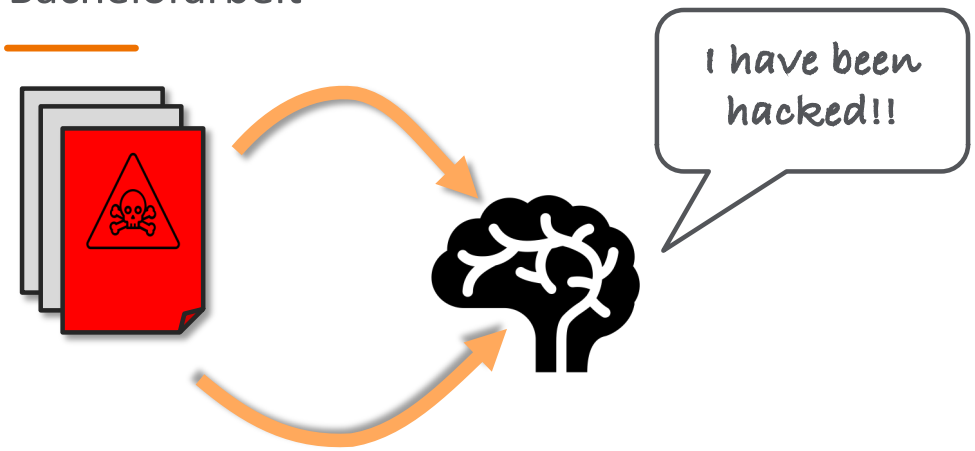
Spurensuche in Chatbot-Angriffen

Bachelorarbeit

OTH Amberg-Weiden



Hintergrund

Viele Chatbots arbeiten als RAG-Systeme mit Dokumenten im Hintergrund. Prominente Beispiele sind **E-Mail-Chatbots**.

Ein E-Mail-Bot informiert mich z.B. über wichtige Neuigkeiten in meiner Inbox, fasst meine Mails zusammen oder beantwortet Fragen zu meinen E-Mails.

Angreifer haben ein unvermeidbares: Sie schicken eine **(manipulierte) E-Mail**.

In dieser Arbeit untersuchen Sie einen eigen-entwickelten Ansatz, um manipulierte Dokumente in RAG-Systemen frühzeitig aufzuspüren und zu erkennen.

Sie können dabei auf **bestehende Arbeiten** aufbauen.

Das Thema bietet Ihnen…

- … Gelengeheit, Erfahrung in angewandter KI und Cybersecurity zu sammeln.
- … das Potential für eine wissenschaftliche Publikation.
- … das Potential zur Ausgründung.

Referenzen

Steindl, Schäfer, Ludwig, Levi. 2024. Linguistic Obfuscation Attacks and Large Language Model Uncertainty. In: *Proc. UncertaiNLP 2024*, St Julians, Malta. ACL

P. Levi. 2025. RAG-Systeme absichern. *iX* 1:50–55.