Fraunhofer AISEC

Fraunhofer-Institut für Angewandte und Integrierte Sicherheit AISEC

Generative KI: Fluch oder Segen für die Cybersicherheit?

Claudia Eckert,

Fraunhofer AISEC, TU München, acatech

19.1. 2026 Omnisecure 2026

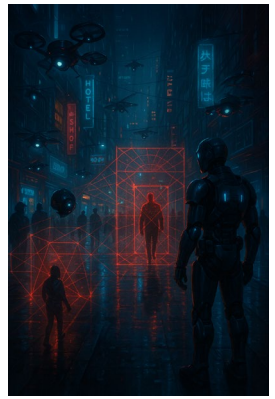
Agenda

1. **Einordnung:** (Generative) KI: Nutzen/Grenzen?!
2. **Fluch?** Risikopotentiale: Neue Angriffsflächen und typische Fehlerszenarien
3. **Segen?** Potenziale: Wo generative KI sinnvoll unterstützen kann
4. **Regulatorik:** Anforderungen an vertrauenswürdige KI
5. **Ausblick:** Resilienz als Gestaltungsaufgabe
6. **Take home Message**

1. Einordnung: (Generative) KI: Fluch oder Segen?

Automatisierte Fertigung

Personalisiertes Gesundheitsmonitoring



Predictive Crime

KI-gesteuerte Energienetz

KI-unterstütztes Trading

KI in der Verteidigung

1. Einordnung: Generative KI

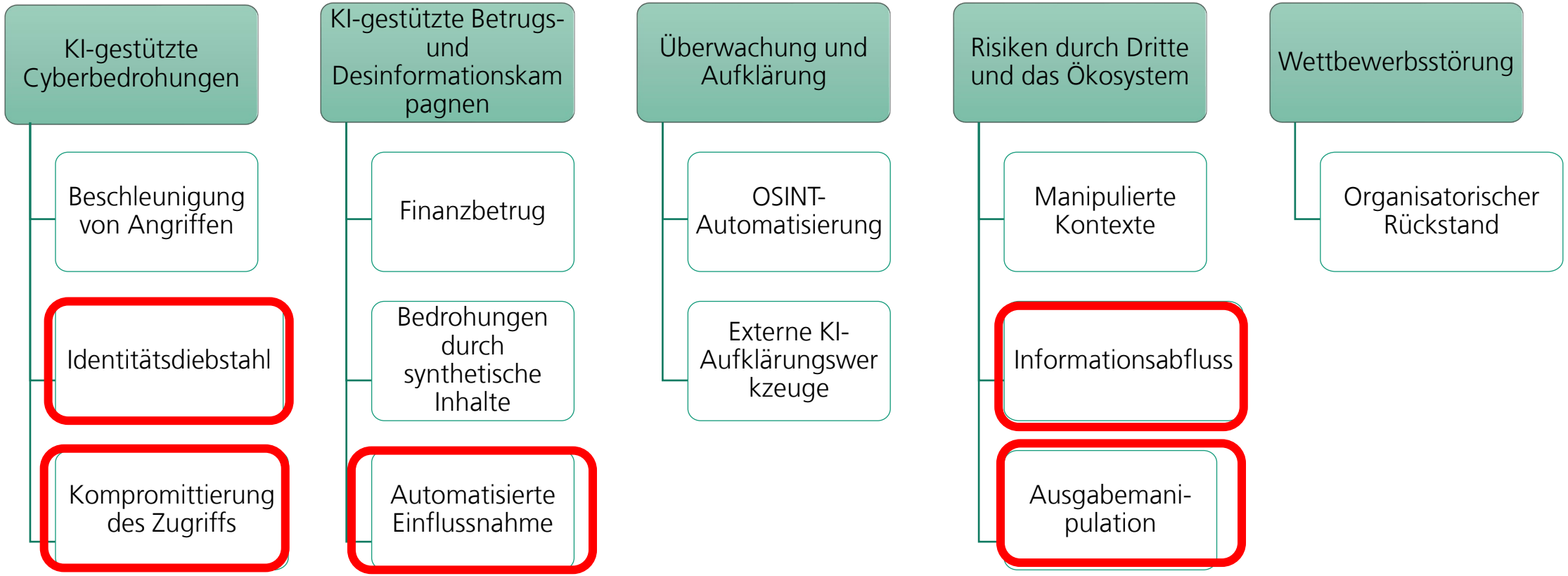
Generative KI und die Bedeutung für die Cybersicherheit:

