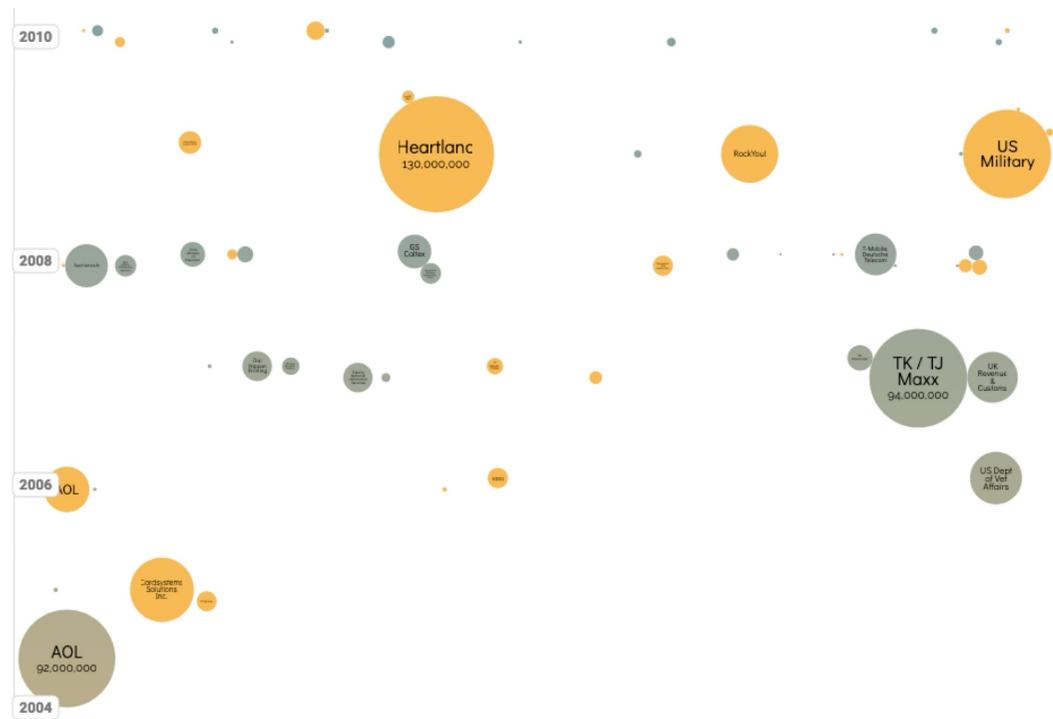


22. Januar 2025 / Dariush Wahdany

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*

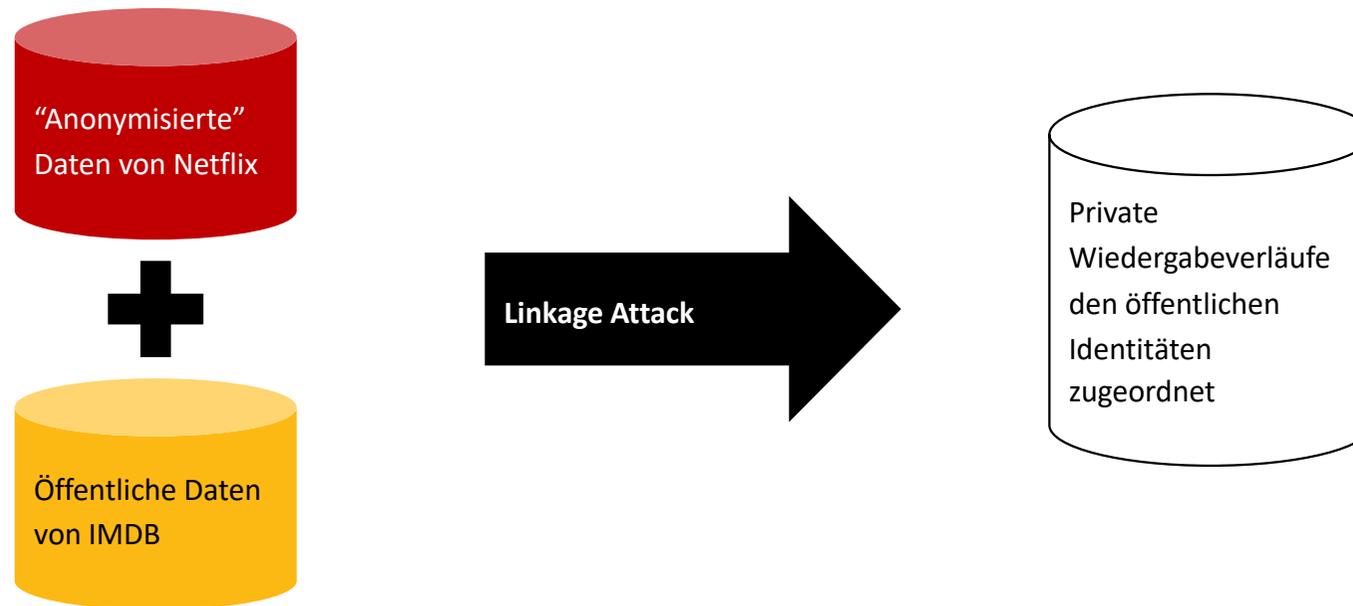
Public



Daten sind ein Wegbereiter für Angriffe

Was heißt das?

Ausnutzen von Hintergrundwissen





Naive Versuche der Privatsphäre **werden gebrochen werden**

Implikationen für Sicherheitskonzepte

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

Sicherheitsfragen

Postleitzahl:

Vorname der Mutter:

Geburtsort:

Straßenname Ihres Arbeitsortes:

Name des Krankenhauses Ihrer Geburt:

