An IT-Security view on privacy-preserving, LLM-based systems with provider exclusion

Telekom MMS

CLOUD & HEAT

Ivan Gudymenko, IT Security Architect

Yewgenij Baburkin, DevOps Engineer

Who we are and what we do

- IT security architect
- Research and Development
- Security Consulting
- Approval process (Zulassung)
- Cloud service provider

WIR SIND
Telekom MMS

Who we are and what we do

- IT-Security Researcher
- DevOps Engineer
- Cloud Provider
- OpenStack integration

Agenda

- LLM-based Systems and Privacy Issues
- Use Case: AI-as-a-Service
- AI and Confidential Computing
- Takeaways

Stanford University
Human-Centered
Artificial Intelligence

About Research Education Policy AI Index

NEWS

Be Careful What You Tell Your AI Chatbot

DATE OCTOBER 15, 2025

TOPICS PRIVACY, SAFETY, SECURITY | GENERATIVE AI | REGULATION, POLICY, GOVERNANCE

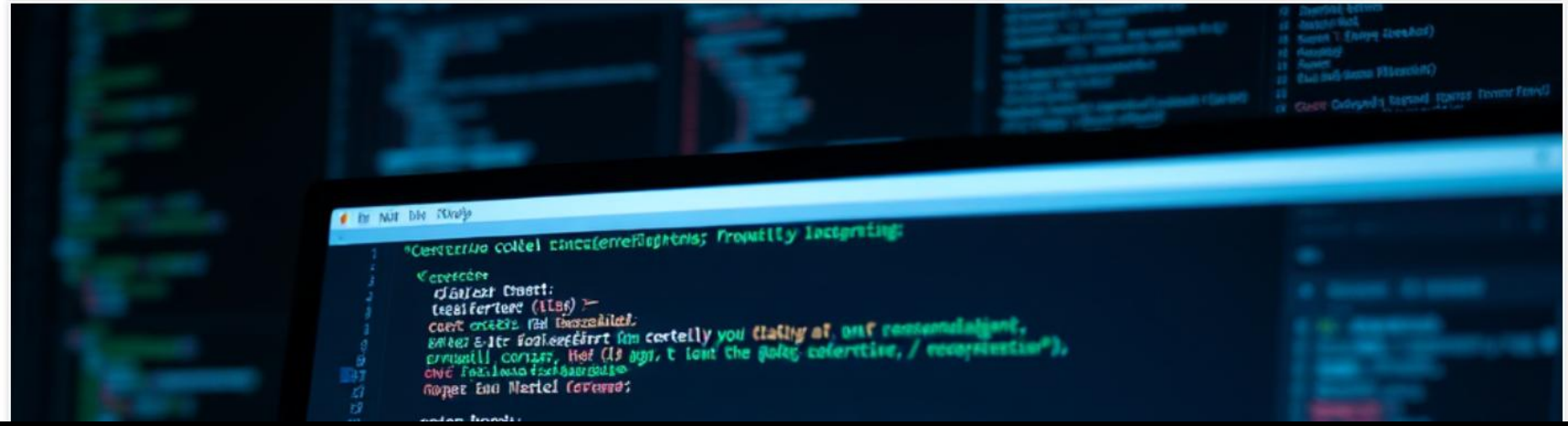
Google is indexing shared ChatGPT links – exposing prompts, responses, and sensitive info you may have unknowingly made public.

Google is indexing public ChatGPT shared links for all but ChatGPT Enterprise customers. This means your AI-generated conversations could be indexed and discoverable via Google Search.

HOME › CYBERSECURITY BLOG › ARTIFICIAL INTELLIGENCE › PROMPT LEAKAGE VIA AUTO-SAVE, LOGGING, AND CHAT HISTORY

Prompt Leakage via Auto-Save, Logging, and Chat History

BY MRJVVXXM *on* JULY 9, 2025 · (0)



A Stanford Study: User Privacy and LLMs

Data Privacy Practice	Amazon	Anthropic	Google	Meta	Microsoft	OpenAI
Chat input used for training by default	Not Specified	Yes	Yes	Yes	Yes	Yes
Mechanism to opt out of chat training	Not Specified	Yes	Yes	Not Specified	Yes	Yes
Chat data retained indefinitely	Yes	No	No	Yes	No	Yes
Chatbot personalization features	Not Specified	Not Specified	Yes	Yes	Yes	Yes
Allows accounts for children 13-18	No	No	Yes	Yes	Yes	Yes