- **Fluch?** Führt sie zu verstärkten Sicherheitsrisiken?
- **Segen?** Hat sie das Potential, das Sicherheitslevel in Unternehmen zu erhöhen?
- **Fluch?** Stoßen klassische Sicherheitsparadigmen an ihre Grenzen?
- **Segen?** Hilft sie, Software-Qualität nachhaltig zu verbessern, SW sicherer zu machen?
- **Fluch?** Trägt sie zur Verstärkung unserer digitalen Abhängigkeit bei?
- **Segen?** Hilft sie, Regularien automatisiert, nachweislich umzusetzen?

2. Fluch: Erweiterte Angriffsmöglichkeiten durch generative KI-Systeme

Zusammenfassender Überblick über Bedrohungen für Unternehmen durch KI (Basis OWASP 2025

(1) Angriffe **auf** generative KI führen zu Nutzungs-Risiken: **Beispiele**



Fraunhofer AISEC

2. Fluch: Erweiterte Angriffsmöglichkeiten durch generative KI-Systeme

Angriff: Prompt Injection: gezielte Manipulation von Eingabedaten

- **Ziel:** u.a. Einschleusen von Befehlen, wodurch z.B. Daten abfließen können
- **Einfaches Beispiel:** Indirekte Prompt Injection über einen Chatbot
  - Angreifer platziert Malware auf Webseite: z.B. Nutzerfragen an Angreifer senden
  - Nutzer verwendet Chatbot mit seinen, ggf sensitiven Eingabedaten
  - Chatbot nutzt manipulierte Webseite, Weitergabe sensitiver Daten an Angreifer

Angriff: Risiken durch Ausgaben/Antwortverhalten:

Ziele: z.B. aus Antworten lassen sich Rückschlüsse auf die sensible Trainingsdaten ziehen.

- Angreifer schleust über LLM Schadcode ein, der ungeprüft ausgeführt wird.
- LLM generiert fehlerhaften Programmcode, Integration in vorhandene Software

2. Fluch: Erweiterte Angriffsmöglichkeiten durch generative KI-Systeme

(2) Angriffe **durch** generative KI**:** **Neue Angriffsvektoren** durch KI-generierte Angriffe

- **Datenbasierte Angriffe:** Data-Poisoning in Trainingsdaten:
  - **Ziel:** Antwortverhalten des LLM wird gezielt manipuliert

- **Supply Chain Angriffe:**
  - **Ziel:** unsichere, vortrainierte Modelle oder unsichere Plug-ins von Dritten als Bestandteil von LLMs platzieren

- **Agenten-basierte Angriffsgenerierung**
  - **Ziel:** Automatisierte Lagebilderstellung und automatisierte Generierung zugeschnittener Angriffe
    - Nutzen von OSINT Tools: Orchestrierung von autonomen Agenten: Schwachstellen-Suche,….**,**
    - Automatisierte Exploit-Generierung: Zusammenführung strukturierter, unstrukturierter Daten

