KI ohne Abhängigkeit:

Wie wir Open-Source-Modelle in der Lehre einsetzen

ludwig.david.lorenz@uni-weimar.de Fachstelle Medientechnologie

Public Domain Presentation

25% +7 ↑

ChatGPT

7%

Microsoft Copilot

6%

Google Gemini
(ehemals Bard)

genutzte KI-Anwendungen in der Bevölkerung ab 14 Jahren. Quelle: D21 Digital-Index 2024/25

HINTERGRUNDBILD https://stadt-bremerhaven.de/wp-content/uploads/2024/05/recall2.jpg

Microsoft Recall (2025)

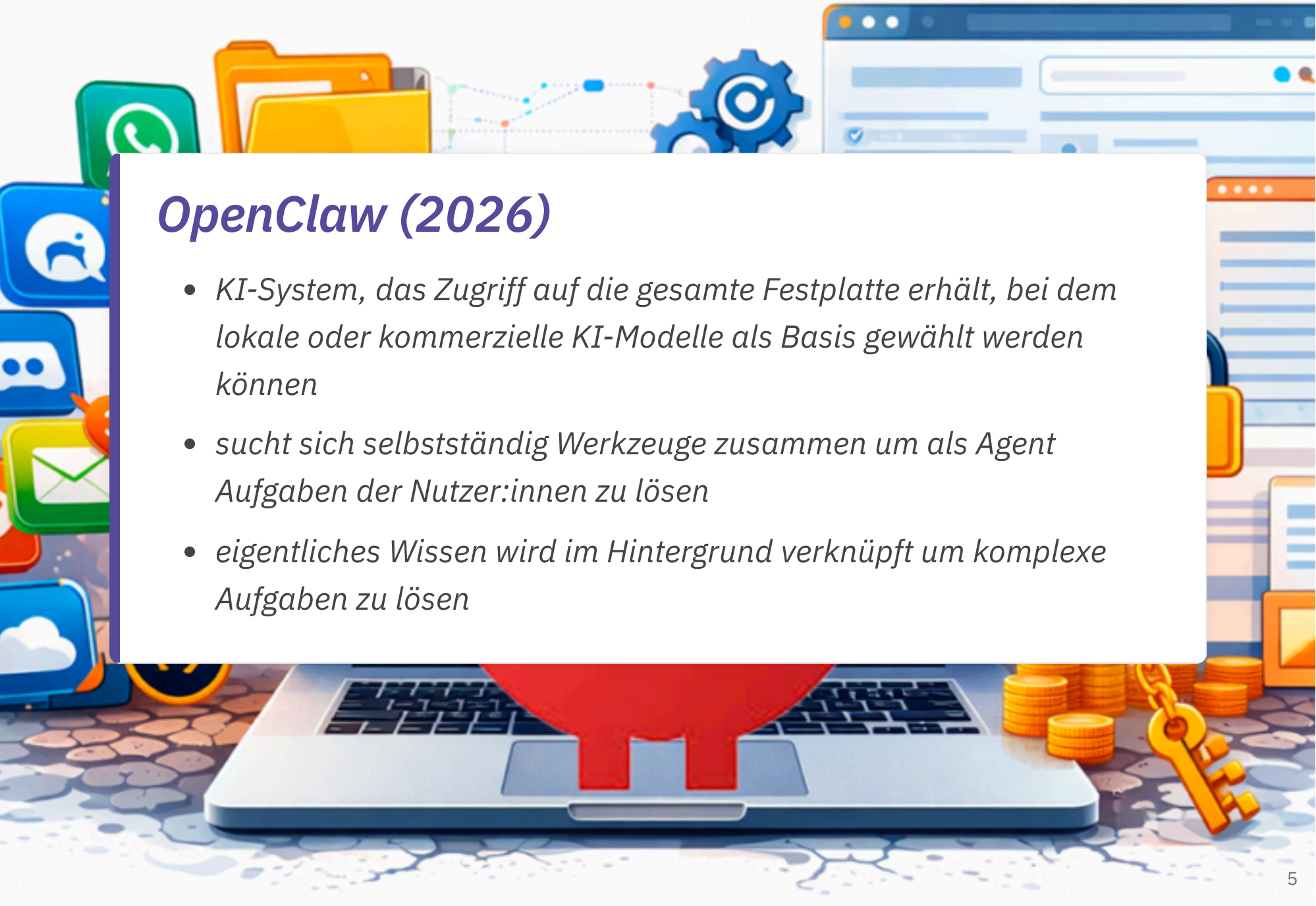
- *umstrittene KI-Funktion in Windows 11, die verschlüsselte Screenshots Ihres Bildschirms speichert, um sie später wieder aufrufen zu können*
- *gescheiterter Versuch, passives PKM mithilfe von KI flächendeckend einzuführen.*

