/C++, ciascuno contenente una singola vulnerabilità CWE specifica.
- Creare 5 programmi non vulnerabili che abbiano la stessa complessità di quelli vulnerabili
- Annotare ogni file con:

  Nome CWE (es. CWE-120: Buffer Copy without Checking Size)

  Breve descrizione della vulnerabilita'

  Strumenti di compilazione/test consigliati (es. gcc, valgrind, clang, ecc.)
- Organizzare i file in un database leggibile e riutilizzabile, con convenzioni di naming e struttura cartelle.

Output Atteso

Per sostenere l'esame, lo studente dovrà consegnare un file .zip contenente il database e un report. Il report dovrà includere:

- per ogni programma vulnerabile, una descrizione del funzionamento del programma, l'indicazione del punto in cui è presente la vulnerabilità, la spiegazione del motivo per cui si verifica e di come può essere sfruttata da un attaccante;
- per ogni programma non vulnerabile, una spiegazione del funzionamento e del motivo per cui il programma non risulta vulnerabile alla CWE scelta.

Il database dovrà essere organizzato in due sottocartelle: una contenente i programmi vulnerabili alla CWE selezionata e una contenente i programmi non vulnerabili.

Analisi di conformità dei siti web

Obiettivi del progetto

Il progetto ha l'obiettivo di analizzare la presenza di possibili violazioni del GDPR e dell'ePrivacy Directive (ePD). L'attenzione si concentra in particolare sulle violazioni relative al consenso, tra cui:

- Violazioni del consenso previo: tracciamento o memorizzazione di cookie prima che l'utente dia il consenso.
- Violazioni del rispetto della scelta dell'utente, ossia casi in cui il sito continua a tracciare nonostante l'utente neghi il consenso.

Descrizione delle Attività

Il progetto utilizza OpenWPM, una piattaforma di misurazione del web ampiamente usata per analizzare il tracciamento online, con lo scopo di raccogliere cookie, traffico di rete e comportamenti dei siti sotto diverse condizioni di consenso.

Lo studente riceverà dal docente un elenco di 10 siti web da analizzare.

Lo studente dovrà implementare uno script Python per raccogliere con OpenWPM per ciascun sito web: i cookies salvati nel browser e il traffico di rete in tre diverse condizioni: non esprimo nessuna scelta, consenso negato e consenso accettato.

Lo studente dovrà implementare uno script che rileva le seguenti violazioni del GDPR:

- V1 – No Consent Before Storing: presenza di cookie non necessari senza aver ottenuto il consenso.
- V2 – No Consent Before Sending: contatto con domini tracker senza aver ottenuto il consenso.
- V3 - se i siti continuano a salvare cookie non necessari,e contattare domini di tracciamento, anche quando l'utente ha negato il consenso.

Lo studente dovrà analizzare quali di queste violazioni sono presenti nei siti da analizzare.

Output Atteso

Per sostenere l'esame lo studente dovrà consegnare via email:

- Uno script OpenWPM in Python per catturare cookie, traffico di rete e screenshot.
- Un script di analisi che rilevi la presenza delle violazioni .

Un report strutturato per ciascun sito con:

- cookie memorizzati prima del consenso,
- domini tracker contattati,
- violazioni rilevate