Letzten fünf Ziffern der Faxnummer:

Agenda

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

KI darf keine sensiblen Informationen preisgeben

Privatsphäre

Sicherheit

Robustheit

Vertraulichkeit

Verantwortlichkeit

Interpretierbarkeit

Erklärbarkeit

Diskriminierungsfreiheit

KI darf keine sensiblen Informationen preisgeben

Privatsphäre

Sicherheit

Verlässlichkeit

Vertraulichkeit

Robustheit

Interpretierbarkeit

Erklärbarkeit

Diskriminierungsfreiheit

Warum Privatsphäre?

Training Set



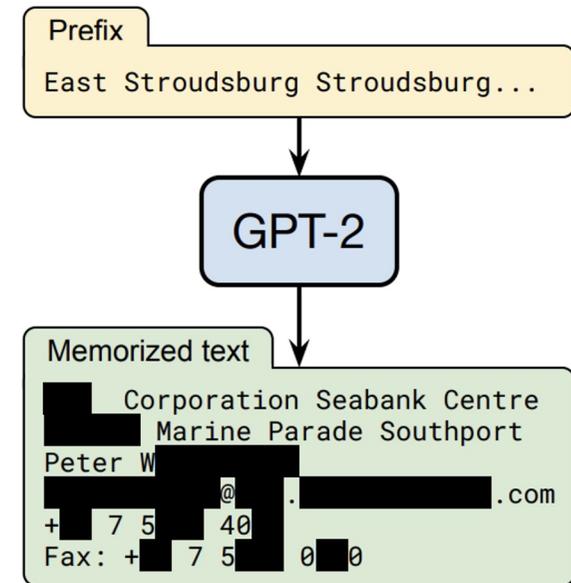
Caption: Living in the light with Ann Graham Lotz