HINTERGRUNDBILD https://stadt-bremerhaven.de/wp-content/uploads/2024/05/recall2.jpg

Microsoft Recall (2025)

- *umstrittene KI-Funktion in Windows 11, die verschlüsselte Screenshots Ihres Bildschirms speichert, um sie später wieder aufrufen zu können*
- *gescheiterter Versuch, passives PKM mithilfe von KI flächendeckend einzuführen.*

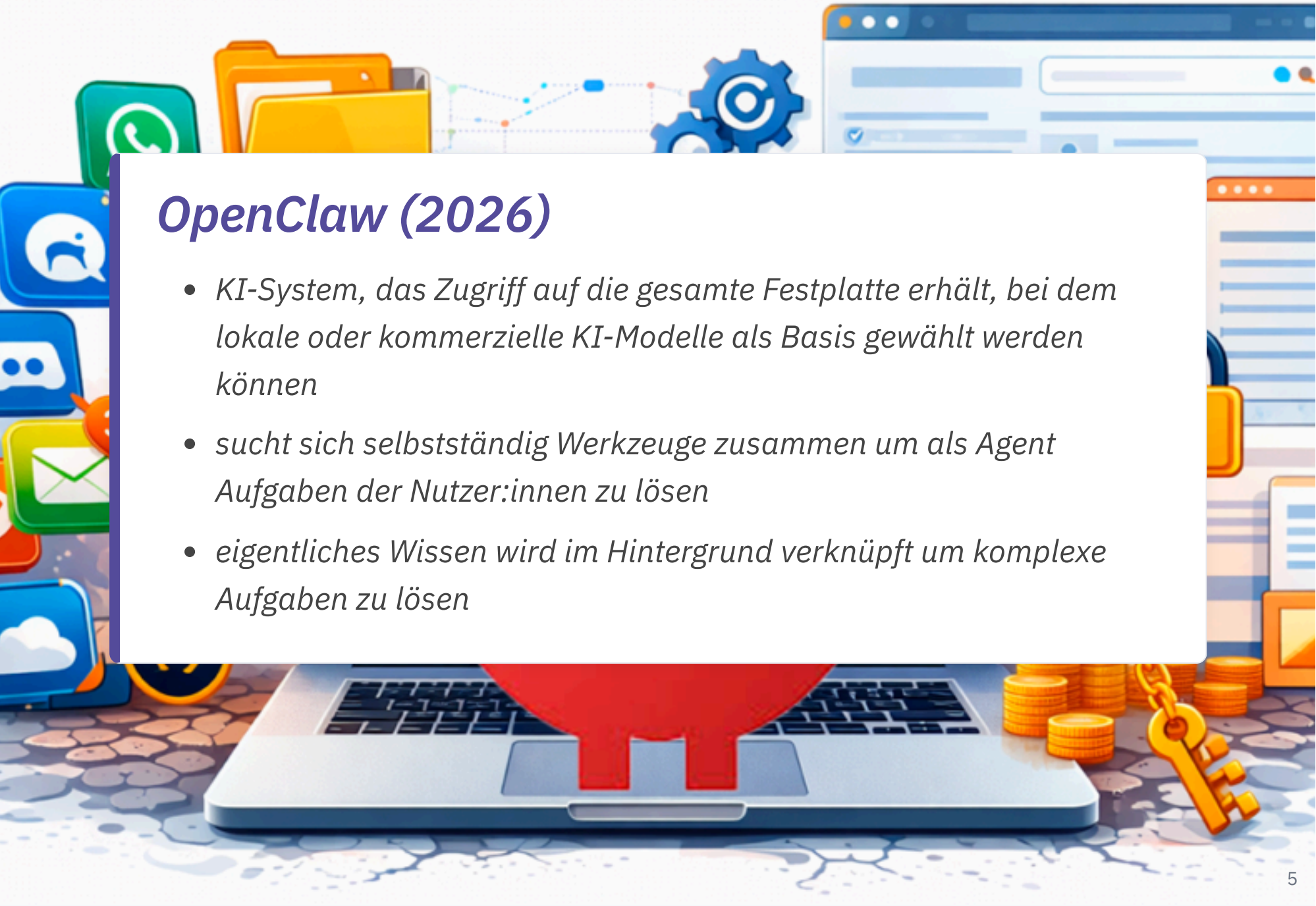
HINTERGRUNDBILD https://tecnobits.com/wp-content/uploads/2026/01/CLAWDBOT-1024x576.png