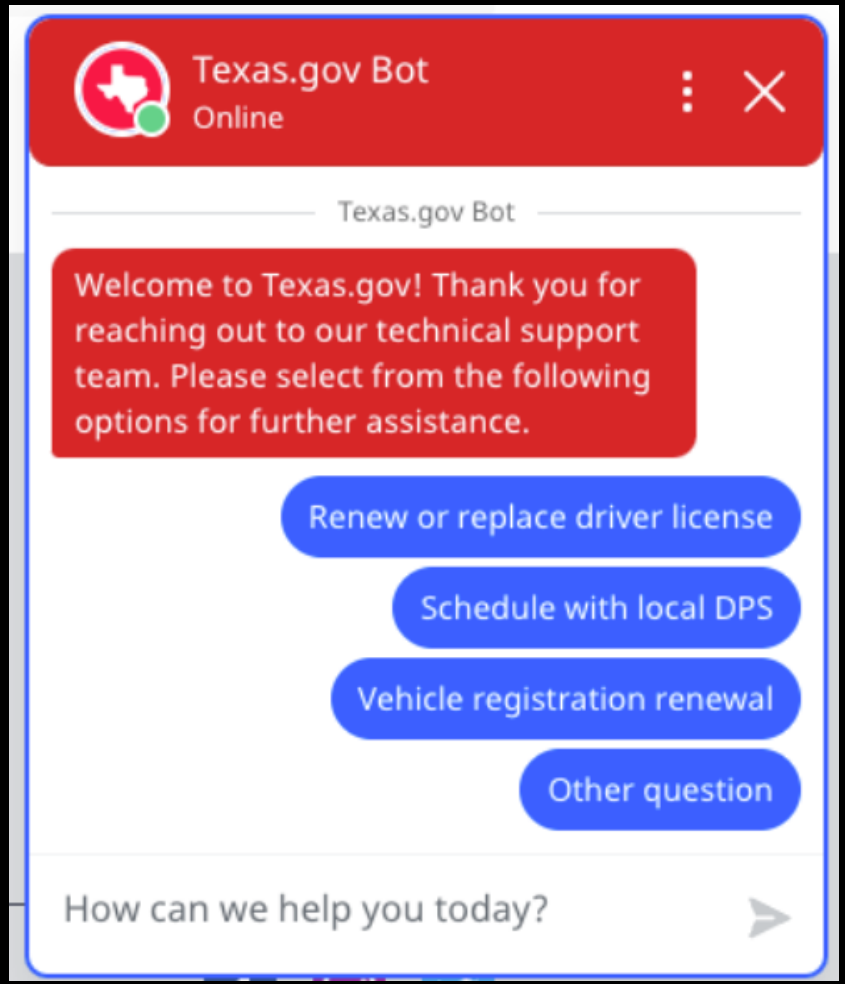
- Are user inputs to chatbots used to train or improve LLMs?
- What sources and categories of personal consumer data are collected, stored, and processed to train or improve LLMs?
- What are the users' options for opting into or out of having their chats used for training?

King, Jennifer, et al. "User privacy and large language models: An analysis of frontier developers' privacy policies." Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society. Vol. 8. No. 2. 2025

Example Use Case: chat bot for public authorities

Sample interaction flow

- Input: a citizen starts conversation
- AI action: guiding the citizen through the process (e.g. fill in some form)
- Output: the process is finalized and the citizen is informed of the current status



https://statescoop.com/government-ai-chatbots-state-local-websites-2024/

Challenge: building trustworthy LLM-based systems

How to ensure citizens' privacy and technically enforce it against (1) operator and (2) provider?