Generated Image



Prompt: Ann Graham Lotz

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

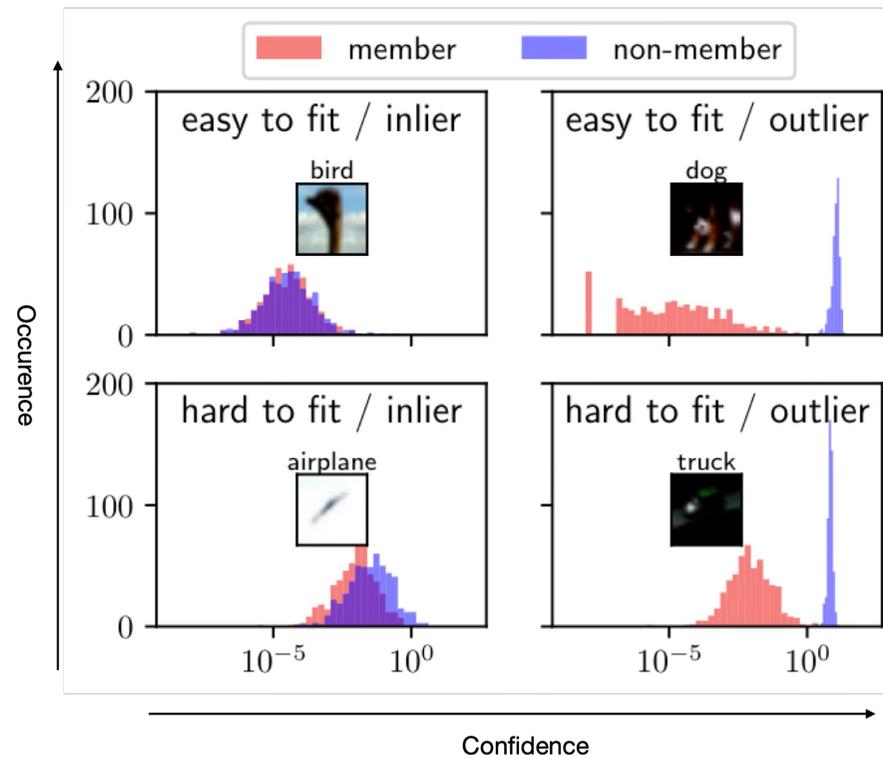


*Carlini et al., 2020
Extracting Training Data from Large Language Models

(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 unbekanntem/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)