OpenClaw (2026)

- *KI-System, das Zugriff auf die gesamte Festplatte erhält, bei dem lokale oder kommerzielle KI-Modelle als Basis gewählt werden können*
- *sucht sich selbstständig Werkzeuge zusammen um als Agent Aufgaben der Nutzer:innen zu lösen*
- *eigentliches Wissen wird im Hintergrund verknüpft um komplexe Aufgaben zu lösen*

HINTERGRUNDBILD https://tecnobits.com/wp-content/uploads/2026/01/CLAWDBOT-1024x576.png

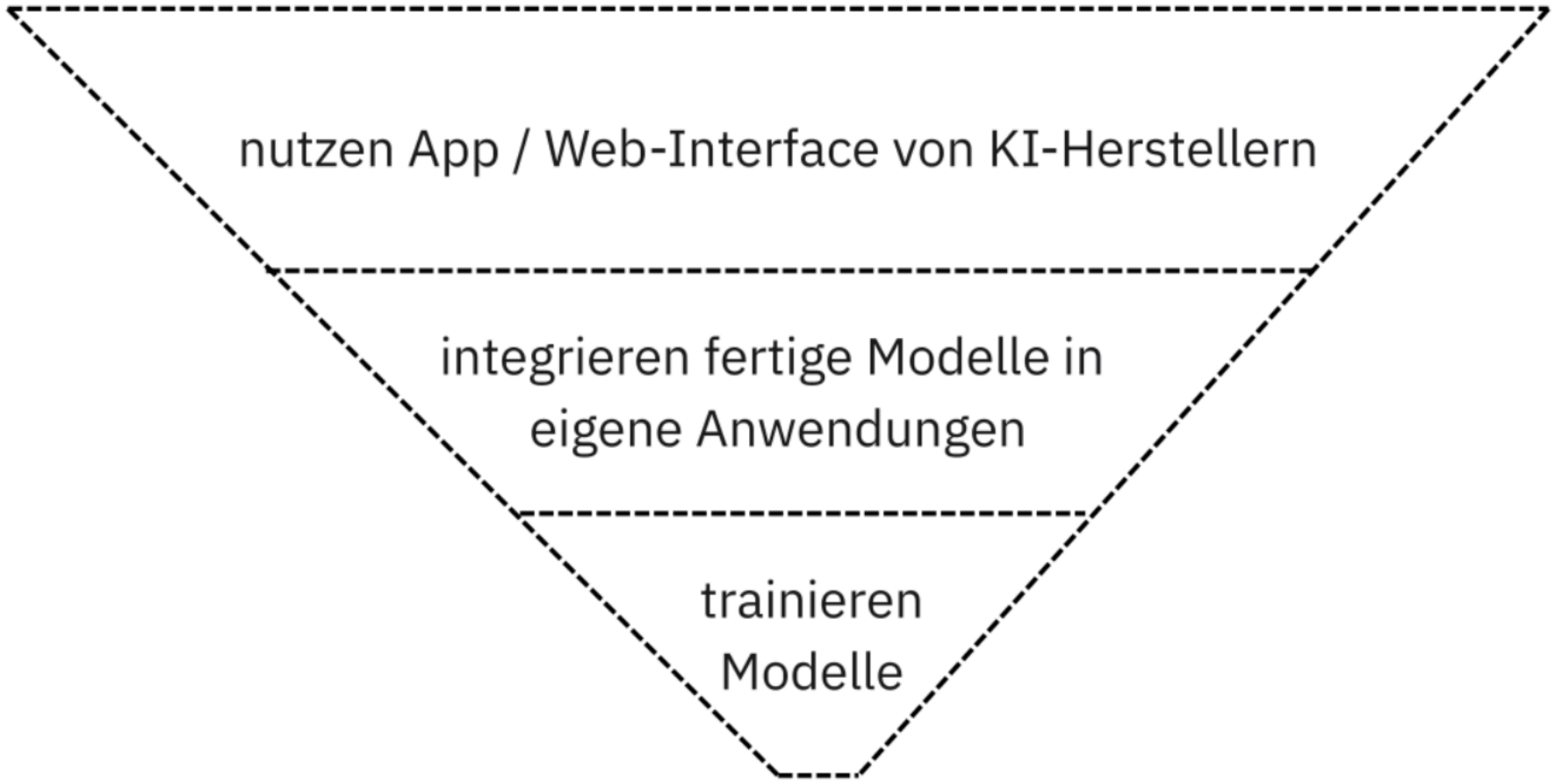


OpenClaw (2026)