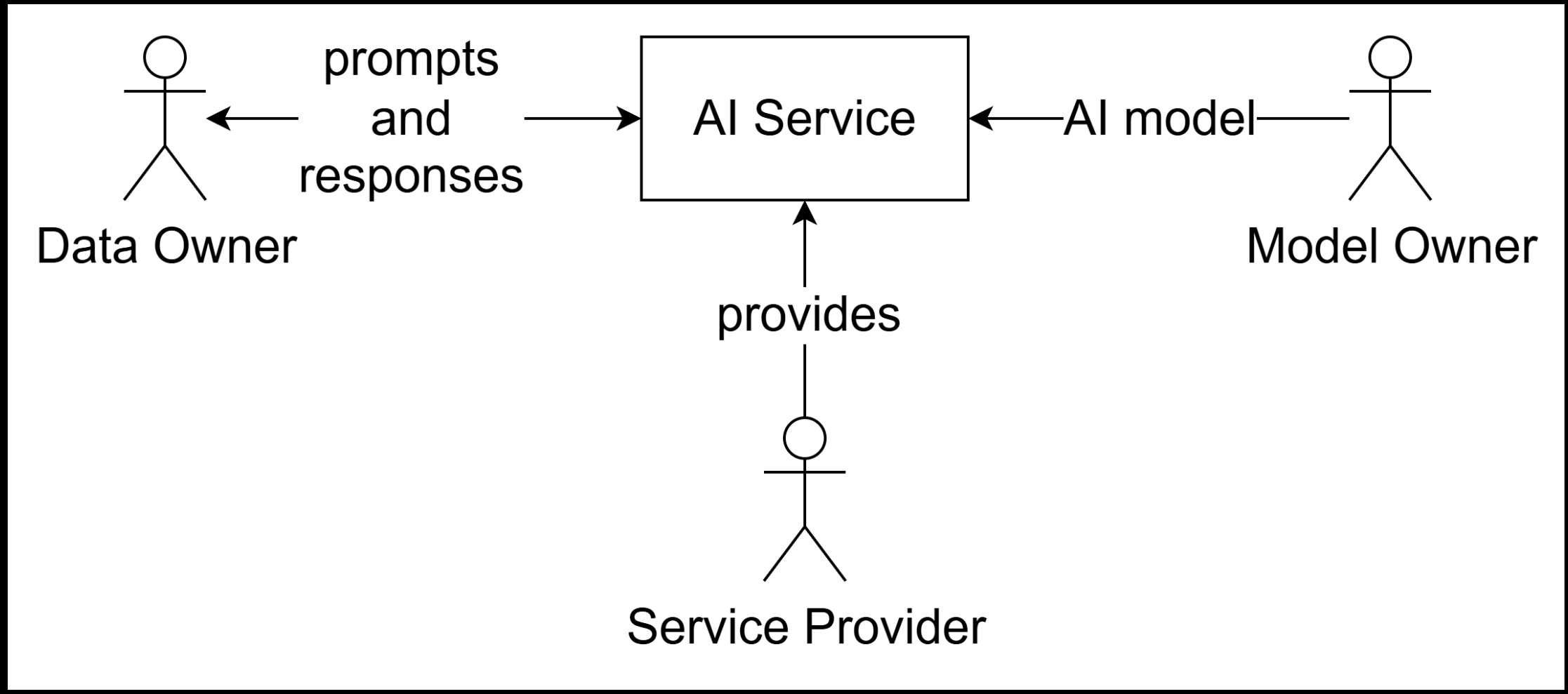
Trustworthy vs Trusted

- After signing off a certain cloud strategy, a cloud provider is *implicitly* trusted (e.g. a cloud-based application)
- It is usually not possible to enforce privacy and security requirements purely by technical means
- Trustworthiness implies further technical controls to ensure that privacy and security requirements are enforced, incl. at the very moment of execution

Questions to be addressed

- How to ensure that the LLM-based application is trustworthy?
- How can the provider and operator be technically excluded from compromizing user privacy? (incl. inadvertent abuses)
- How can the provider and operator be technically excluded from extracting model weights (intellectual property risks)