Week Of	Long waits	Turned off	work week	Training	Testing
Monday	hours	hours	over weekend	loss	loss
Monday	hours	Play Days	week	Sunday night	Total
1	0.0	0.0	0.0	0.0	0.0
2	0.0	0.0	0.0	0.0	0.0
3	0.0	0.0	0.0	0.0	0.0
4	0.0	0.0	0.0	0.0	0.0
5	0.0	0.0	0.0	0.0	0.0
6	0.0	0.0	0.0	0.0	0.0
7	0.0	0.0	0.0	0.0	0.0
8	0.0	0.0	0.0	0.0	0.0
9	0.0	0.0	0.0	0.0	0.0
10	0.0	0.0	0.0	0.0	0.0
11	0.0	0.0	0.0	0.0	0.0
12	0.0	0.0	0.0	0.0	0.0
13	0.0	0.0	0.0	0.0	0.0
14	0.0	0.0	0.0	0.0	0.0
15	0.0	0.0	0.0	0.0	0.0
16	0.0	0.0	0.0	0.0	0.0
17	0.0	0.0	0.0	0.0	0.0
18	0.0	0.0	0.0	0.0	0.0
19	0.0	0.0	0.0	0.0	0.0
20	0.0	0.0	0.0	0.0	0.0
21	0.0	0.0	0.0	0.0	0.0
22	0.0	0.0	0.0	0.0	0.0
23	0.0	0.0	0.0	0.0	0.0
24	0.0	0.0	0.0	0.0	0.0
25	0.0	0.0	0.0	0.0	0.0
26	0.0	0.0	0.0	0.0	0.0
27	0.0	0.0	0.0	0.0	0.0
28	0.0	0.0	0.0	0.0	0.0
29	0.0	0.0	0.0	0.0	0.0
30	0.0	0.0	0.0	0.0	0.0
31	0.0	0.0	0.0	0.0	0.0
32	0.0	0.0	0.0	0.0	0.0
33	0.0	0.0	0.0	0.0	0.0
34	0.0	0.0	0.0	0.0	0.0
35	0.0	0.0	0.0	0.0	0.0
36	0.0	0.0	0.0	0.0	0.0
37	0.0	0.0	0.0	0.0	0.0
38	0.0	0.0	0.0	0.0	0.0
39	0.0	0.0	0.0	0.0	0.0
40	0.0	0.0	0.0	0.0	0.0
41	0.0	0.0	0.0	0.0	0.0
42	0.0	0.0	0.0	0.0	0.0
43	0.0	0.0	0.0	0.0	0.0
44	0.0	0.0	0.0	0.0	0.0
45	0.0	0.0	0.0	0.0	0.0
46	0.0	0.0	0.0	0.0	0.0
47	0.0	0.0	0.0	0.0	0.0
48	0.0	0.0	0.0	0.0	0.0
49	0.0	0.0	0.0	0.0	0.0
50	0.0	0.0	0.0	0.0	0.0
51	0.0	0.0	0.0	0.0	0.0
52	0.0	0.0	0.0	0.0	0.0
53	0.0	0.0	0.0	0.0	0.0
54	0.0	0.0	0.0	0.0	0.0
55	0.0	0.0	0.0	0.0	0.0
56	0.0	0.0	0.0	0.0	0.0
57	0.0	0.0	0.0	0.0	0.0
58	0.0	0.0	0.0	0.0	0.0
59	0.0	0.0	0.0	0.0	0.0
60	0.0	0.0	0.0	0.0	0.0
61	0.0	0.0	0.0	0.0	0.0
62	0.0	0.0	0.0	0.0	0.0
63	0.0	0.0	0.0	0.0	0.0
64	0.0	0.0	0.0	0.0	0.0
65	0.0	0.0	0.0	0.0	0.0
66	0.0	0.0	0.0	0.0	0.0
67	0.0	0.0	0.0	0.0	0.0
68	0.0	0.0	0.0	0.0	0.0
69	0.0	0.0	0.0	0.0	0.0
70	0.0	0.0	0.0	0.0	0.0
71	0.0	0.0	0.0	0.0	0.0
72	0.0	0.0	0.0	0.0	0.0
73	0.0	0.0	0.0	0.0	0.0
74	0.0	0.0	0.0	0.0	0.0
75	0.0	0.0	0.0	0.0	0.0
76	0.0	0.0	0.0	0.0	0.0
77	0.0	0.0	0.0	0.0	0.0
78	0.0	0.0	0.0	0.0	0.0
79	0.0	0.0	0.0	0.0	0.0
80	0.0	0.0	0.0	0.0	0.0
81	0.0	0.0	0.0	0.0	0.0
82	0.0	0.0	0.0	0.0	0.0
83	0.0	0.0	0.0	0.0	0.0
84	0.0	0.0	0.0	0.0	0.0
85	0.0	0.0	0.0	0.0	0.0
86	0.0	0.0	0.0	0.0	0.0
87	0.0	0.0	0.0	0.0	0.0
88	0.0	0.0	0.0	0.0	0.0
89	0.0	0.0	0.0	0.0	0.0
90	0.0	0.0	0.0	0.0	0.0
91	0.0	0.0	0.0	0.0	0.0
92	0.0	0.0	0.0	0.0	0.0
93	0.0	0.0	0.0	0.0	0.0
94	0.0	0.0	0.0	0.0	0.0
95	0.0	0.0	0.0	0.0	0.0
96	0.0	0.0	0.0	0.0	0.0
97	0.0	0.0	0.0	0.0	0.0
98	0.0	0.0	0.0	0.0	0.0
99	0.0	0.0	0.0	0.0	0.0
100	0.0	0.0	0.0	0.0	0.0

Datenschutzrisiken von Künstlicher Intelligenz



DSGVO definiert drei Datenschutzrisiken:

- **Linkability:** Verknüpfen von Datenpunkten
- **Singling Out:** Identifizieren aller zu einer Person gehörenden Daten
- **Inference:** Ableiten von unbekanntem/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
- Verarbeitung 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