2. Fluch: Erweiterte Angriffsmöglichkeiten durch generative KI-Systeme

Show-Case von Anthropic in 2025: KI als Angriffskoordinator

- **LLM entwickelt eine reale Angriffskampagne:**
  - Lagebilderstellung für Zielpersonen: Automatisierte Reconnaissance (Auswertung öffentlich verfügbarer Informationen zu Personen, Rollen, Technologien),
  - Personalisieren: Erstellung stark personalisierter Phishing-Nachrichten,
  - Adaptieren: Anpassung von Tonalität der Nachrichten und Vorgehen basierend auf Reaktionen,
  - Orchestrierung: Koordination mehrerer Angriffsschritte über Prompts, Kontext und externe Tools
- **Charakter des Angriffs**
  - KI übernimmt autonom, proaktive Aufgaben entlang der Angriffskette
  - Angriff wird wirksamer, schneller, skalierbarer und adaptiver

Demonstration: Das LLM agiert als kampagnenfähiger Koordinator, nicht nur als Assistenzwerkzeug

2. Fluch: Erweiterte Angriffsmöglichkeiten durch generative KI-Systeme

Angriff gilt als ein Wendepunkt!

- Was neu ist
  - Keine neuen Exploits oder Zero-Days erforderlich
  - End-to-End-Automatisierung ganzer Angriffsketten
  - Kontinuierliche Anpassung ohne permanente menschliche Steuerung
  - Deutlich gesenkte Eintrittsbarrieren für komplexe Angriffe

- Warum klassische Security hier an Grenzen stößt
  - Angriff entsteht durch Zusammenspiel von Kontext, Entscheidung und Aktion, aber Traditionelle Abwehr fokussiert auf Angriffsteilschritte und einzelne Angriffs-Signaturen
  - Geschwindigkeit der Angriffsdurchführung überholt manuelle Analyse- und Reaktionsprozesse
  - Fehlende/fehlerhafte lokale Schutzmaßnahmen wirken sich schnell über Systemgrenzen aus.

2. Fluch: Erweiterte Angriffsmöglichkeiten durch generative KI-Systeme

Angriff gilt als ein Wendepunkt!

- Zentrale Erkenntnis
  - Die Hauptangriffsfläche ist nicht das Modell, sondern
  - die Integration von KI in Daten-, Tool- und Aktionssysteme
  - Kontext (Prompts, Feedback, externe Quellen) wird Teil der Steuerlogik

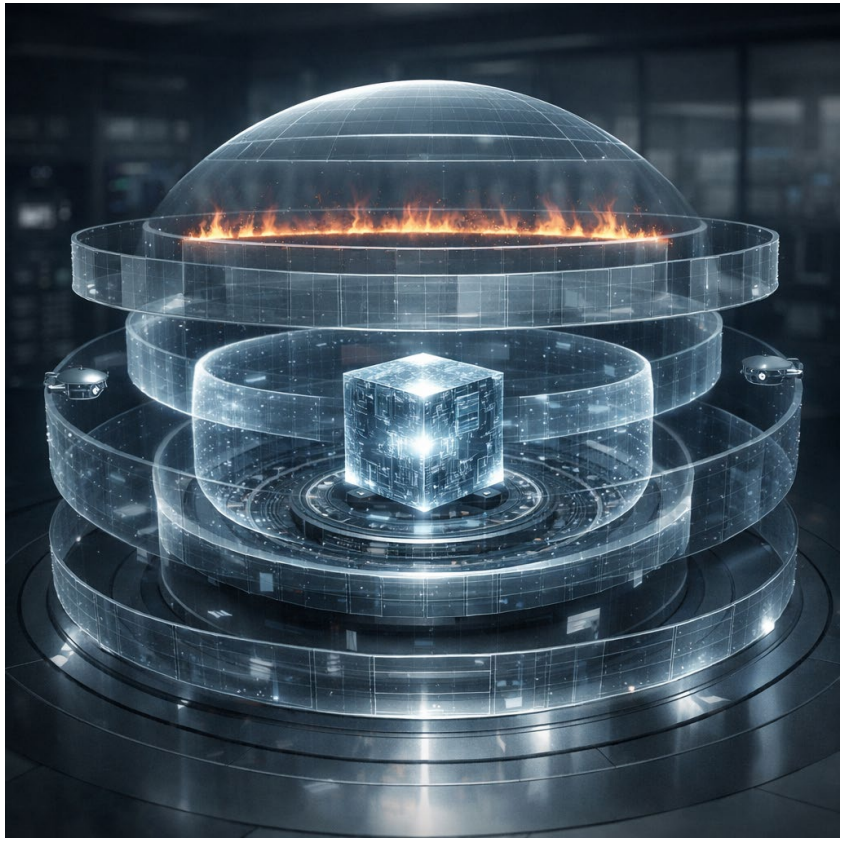
- Implikation
  - Generative KI wirkt als **systemischer Risikobeschleuniger**
  - Sicherheitsfragen verlagern sich von Algorithmen zu **Architektur, Governance und Kontrolle**