AI-as-a-Service at a High-level

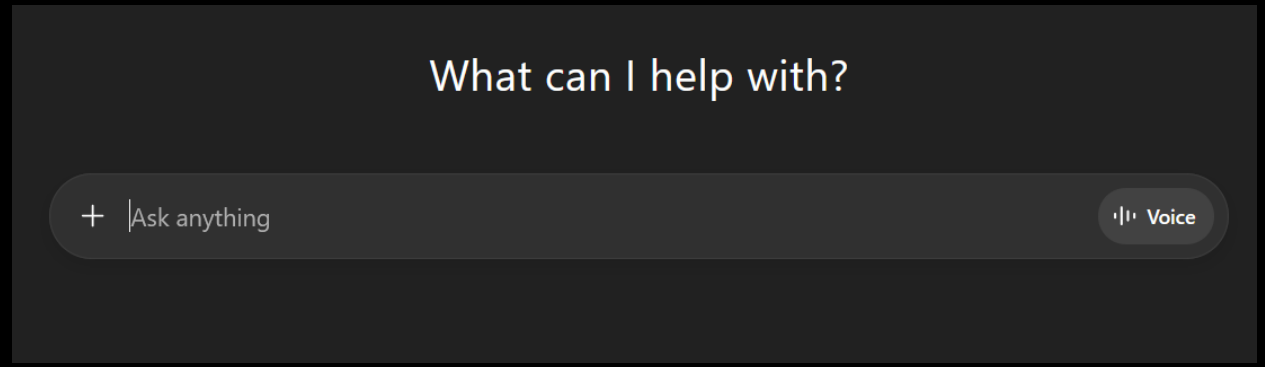


Threat Modeling is Essential

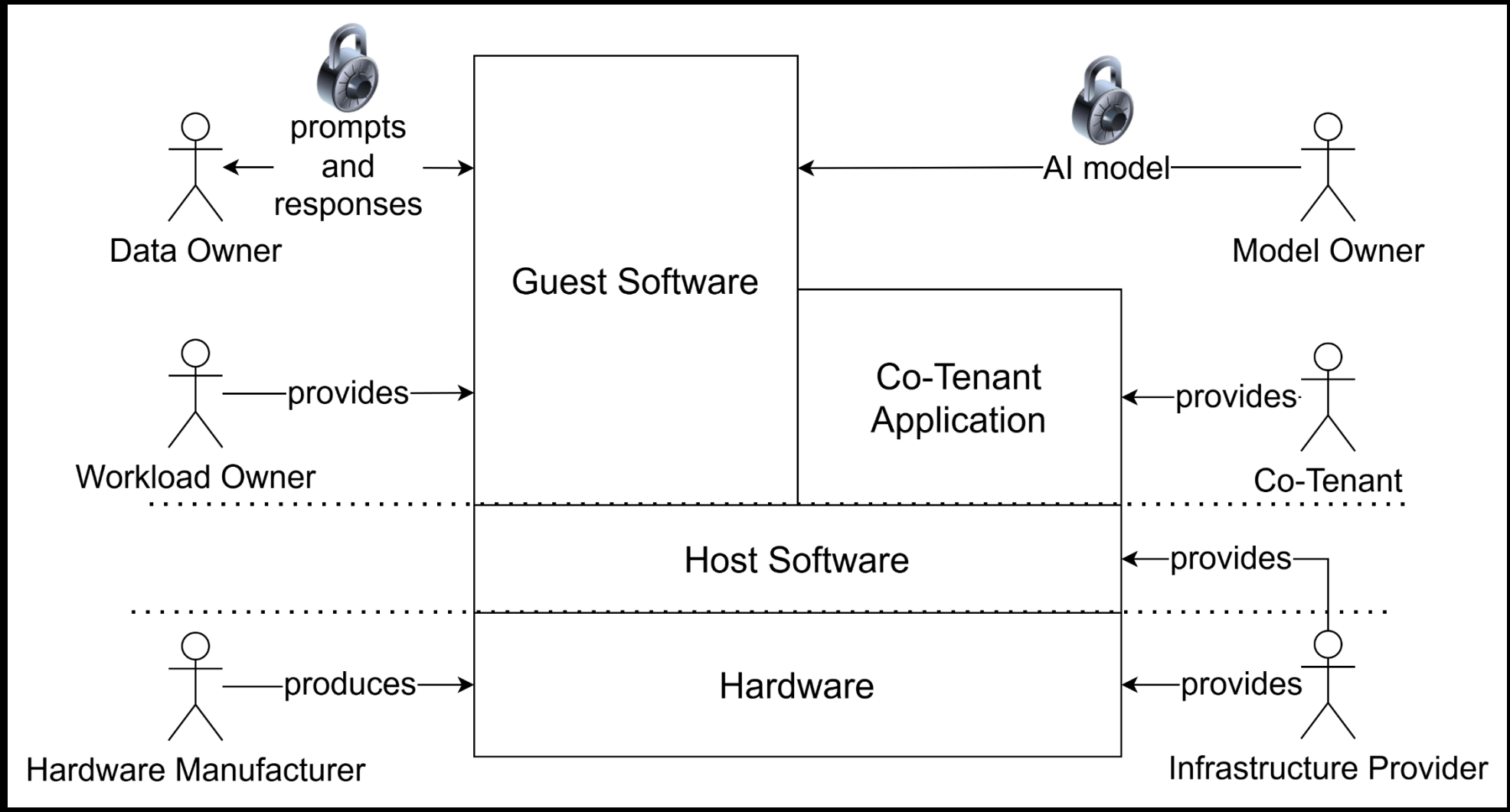
- Data Flow Diagram: what is being developed
- What are the critical assets?
- Security Threats: What can go wrong
- Countermeasures: What should we do
- Implementation
- Security concept

