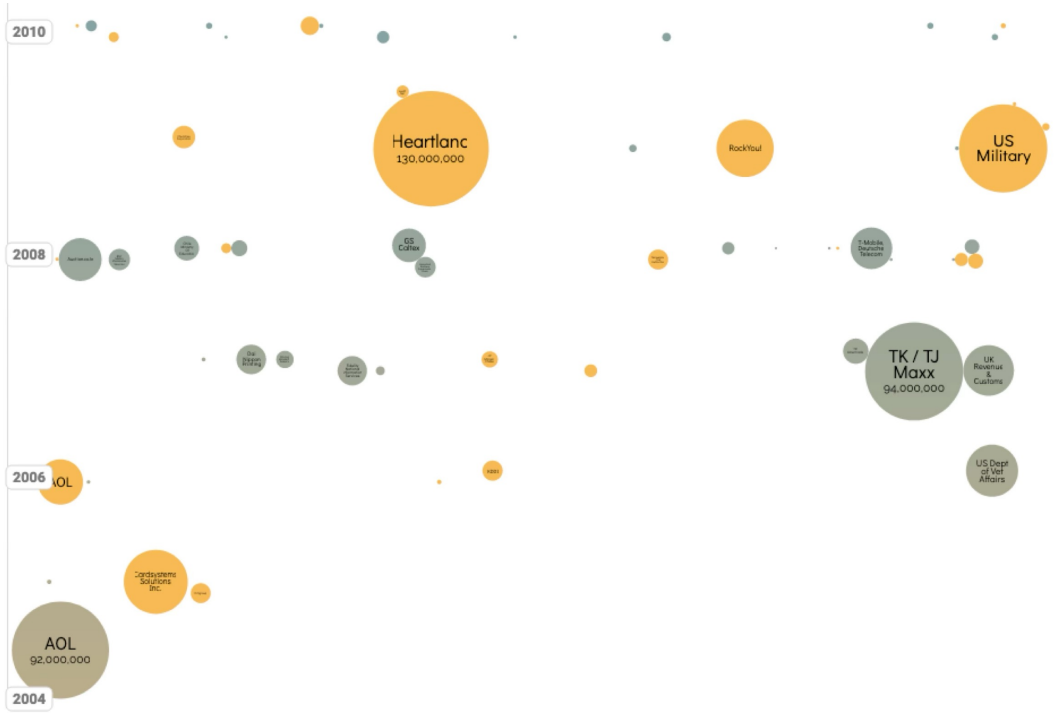
Fraunhofer AISEC

22. Januar 2025 / Dariush Wahdany

Privatsphäre-bewahrende KI in der öffentlichen Verwaltung



David McCandless & Tom Evans
Information is Beautiful

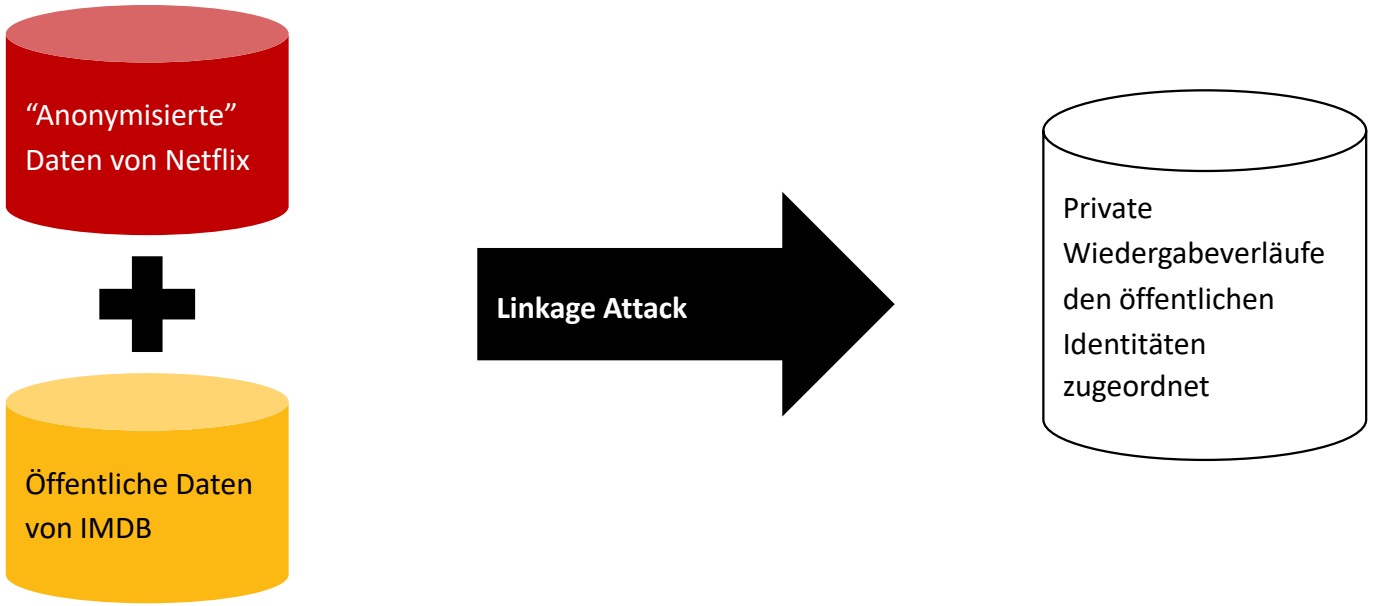
sources: New York Times, Forbes, The Guardian, Tech Radar, BBC, PC Mag, Tech Crunch & others
see the data