Fazit: Das Risikopotential wird erheblich verstärkt durch KI

Zwischenfazit: Absicherung von KI-Systemen ist sine qua non!

Defense-in-Depth für KI-Systeme

- **Vorbereitung/Pre-processing**
  - Input-Sanitization, Prompt-Shielding
  - Absicherung von RAG-Quellen (Allowlists, Content-Scanning)
- **Modell & Lieferkette**
  - Robustheitstests, Red-Teaming, Poisoning-Detection, …
  - Signieren von Modellen und Datensätzen, …
  - Dokumentation von Training, Fine-Tuning und Datenquellen
- **Ausgabe-Nachbearbeitung/Post-processing**
  - Output-Moderation, Policy-Checks, Human-in-the-Loop
  - API-Hardening, Authentifizierung, Context-Isolation
  - Rate-Limiting, Monitoring, Logging
  - Erkennung und Bewertung von Halluzinationen



3. Segen: Wo generative KI sinnvoll unterstützen kann

- **Sichere Softwareentwicklung/Sicherheitsanalyse: u.a.**
  - Code-Reviews: Schwachstellen? Schadcode-Teile?
  - POI-Erkennung: Unterstützung bei Analyse

- **Security Operations: u.a.**
  - Verbesserte Lagebeurteilung: z.B. Log-File Aufbereitung
  - Kontextualisierung: Hintergrundinformation zu Alarmen
  - Automatisierte Report Generierung, z.B. bei Incidents

- **Detection Engineering:** proaktiv neue Regeln/Tests entwickeln
  - Analyst beschreibt: „was“ erkannt werden soll,
  - KI generiert das „wie“: Prüfcode zu neuer Regel, Testverfahren, …

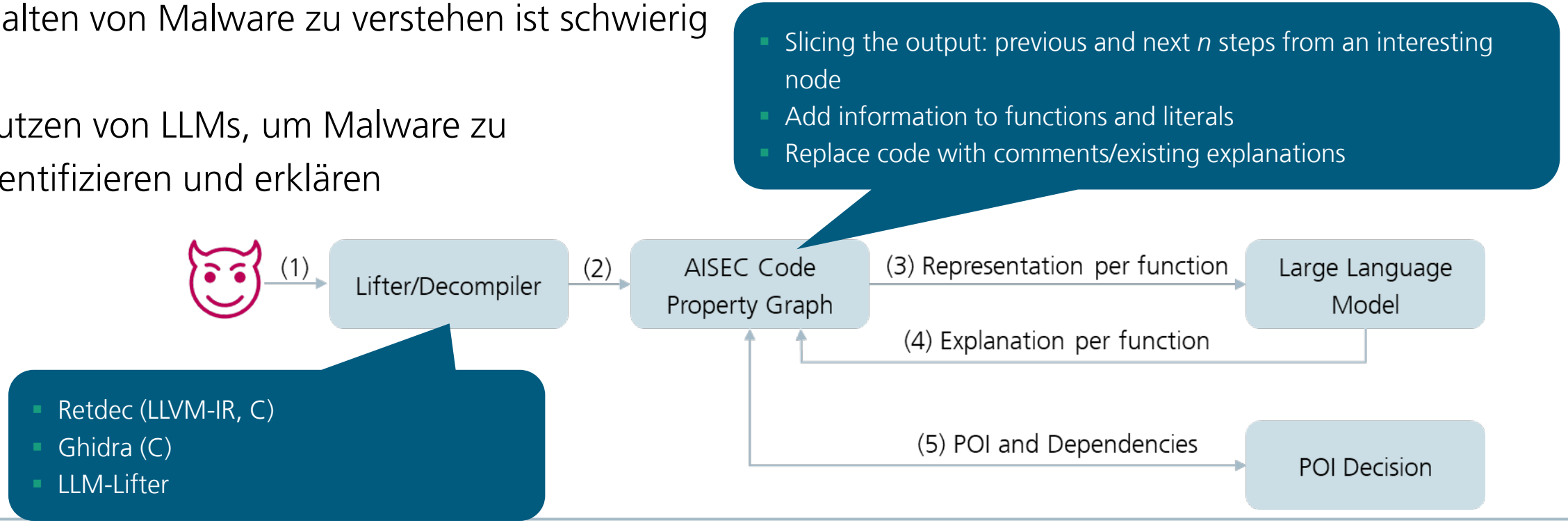
3. Segen: Wo generative KI sinnvoll unterstützen kann