Data security as an example protection goal

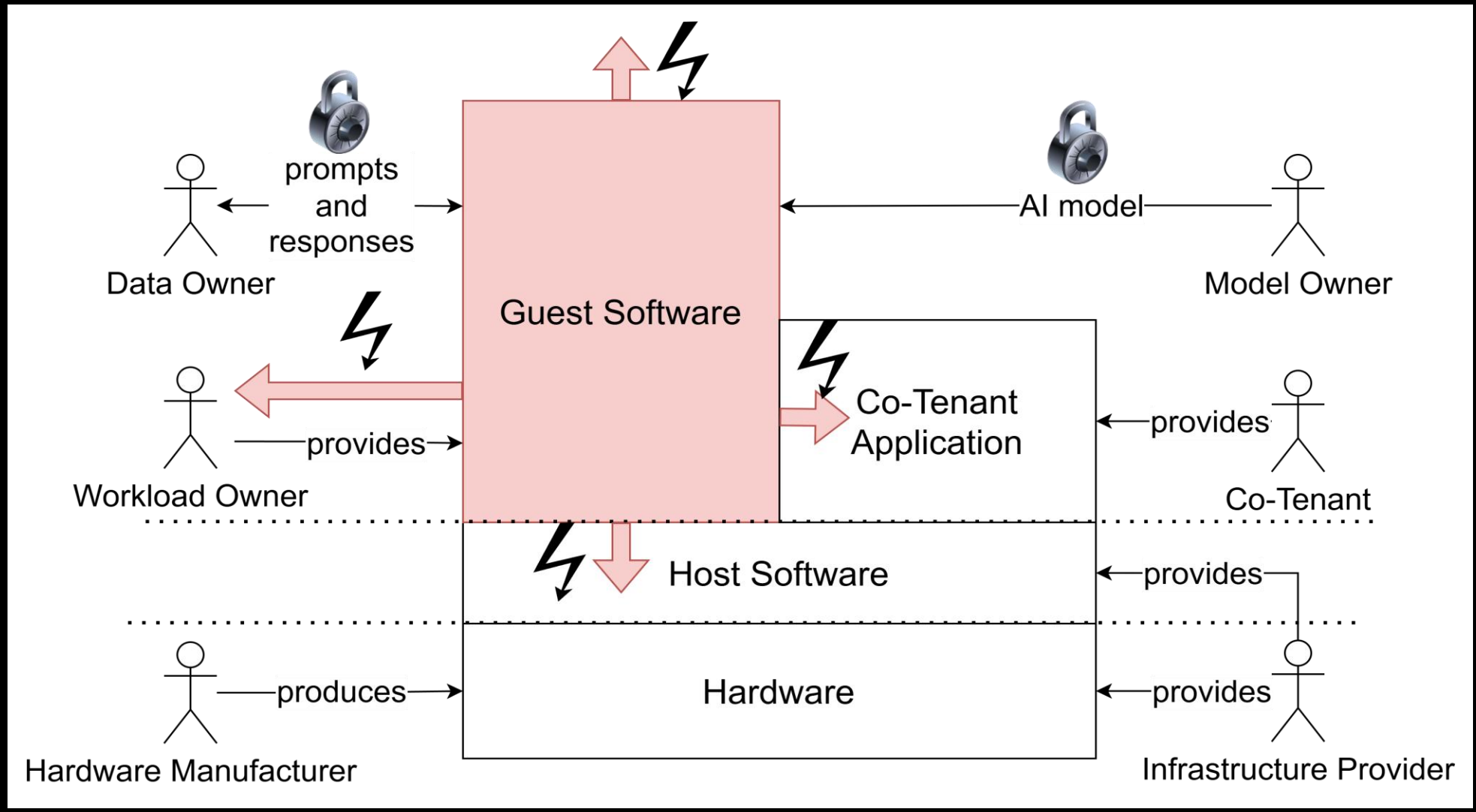
- In real-world systems, multiple protection goals are addressed
- In this talk, we focus data security in terms of prompts content
- and model weights confidentiality