MADE WITH VIZsweet

Daten sind ein Wegbereiter für Angriffe

Was heißt das?

Ausnutzen von Hintergrundwissen



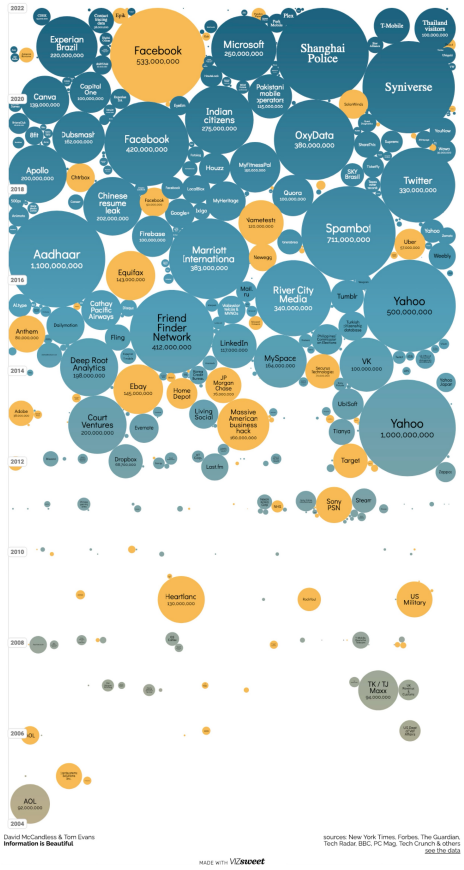
Naive Versuche der Privatsphäre **werden gebrochen werden**

Hintergrundwissen nimmt zu

Ausmaß und Häufigkeit von Datenlecks nimmt zu’

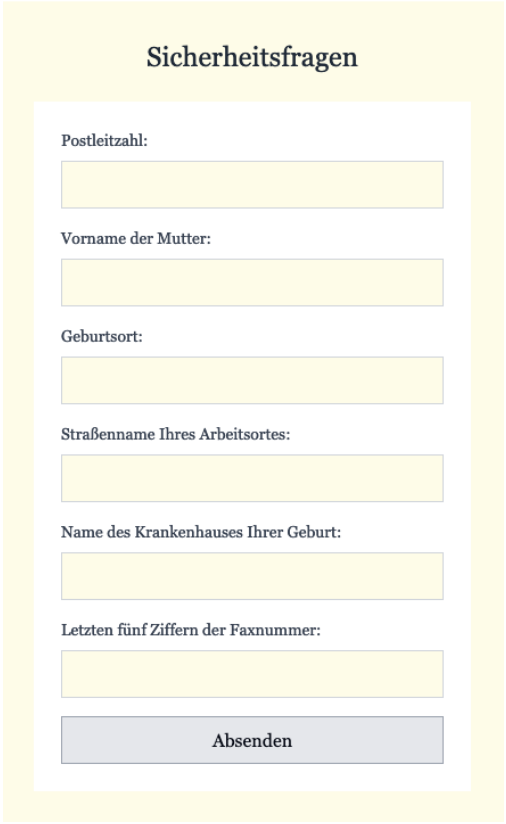
Q3’24 etwa 14,6 Millionen gehackte Nutzerkonten in Deutschland

