NORDAKADEMIE
HOCHSCHULE DER WIRTSCHAFT

Wie lassen sich ML- und LLM-Systeme angesichts Angriffsformen in den Bereichen Inferenz, Training und RAG wirksam absichern, und wie ergänzen regulatorische Governance-Mechanismen die Grenzen technischer Verteidigungsstrategien?

Bearbeitungszeit der Hausarbeit:

Sonntag, den 30.11.2025 bis Sonntag, den 21.12.2025

Erarbeitet von:

Gruppe A

Anastasia Klat, Matrikelnummer: 13275

Felix Pommerening, Matrikelnummer: 13243

Julian Falk, Matrikelnummer: 13498

Studiengang: Technische Informatik

Zenturie: T23a

WPM - Einführung in die KI

Dozenten und Prüfer:

Prof. Dr. Arne Ewald

Dr. Fereshta Yazdani

Zur besseren Lesbarkeit wird in dieser Hausarbeit das generische Maskulinum verwendet. Die in dieser Arbeit verwendeten Personenbezeichnungen beziehen sich - sofern nicht anders kenntlich gemacht - auf alle Geschlechter.

Inhaltsverzeichnis

Abkürzungsverzeichnis

FCFS	First-Come-First-Serve

1 Einleitung

2 Grundlagen

2.1 Begriffserklaerungen

2.1.1 Deep Learning, ML und LLMs

2.1.2 Adversarial Attacks

2.1.3 Daten

2.2 Angriffsziele

3 Angriffe in der Inferenzphase

4 Angriffe in der Trainigsphase

5 RAG - spezifische Angriffe

6 Verteidigungsstrategien

7 Governance und ethische Ueberlegungen

8 Fazit und Ausblick

Praktisches?

Falls am Anfang nicht gepullt wurde und es lokale Änderungen gibt, kann es zu Konflikten kommen. Um das zu vermeiden, können lokale Änderungen vorher weggespeichert werden:

```
1    git status                          # nur zum Ueberblick
     git stash push -m "mein Stand"      # lokale Aenderungen wegpacken
     git pull                            # jetzt kannst du gefahrlos pullen
     git stash pop                       # deine Aenderungen wieder
        draufspielen
```

Struktur

Erste Ideen für die Gliederung der Hausarbeit:

```
1. Einleitung(1 - 1,5 Seiten) -- alle (file01)
    1.1 Motivation
    1.2 Ziel der Arbeit
    1.3 Aufbau der Arbeit
2. Grundlagen(2 - 2,5 Seiten) -- alle
    2.1 Begriffserklaerungen
        2.1.1 Machine Learning, Deep Learning, LLMs -- Anastasia(
           file02)
        2.1.2 Adversarial Attacks -- Felix(file03)
        2.1.3 Daten -- Julian(file04)
    2.2 Angriffsziele --Julian(file05)
        2.2.1 Confidentiality, Integrity, Availability (CIA-Triad)
3. Angriffe in der Inferenzphase(3 Seiten) -- Felix(file06)
    3.1 Adversarial Examples (Vision, Text, Multimodal)
        3.1.1 Gradient-based Attacks (FGSM, PGD)
        3.1.2 Transferability
    3.2 Prompt Injection auf LLMs
        3.2.1 Jailbreaks, Role Overrides, Indirect Prompt Injection
    3.3 Model Extraction
        3.3.1 API-Abfragebasierte Rekonstruktion
        3.3.2 Model Inversion \& Membership Inference
        3.3.3 Datenschutzrelevante Risiken
4. Angriffe in der Trainingsphase(3 Seiten) -- Julian(file07)
    4.1 Data Poisoning
    4.2 Backdoor Attacks
    4.3 Supply-Chain Risiken
5. RAG-spezifische Angriffe(3 Seiten) -- Anastasia(file08)
    5.1 Retrieval Layer als Angriffsflaeche
    5.2 Malicious Documents in Vektordatenbanken
    5.3 Tool-Injection / Action Hijacking
    5.4 Retrieval-Sanitization als Verteidigungsstrategie
6. Verteidigungsstrategien(1-2 Seiten) -- Julian und Felix(file09-12)
    / erstmal notizen
    6.1 Adversarial Training
    6.2 Input Sanitization & Anomaly Detection
    6.3 Robustness Evaluation & Zertifizierung
    6.4 Monitoring & Incident Response
7. Governance und ethische Ueberlegungen(1 Seite) -- Anastasia(file13
```

```
    )
      7.1 Regulatorische Anforderungen
      7.2 Ethische Implikationen von Adversarial Attacks
      7.3 Verantwortungsbewusster Umgang mit KI-Sicherheit
  8. Fazit und Ausblick(1-1,5 Seiten) -- alle(file14)
```

Eigenständigkeitserklärung

Mit meiner Unterschrift versichere ich, dass ich die hier vorliegende Arbeit selbständig, ohne fremde Hilfe und nur mit den angegebenen Hilfsmitteln verfasst habe und meine Angaben zu den verwendeten Quellen der Wahrheit entsprechen und vollständig sind. Alle Quellen, aus denen ich wörtlich oder sinngemäß übernommen habe, habe ich als solche gekennzeichnet. Darüber hinaus versichere ich,

dass ich sämtliche Teile der vorliegenden Arbeit, die unter Zuhilfenahme künstlicher Intelligenz (KI) generiert wurden, als solche gekennzeichnet habe und deren Entstehung in einer beigefügten Prozessdokumentation nachgewiesen habe. Ich habe zur Kenntnis genommen, dass zuwiderlaufendes

Verhalten als Täuschungsversuch gewertet wird und zu den in der geltenden Prüfungsverfahrensordnung genannten Konsequenzen führen wird.

Ort, Datum	Unterschrift
Ort, Datum	Unterschrift
Ort, Datum	Unterschrift