- *KI-System, das Zugriff auf die gesamte Festplatte erhält, bei dem lokale oder kommerzielle KI-Modelle als Basis gewählt werden können*
- *sucht sich selbstständig Werkzeuge zusammen um als Agent Aufgaben der Nutzer:innen zu lösen*
- *eigentliches Wissen wird im Hintergrund verknüpft um komplexe Aufgaben zu lösen*

Erkenntnis

Die Zukunft der KI-Modelle basiert auf persönlichen Daten.



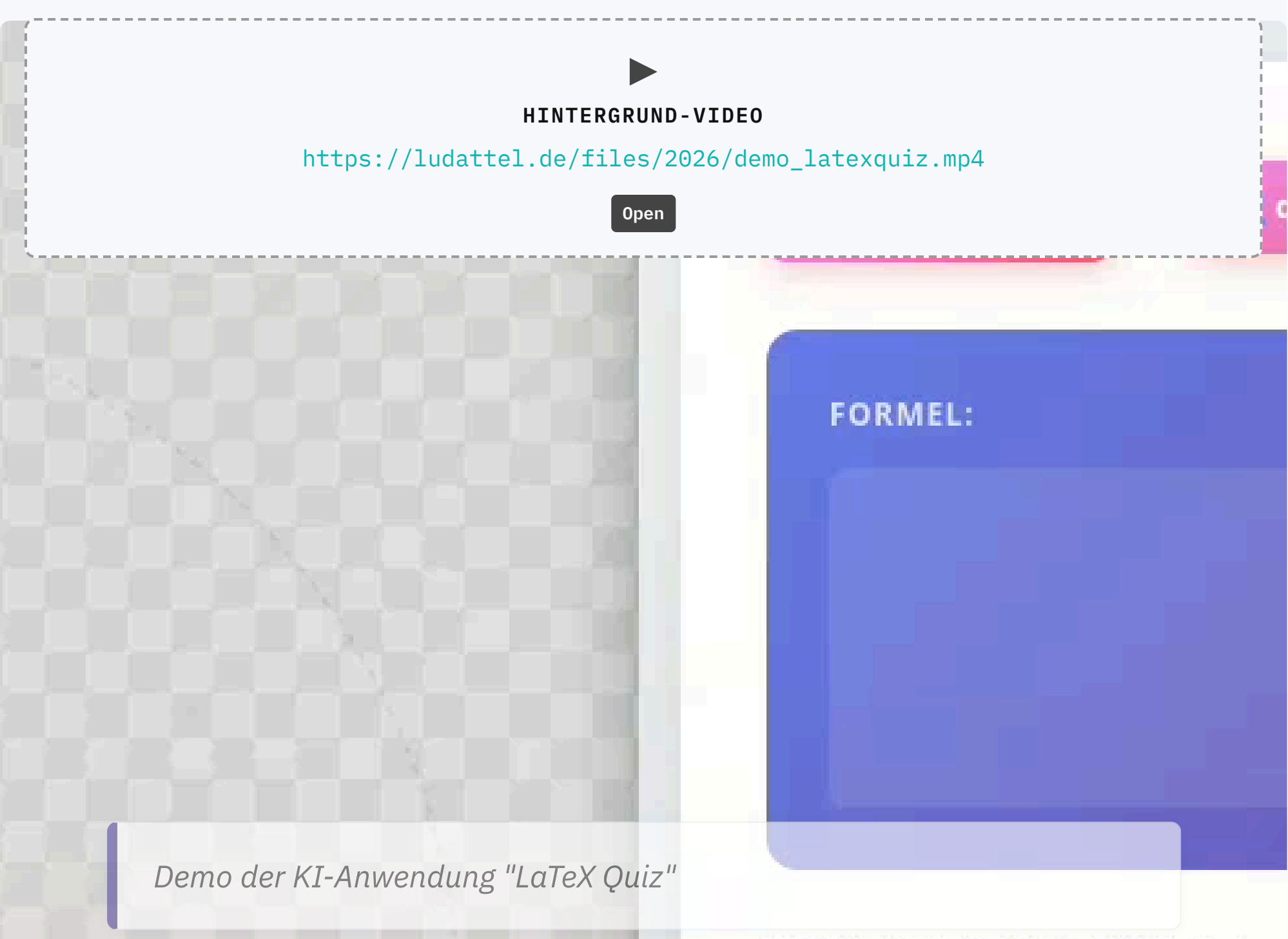
Anwendungsgruppen bei Künstlicher Intelligenz