Vermutlich über den Großteil der Bundesbürger vielfältiges Hintergrundwissen in Umlauf



Implikationen für Sicherheitskonzepte

"Sicherheits"-Faktoren basierend auf Wissen über persönliche Daten werden unsicherer



Agenda

1. Kontext
2. Künstliche Intelligenz
3. Schutzmaßnahmen & Lösungen

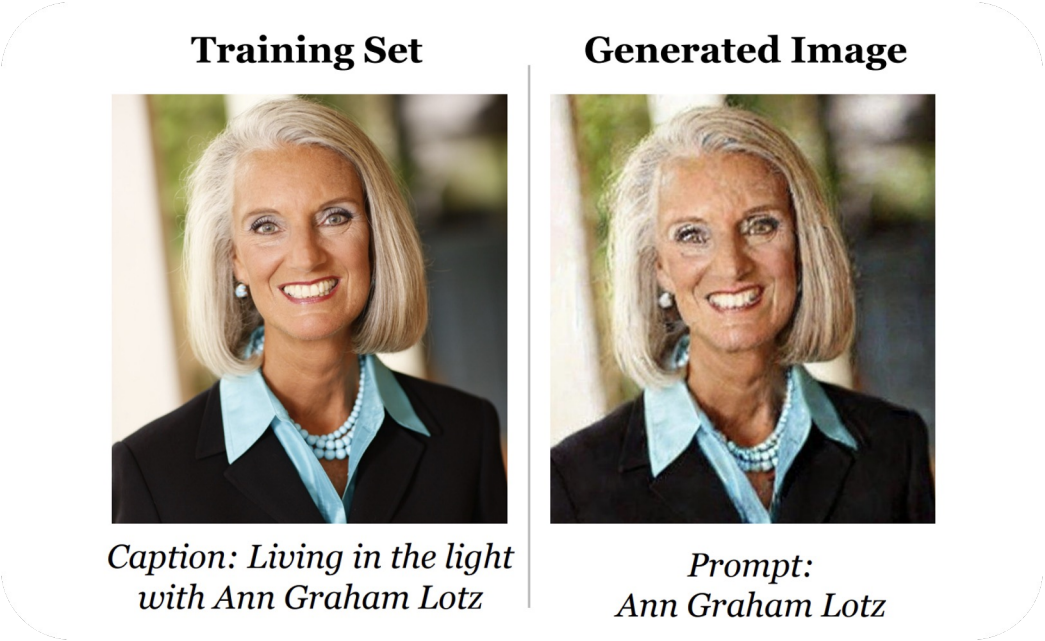
KI darf keine sensiblen Informationen preisgeben

Privatsphäre	Sicherheit	Robustheit	Vertraulichkeit
Veranwortlichkeit	Interpretrierbarkeit	Erklärbarkeit	Diskriminierungsfreiheit

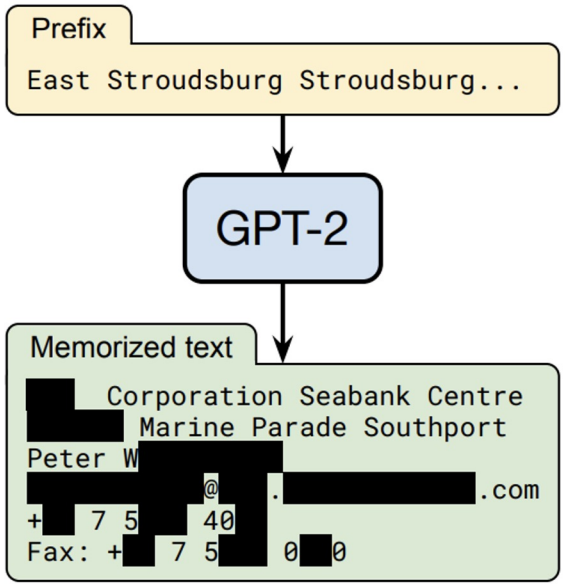
KI darf keine sensiblen Informationen preisgeben

Privatsphäre	Sicherheit	Verlässlichkeit	Vertraulichkeit
Robustheit	Interpretrierbarkeit	Erklärbarkeit	Diskriminierungsfreiheit

Warum Privatsphäre?