Threat Actors

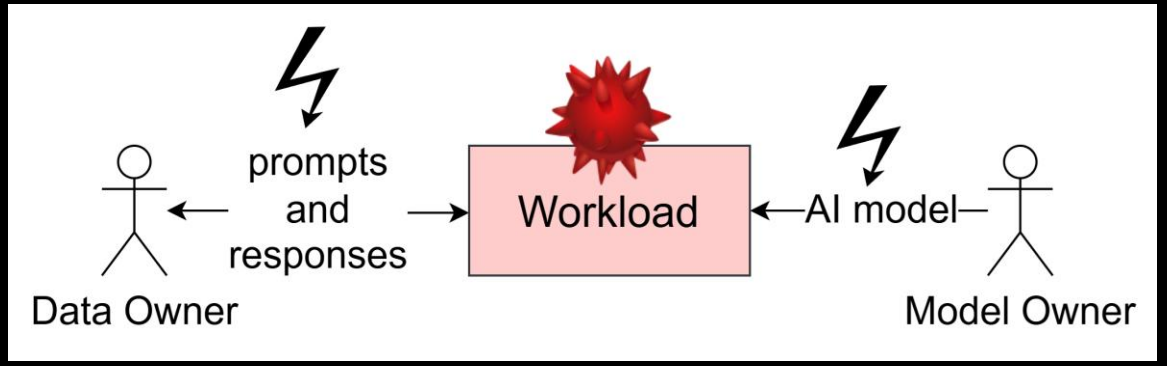


Threat Model



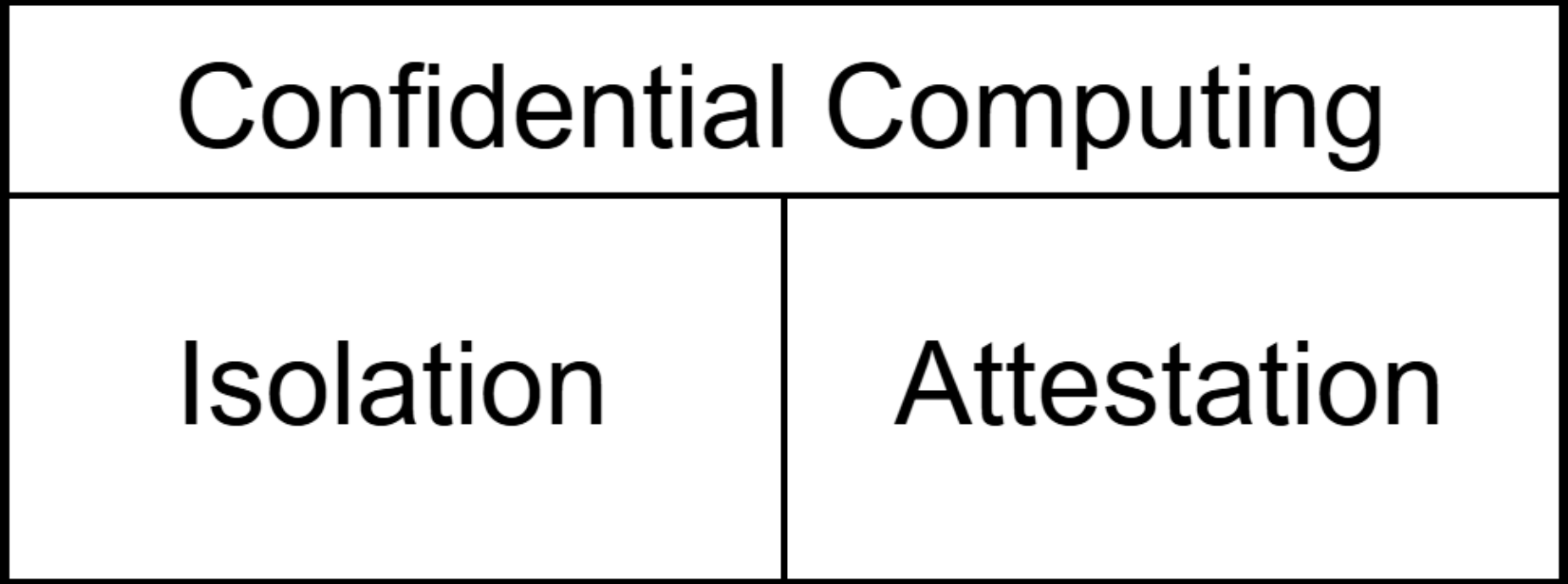
Workload Substitution

- Through Workload Owner or other Actors
- Data Owner may disclose prompts and responses
- Model Owner may disclose AI model

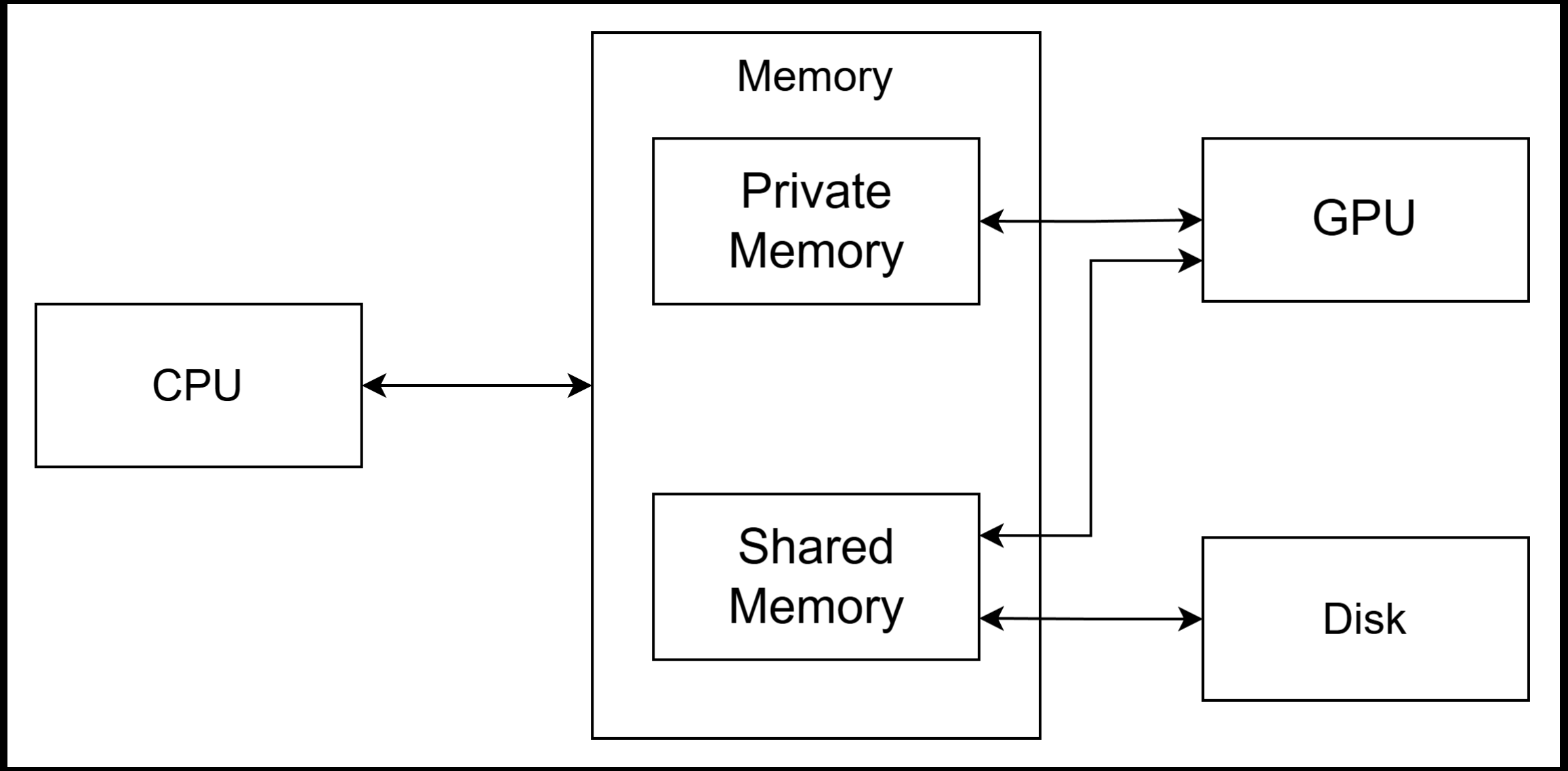


Can Confidenial Computing help in technically enforcing data privacy?