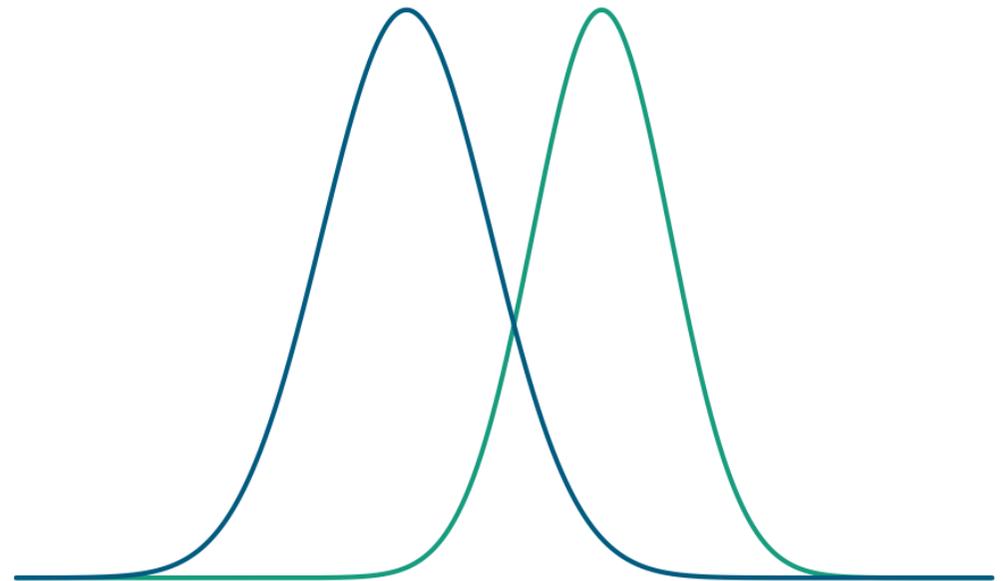
Unterstützung von Entscheidungsträgern

- Analyse von komplexen Daten
- Aufbereitung von Daten



Quantifizierbare Privatsphäregarantien mit ϵ -Differential Privacy

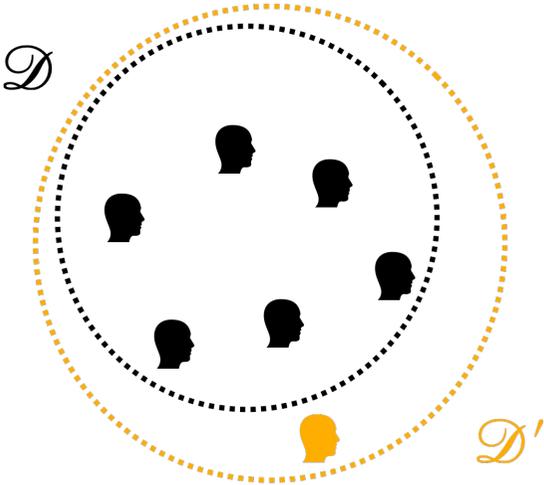
Mathematische Grundlage für Privatsphäre