*Carlini et al., 2023
Extracting Training Data from Diffusion Models

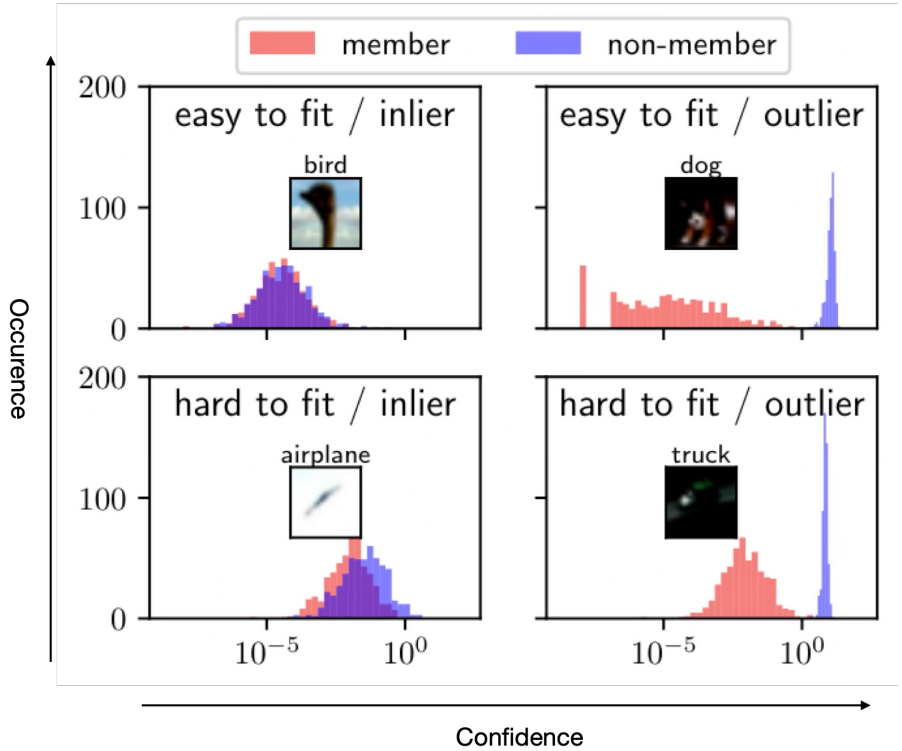


*Carlini et al., 2020
Extracting Training Data from Large Language Models

(Generative) KI reproduziert Trainingsdaten

Warum Privatsphäre?

Carlini, N. *et al*. Membership Inference Attacks From First Principles.
Preprint at http://arxiv.org/abs/2112.03570 (2022).



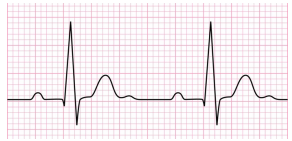
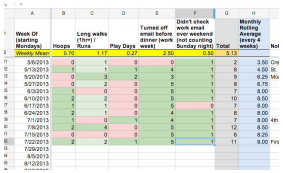
Auch nicht-generative KI gibt Informationen preis

Datenschutzrisiken von Künstlicher Intelligenz

DSGVO definiert drei Datenschutzrisiken:

- **Linkability**: Verknüpfen von Datenpunkten
- **Singling Out**: Identifizieren aller zu einer Person gehörenden Daten
- **Inference**: Ableiten von unbekannten/sensiblen Attributen aus anderen Daten

Angriffe

Model Inversion Attack	**Rekonstruktion sensibler Daten**	▪ Biometrische Erkennung	
Membership Inference Attack	**Sensible Attribute**	▪ Medizinische Klassifizierungsmodelle	
Property Inference Attack	**Offenlegung von Datensatzeigenschaften**	▪ Spracherkennung	
Prompt Injection Attack	**Weitergabe von internen Daten**	▪ Sprachmodelle (LLMs)	

Datenschutzrisiken von Künstlicher Intelligenz

DSGVO definiert drei Datenschutzrisiken:

- **Linkability**: Verknüpfen von Datenpunkten
- **Singling Out**: Identifizieren aller zu einer Person gehörenden Daten
- **Inference**: Ableiten von unbekannten/sensiblen Attributen aus anderen Daten

KI birgt diverse Datenschutzrisiken:

- Rekonstruktion sensibler Daten
- Offenlegung von Datensatzeigenschaften
- Weitergabe von internen Daten

- Warum wollen wir KI dann nutzen?