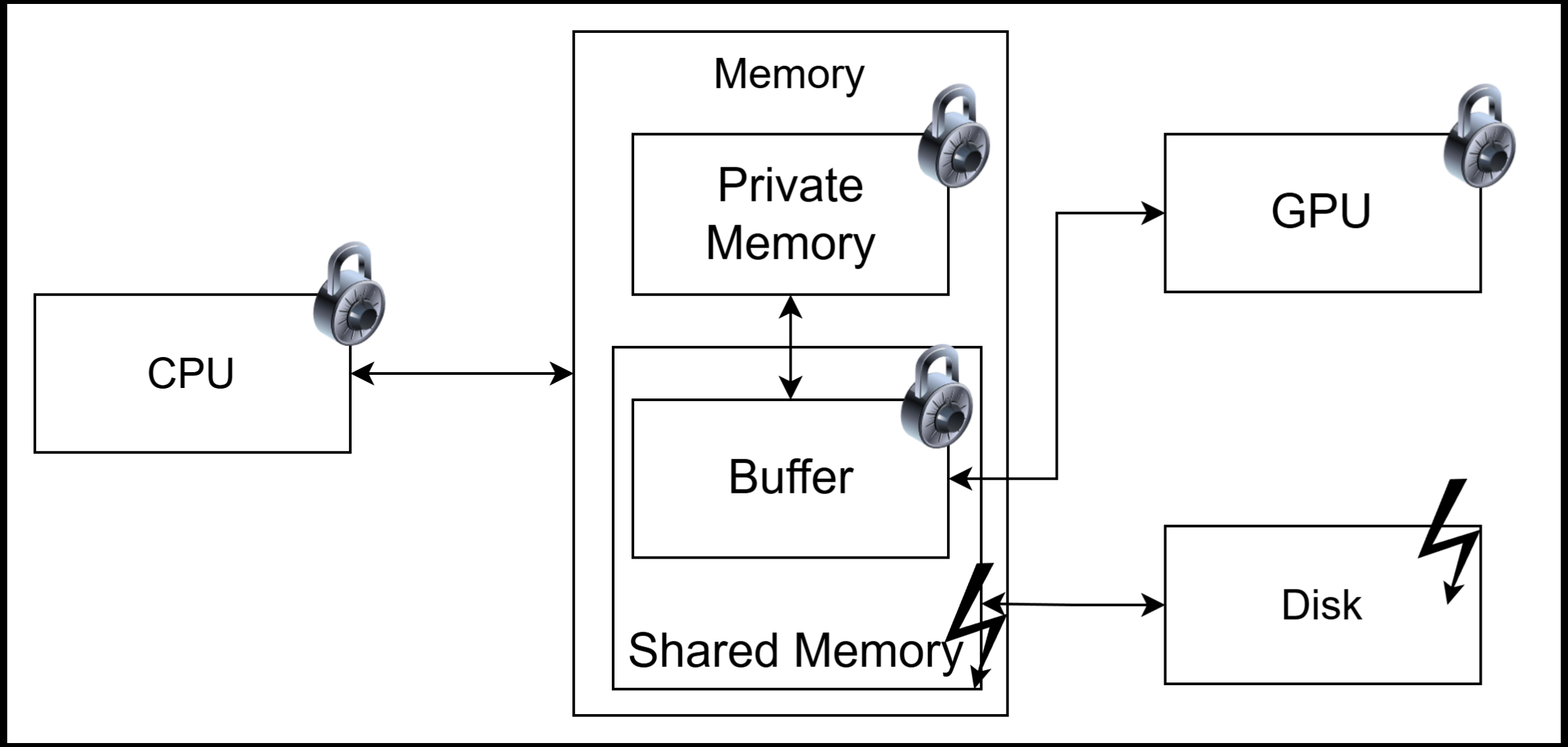
Isolation and Attestation



Isolation: Memory Mapping



Isolation: Memory Mapping, CPU+GPU CC, Buffer

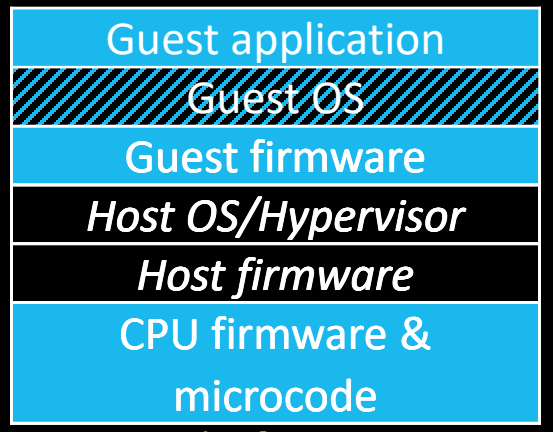


Attestation

- Proof of isolation (TEE usage)
- Proof of Workload identity
- Needs to be verified
- Secret injection/service consumption on success