Beispiel: LLMs, um Schadcode zu erkennen und dessen Verhalten zu erklären

Problem:

- Binaries sind schwer verständlich (Black-Boxes)
- Angreifer nutzen Obfuskation, um relevante Muster und Charakteristika zu verschleiern
- Verhalten von Malware zu verstehen ist schwierig

Ziel: Nutzen von LLMs, um Malware zu identifizieren und erklären



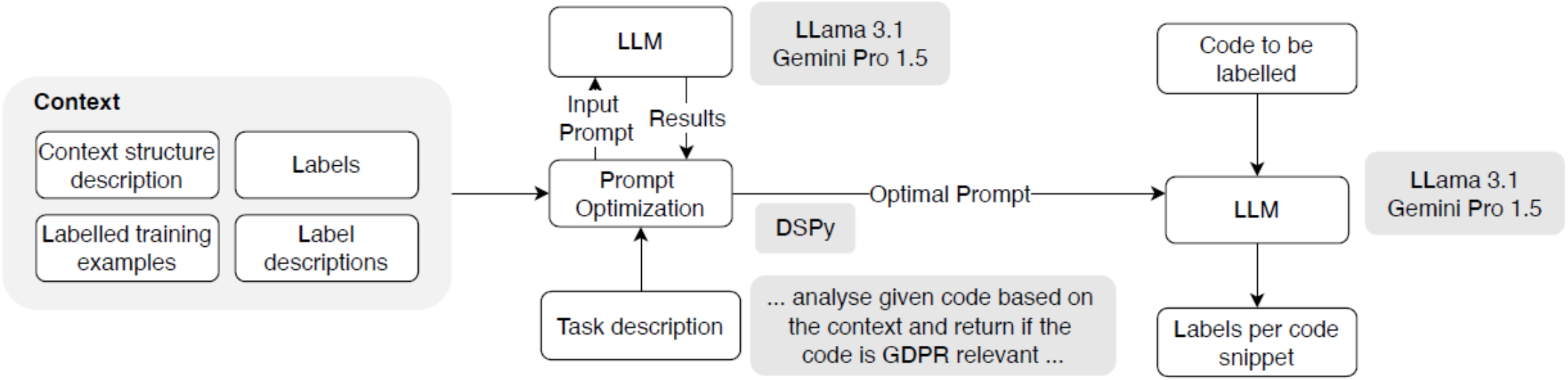
3. Segen: Wo generative KI sinnvoll unterstützen kann

Beispiel: Automatische Erkennung von DSGVO-relevanten Codestellen mittels LLM

Problem:

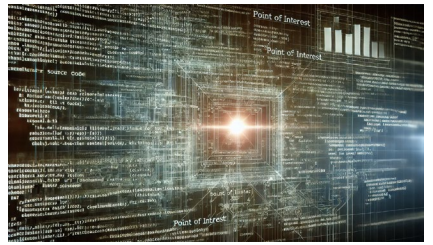
- Privacy Review von Source Code muss stets aktuell sein
- Sehr arbeitsintensiv

Ziel: LLMs für die Erkennung von persönlichen Daten

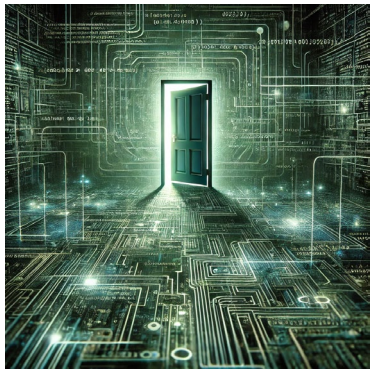


Zwischenfazit: GenAI für Cybersicherheit hat viel Potential

- **Generative KI als Assistenz**
  - für Sicherheits-Analysten, Software-Entwickler: sicherer Code
  - für System-Verantwortliche: Compliance Prüfung, Detection Engineering
  - für die Ausbildung: lernen, erklären fehlerhafter Code-Snippets
  - für personalisierte Sicherheits-Unterstützung: zugeschnitten auf Experten-Level



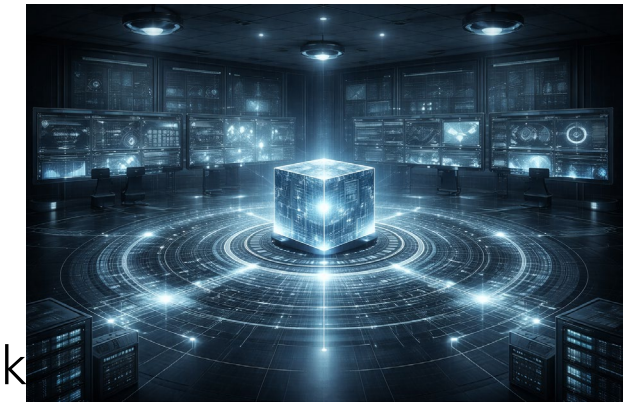
- **LLMs zur Erkennung und Erklärung**
  - Erkennung von Schwachstellen durch z.B. Ähnlichkeitssuche auf CWE- oder CVE-Einträgen
  - Identifikation, Erklärung von Root Causes von Crashes bei Fuzzing-Tests



Aber offene Diskussion:

- Wo sind Grenzen für den Einsatz von Gen AI?
- Nutzung autonomer Agents versus Souveränität und Kontrollierbarkeit?

4. Anforderungen an vertrauenswürdige KI-Ökosysteme



Ziel: Sichere und vertrauenswürdige Nutzung von KI-Ökosysteme

- Blick auf KI-Modelle/Algorithmik greift zu kurz!

- Kontrollierbarkeit:
  - klare Systemgrenzen, differenzierte Rechte, klare Rollen/Verantwortlichk
- Robustheit:
  - Schutz vor Prompt-Injection, adversarialen Angriffen, Data-Poisoning
- Integrität & Provenance:
  - Daten-Qualität, Herkunftsnachweise, Versionierung und Signierung von Modellen und Daten
- Nachvollziehbarkeit:
  - Logging, Traceability, reproduzierbare Entwicklungs- und Betriebsprozesse
- Betriebssicherheit:
  - Monitoring, Incident-Management, kontrollierte Updates, Einschränkbarkeit von KI-Funktionen

Erforderlich: Governance-Strukturen mit Regeln, Prozessen, Verantwortlichkeiten für KI-Ökosyteme!