Vorteile und Anwendungsfelder von KI

Neue Anwendungsfelder:

- Semantische Suche
- Generierung natürlicher Sprache, Bilder und Videos
- Vearbeitung natürlicher Sprache
- Mehrschrittige Agenten

Interaktive Anwendungen

Personalisierung

Vereinfachung

Automatisierung

Beispielhafte Anwendungen

Betrugserkennung

- Steuerhinterziehung
- Sozialbetrug

Korruptionsbekämpfung

- Vergabeverfahren
- Beschaffungsvorgänge

Verwaltungsautomatisierung

- Bearbeitung von Anträgen
- Klassifizierung und Weiterleitung von Anfragen

Serviceverbesserungen

- Rund um die Uhr personalisierte Informationen
- Hilfe bei Anträgen

Übersetzung von Dokumenten

- Mehrsprachige Kommunikation
- Barrierefreiheit

Unterstützung von Entscheidungsträgern

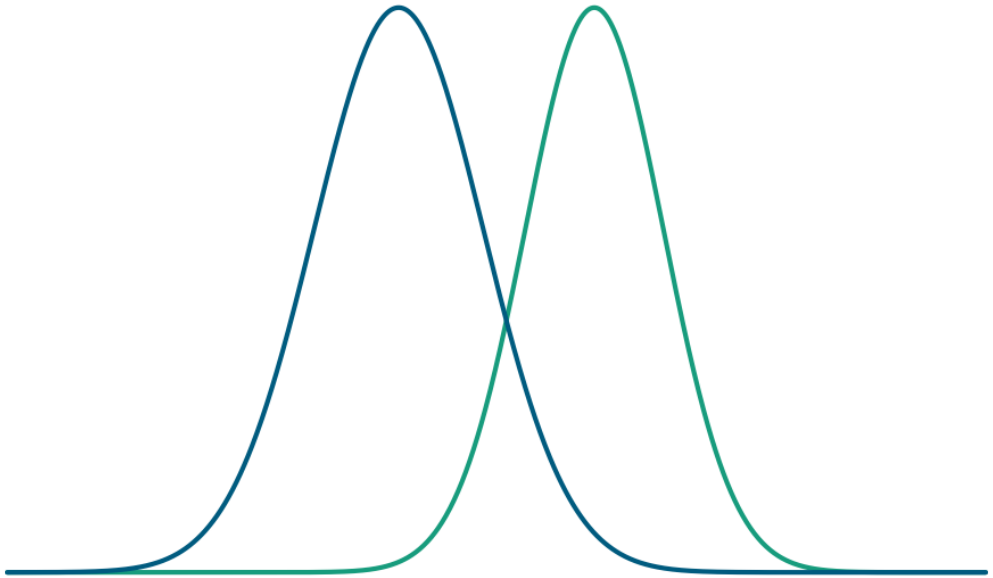
- Analyse von komplexen Daten
- Aufbereitung von Daten


VLZ 6000-00

Quantifizierbare Privatsphäregarantien mit

ε-Differential Privacy

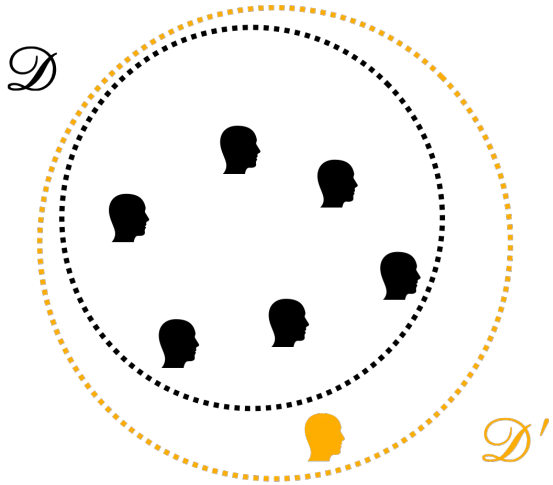
Mathemathische Grundlage für Privatsphäre