Attestation Levels

- Levels build upon each other
- Goal is to achieve Level 4
- Out-of-the-box experience only goes up to Level 2
- CSP adoption varies



Boot Hierarchy of a system running a CVM. Components in italics signify exclusion from the TCB.

Hierarchy of CVM attestation levels, Scopelliti et al., 2024

CSP adoption

- CSPs‘ CC definition and implementation differs

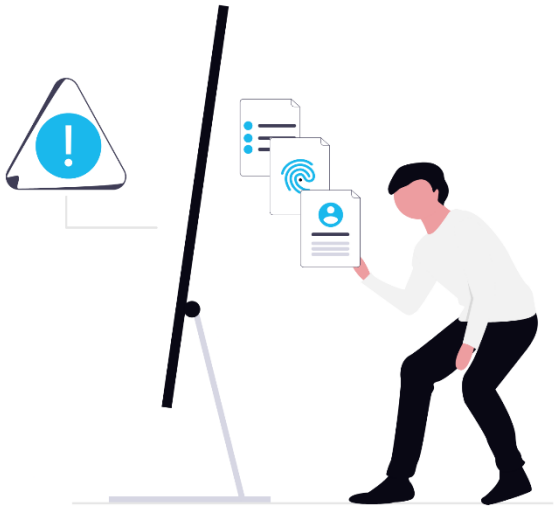
	AWS	Azure	GCP
TEE	●	◐	●
Firmware	●	◐	◐
Kernel	◐	◐	◐
Root FS	◐	◐	◐
Nominal AL	**4**	**4**	**4**
Trustworthy AL	**2**	**0**	**1**

● = Verifiable w/o trust; ◐ = Verifiable w/ trust; ○ = Not verifiable

Scopelliti et al., 2024,
https://doi.org/10.1109/EuroSPW61312.2024.00023

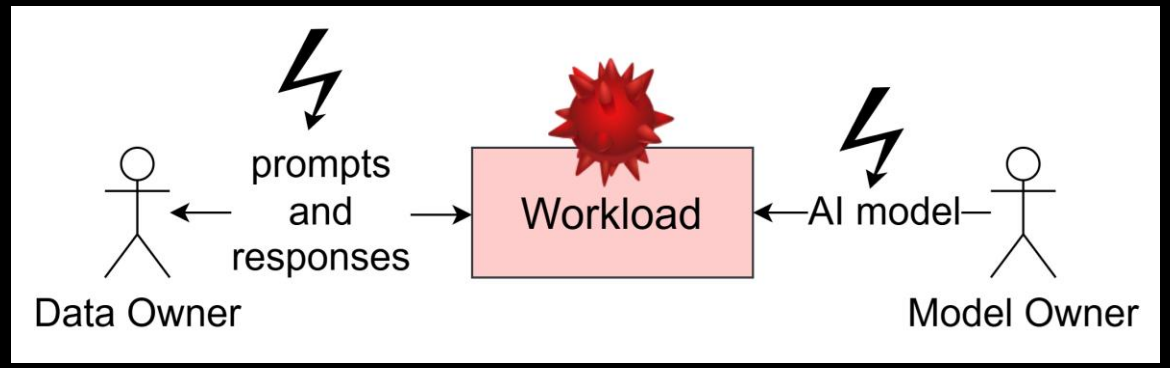
Threat Model: Confidential Computing

- Mitigation status varies across technologies
- Examples:
  - Mitigated: TEE Memory Extraction, DRAM Extraction
  - Partially Mitigated: Workload Substitution, Software Side-Channel Attacks
  - Not Mitigated: Fault Injection, Physical Side-Channel Attacks



Callback: Workload Substitution

- Through Workload Owner or other Actors
- Data Owner may disclose prompts and reponses
- Model Owner may disclose AI model
- **<u>Attestation Level 4 is needed for mitigation</u>**



Things to Keep In Mind

- Threat models must align
- Choosing the right CC technology
- CSP adoption
- Attestation and its consumption

Summary

- LLM-based systems pose inherent risks if used in uncontrolled manner
- Threat modelling is essential
- Achieving complete attestation (AL4)
- Bootstrapping persistent storage
- Confidential Computing as Defence in Depth

Thank you very much for your attention!

Ivan Gudymenko, IT Security Architect, ivan.gudymenko@telekom.de

Yewgenij Baburkin, DevOps Engineer, yewgenij.baburkin@cloudandheat.com

Telekom MMS

CLOUD & HEAT