ϵ Differential Privacy

ϵ Differential Privacy

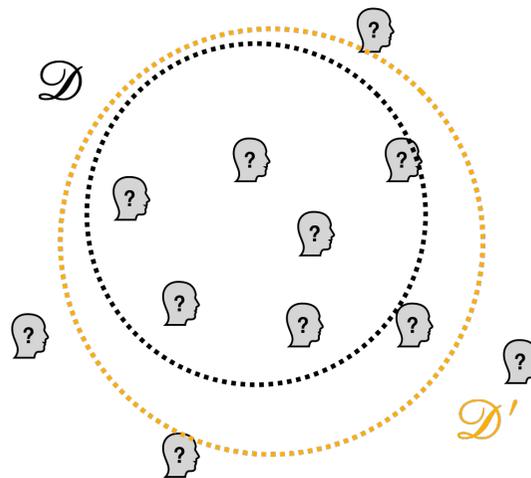
ϵ Differential Privacy



ϵ Differential Privacy

- ϵ : Stärke der Privatsphäre

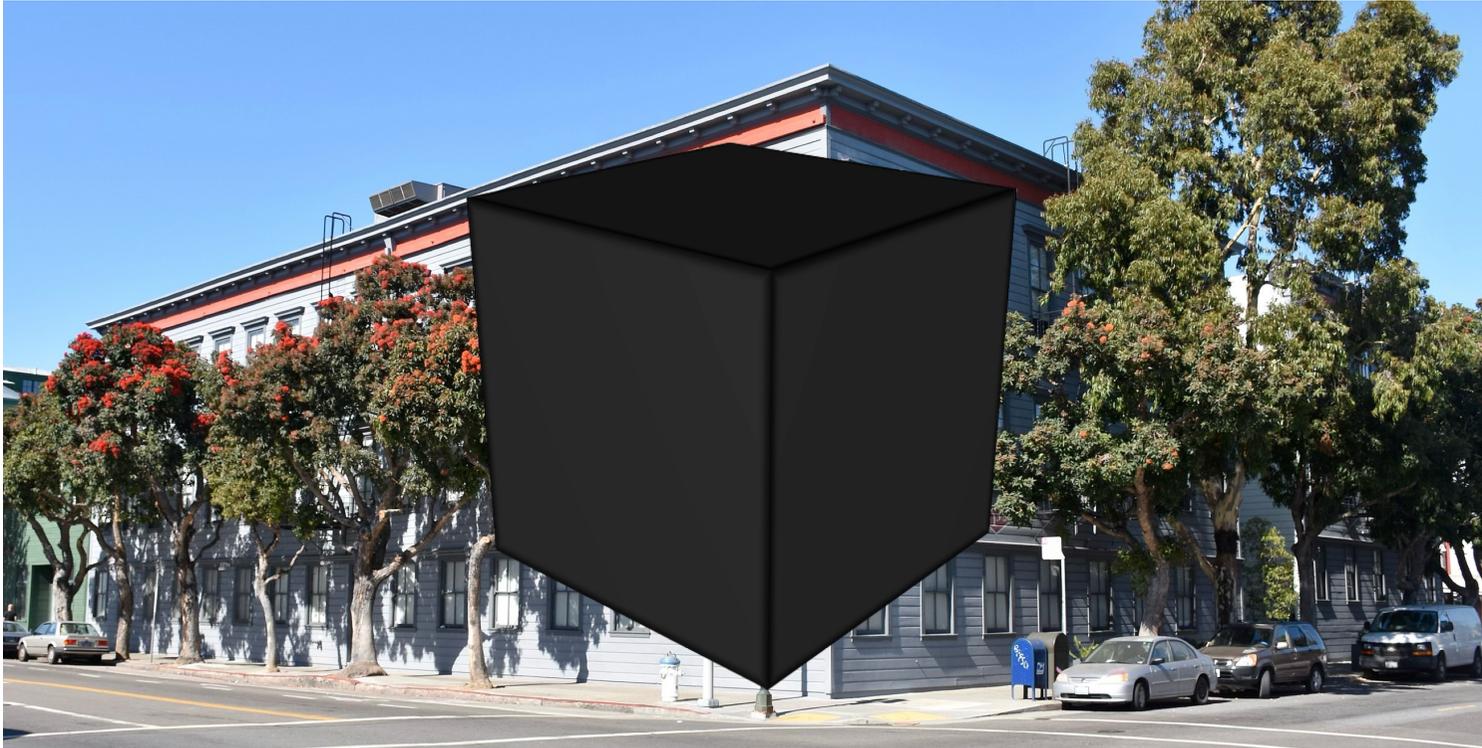
$$\Pr(\mathcal{A}(D) \in S) \leq e^\epsilon \cdot \Pr(\mathcal{A}(D') \in S)$$



- **Mathematische
Privatsphäregarantie**

- 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. Copyright 2024 by the author(s).