ε Differential Privacy

ε Differential Privacy

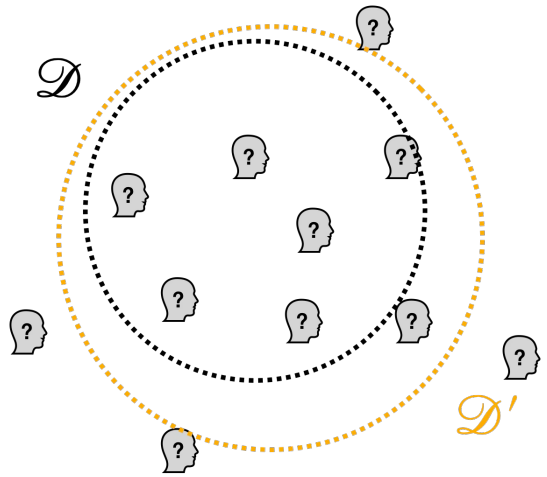
ε Differential Privacy



ε Differential Privacy

- ϵ : Stärke der Privatsphäre

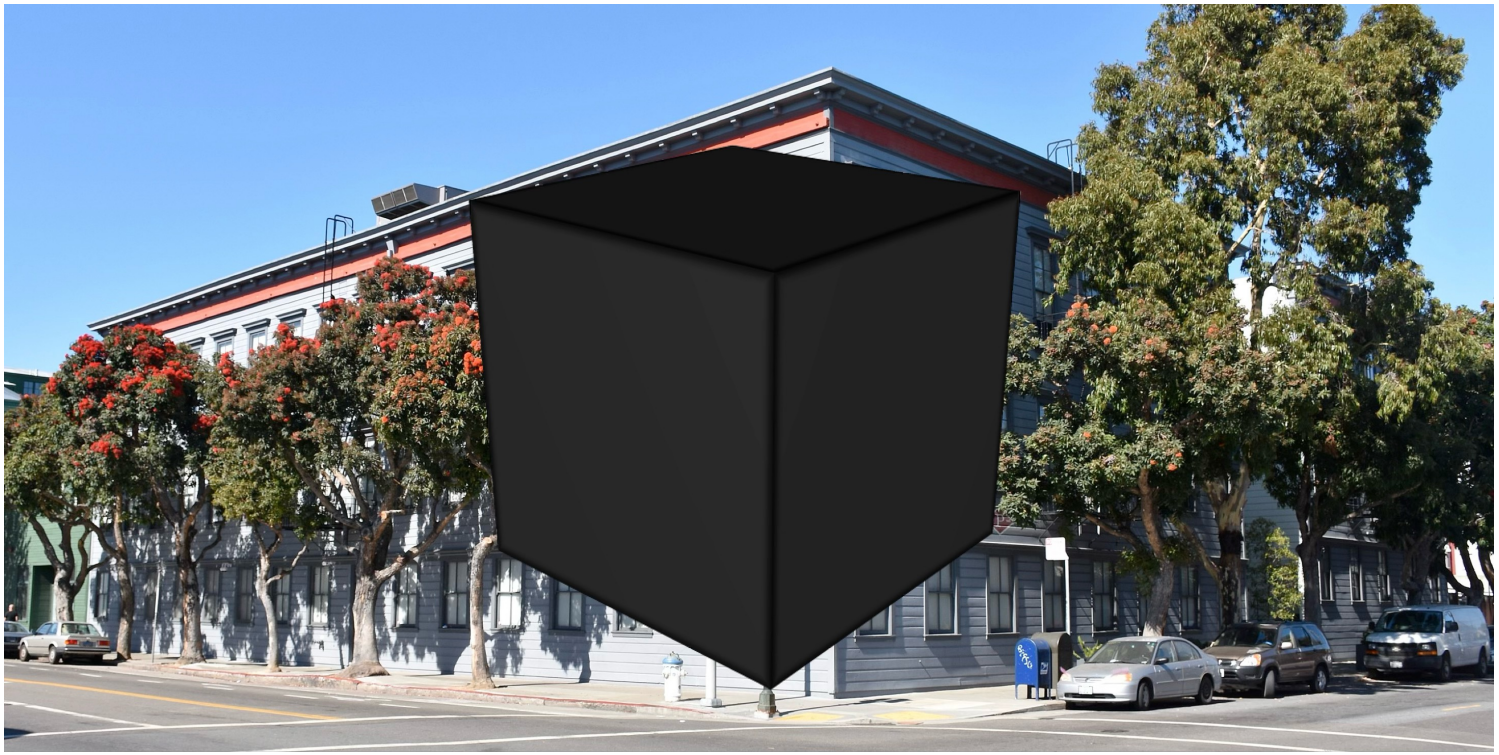
$$\Pr(\mathcal{A}(D) \in S) \leq e^{\epsilon} \cdot \Pr(\mathcal{A}(D') \in S)$$