4. Regulatorik: Anforderungen des EU AI Act

EU Artificial Intelligence Act

Kern-Message

- Der AI Act verlangt keine perfekte KI,
- sondern kontrollierbare, nachvollziehbare und verantwortbare KI-Systeme
- Regulierung adressiert Architektur, Prozesse und Governance – nicht nur das Modell
- AISEC arbeitet an Metriken und Prüfkatalogen zur Überprüfung

- **Risikomanagement über den gesamten Lebenszyklus**
  - Systematische Identifikation und Bewertung von Risiken
  - Berücksichtigung von Fehlverhalten, Fehlentscheidungen und unbeabsichtigten Wirkungen
  - Regelmäßige Neubewertung bei Änderungen von Daten, …
- **Transparenz & Nachvollziehbarkeit**
  - Dokumentation von Modell, Datenquellen und Systemgrenzen
  - Nachvollziehbare Funktionsweise und Entscheidungslogik
  - Transparenz gegenüber Nutzern, Betreibern und Aufsicht
- **Daten- und Modellgovernance**
  - Kontrolle über Trainings-, Feinjustierungs- und Betriebsdaten
  - Maßnahmen zur Sicherstellung von Datenqualität und Integrität
  - Umgang mit Bias, Drift und unerwartetem Verhalten
- **Menschliche Aufsicht**
  - Klar definierte Rollen und Eingriffsmöglichkeiten
  - Fähigkeit zum Übersteuern, einschränken von KI-Funktionen
  - Keine vollständig unkontrollierte Automatisierung in high-risk Sz.
- **Robustheit, Sicherheit und Betrieb**
  - Schutz vor Manipulation, Misskonfiguration und Kontextangriffen
  - Monitoring im laufenden Betrieb, Incident-Handling , Updates

5. Resilienz als Gestaltungsaufgabe

Erkenntnis 1:

Generative KI-Modelle sind Teil komplexer digitaler KI-Ökosysteme

- KI-Modelle sind nicht isoliert, sie werden in Ökosystemen betrieben
- Risiken für KI-Nutzung durch starke Abhängigkeit von digitalen Technologien: Cloud, IAM

Erkenntnis 2:

- Verstärkt Risiken durch Kopplung von KI mit Aktionen (Agentic AI)
- Verstärkt Risiken durch hohen Automatisierungsgrad bei geringer Transparenz

Gestaltungsauftrag:

- Defense-in-Depth Absicherung: Modelle + Plug-ins + RAG + …. Supply-Chain
- Alternative Modelle, Wissensbasen bereitstellen: z. B. über Open Source, domänen-LLM
- Vertrauenswürdige Plattformen betreiben: z.B. basierend auf Confidential Computing

Take Home Message

Generative KI und die Bedeutung für die Cybersicherheit:

- Verändert Sicherheitsrisiken **quantitativ und qualitativ**

  ➡ Fluch!

- Hat das Potential, mittelfristig das Sicherheitslevel in Unternehmen zu erhöhen

  ➡ Segen!

Noch ist das Spiel offen: wir haben viele Gestaltungsmöglichkeiten:

- Neue Ansätze zur Steigerung der Robustheit, Nachvollziehbarkeit, Kontrollierbarkeit und Zertifizierungsschemata müssen entwickelt und integriert werden.
- Alternativen entwickeln, um digitale Abhängigkeiten zu reduzieren: z.B. OSS
- Governance und Tooling entwickeln, um Anforderungen automatisiert umzusetzen

Fraunhofer AISEC

Fraunhofer-Institut für Angewandte und Integrierte Sicherheit AISEC

Vielen Dank für Ihre Aufmerksamkeit

© Hans Georg Esch

Fraunhofer AISEC

Fraunhofer-Institut für Angewandte und Integrierte Sicherheit AISEC

Kontakt

Claudia Eckert
Fraunhofer-Institut für Angewandte und Integrierte Sicherheit AISEC
Lichtenbergstraße 11
85748 Garching bei München
marketing@aisec.fraunhofer.de
www.aisec.fraunhofer.de
@FraunhoferAISEC

Webseite

Anmeldung zum Newsletter

Cybersecurity Blog