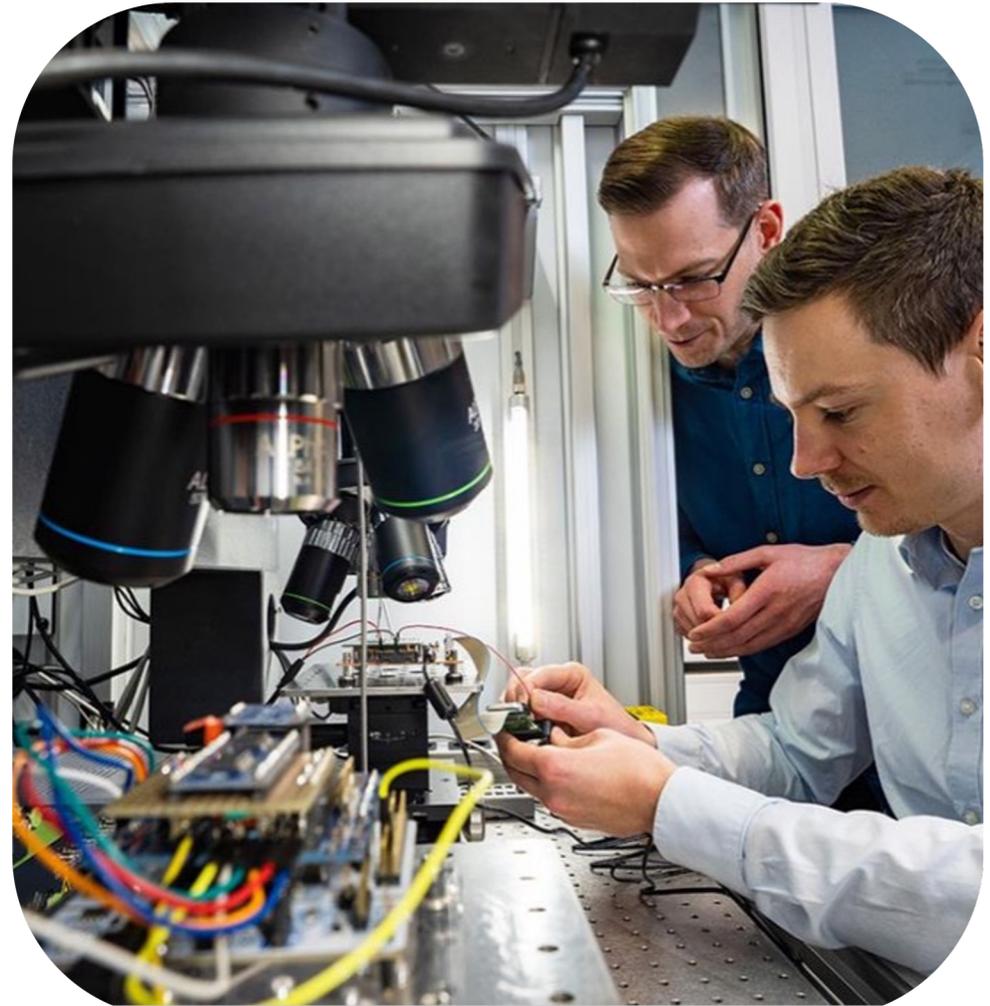
Open LLMs are Necessary for Private Adaptations and Outperform their Closed Alternatives

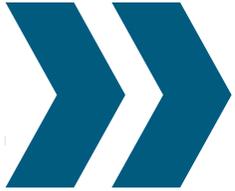
Vincent Hanke¹ Tom Blanchard¹ Franziska Boenisch¹
Iyiola E. Olatunji¹ Michael Backes¹ Adam Dziedzic¹

¹CISPA 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 privacy-preserving LLM adaptations that yield high performance and more privacy at lower costs, one should use open LLMs.

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