- Mathematische Privatsphäre**garantie**

- Schutzmaßnahmen erfordern häufig **lokale Modelle**

Blackbox AI

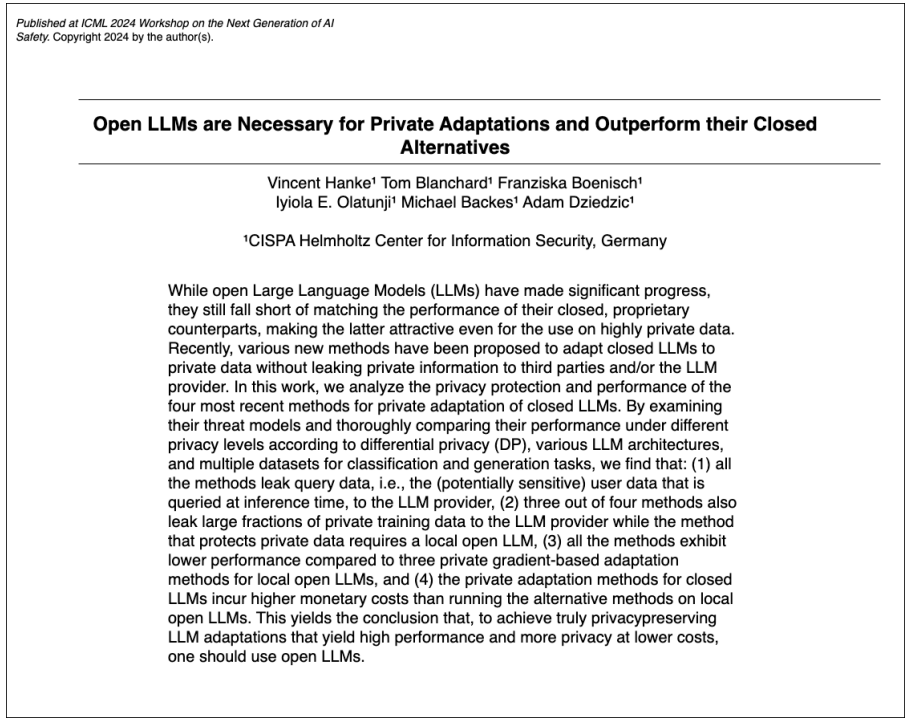


Von HaeB, CC BY-SA 4.0, https://commons.wikimedia.org/w/index.php?curid=87849278

Lokale Modelle

Lokale Modelle sind für viele Anwendungen

- Effektiver
- Günstiger
- Privatsphäre-bewahrender



Published at ICML 2024 Workshop on the Next Generation of AI Safety.

Open LLMs are Necessary for Private Adaptations and Outperform their Closed Alternatives

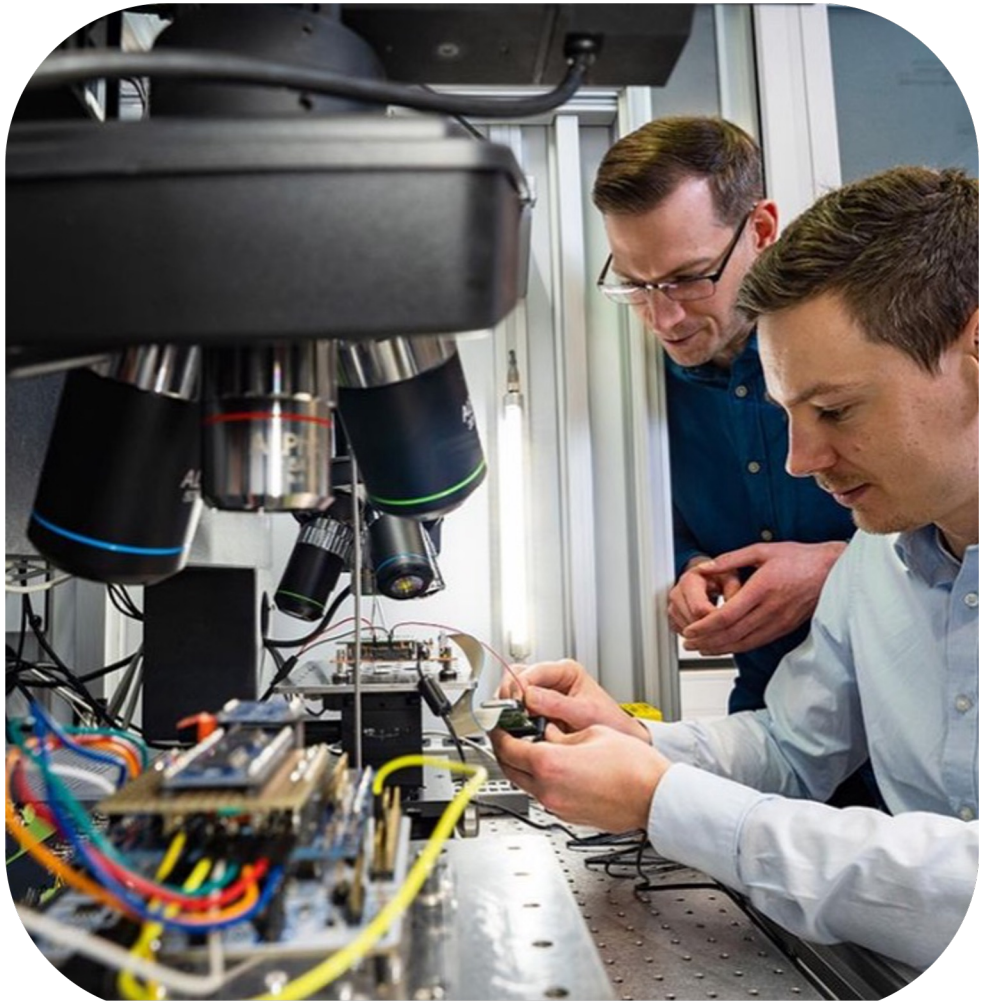
Vincent Hanke[1] Tom Blanchard[1] Franziska Boenisch[1]
Iyiola E. Olatunji[1] Michael Backes[1] Adam Dziedzic[1]

[1]CISPA Helmholtz Center for Information Security, Germany

While open Large Language Models (LLMs) have made significant progress, they still fall short of matching the performance of their closed, proprietary counterparts, making the latter attractive even for the use on highly private data. Recently, various new methods have been proposed to adapt closed LLMs to private data without leaking private information to third parties and/or the LLM provider. In this work, we analyze the privacy protection and performance of the four most recent methods for private adaptation of closed LLMs. By examining their threat models and thoroughly comparing their performance under different privacy levels according to differential privacy (DP), various LLM architectures, and multiple datasets for classification and generation tasks, we find that: (1) all the methods leak query data, i.e., the (potentially sensitive) user data that is queried at inference time, to the LLM provider, (2) three out of four methods also leak large fractions of private training data to the LLM provider while the method that protects private data requires a local open LLM, (3) all the methods exhibit lower performance compared to three private gradient-based adaptation methods for local open LLMs, and (4) the private adaptation methods for closed LLMs incur higher monetary costs than running the alternative methods on local open LLMs. This yields the conclusion that, to achieve truly privacypreserving LLM adaptations that yield high performance and more privacy at lower costs, one should use open LLMs.

Fraunhofer
AISEC

Umsetzung von IT Sicherheitsforschung in anwendungs-orientierte Lösungen



"It's going to be interesting to see how society deals with artificial intelligence, but it will definitely be cool."

Colin Angle

CEO/Co-Founder von iRobot

Kontakt

Dariush Wahdany, M. Sc.
Secure Systems Engineering
Tel. +49 89 3229986 250
dariush.wahdany@aisec.fraunhofer.de

Fraunhofer AISEC
Breite Straße 12
14199 Berlin
www.aisec.fraunhofer.de

Vielen Dank für Ihre Aufmerksamkeit

Fraunhofer AISEC