▶

HINTERGRUND-VIDEO

https://ludattel.de/files/2026/demo_latexquiz.mp4

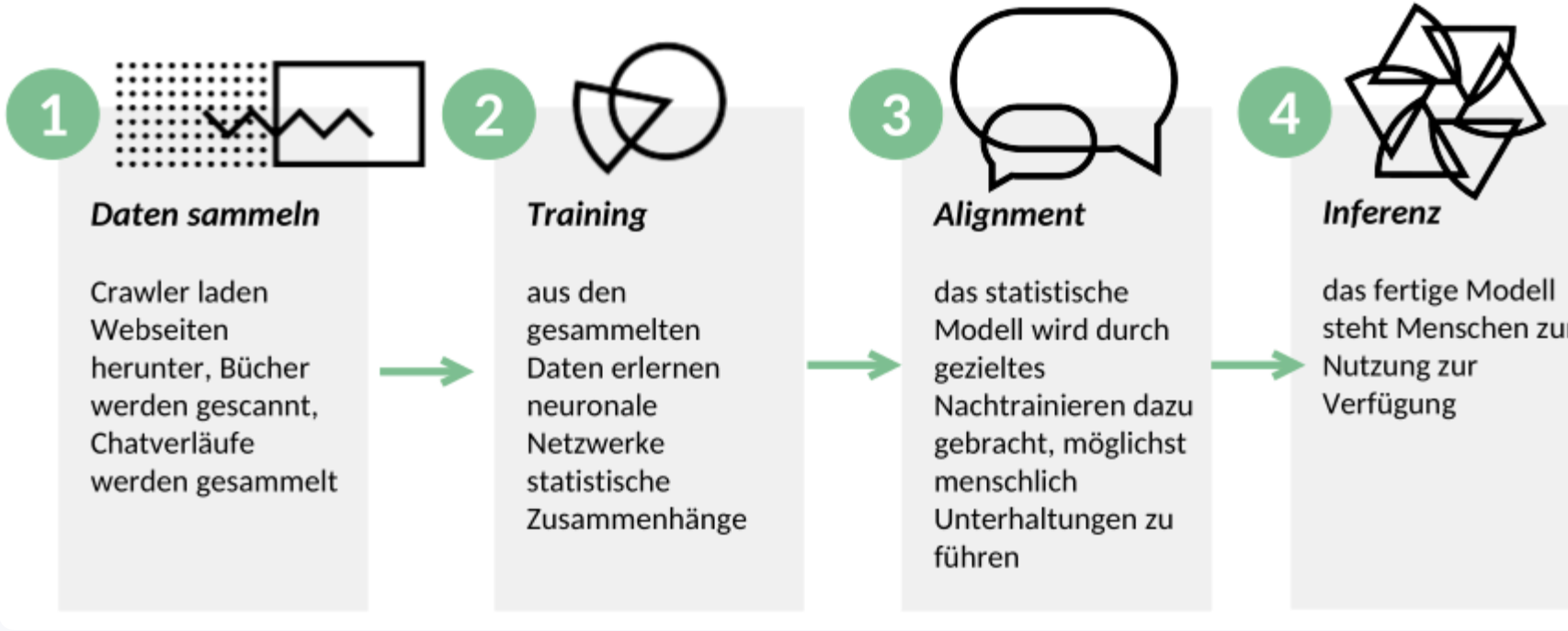
Open



Demo der KI-Anwendung "LaTeX Quiz"

Recap: Was ist ein Modell?

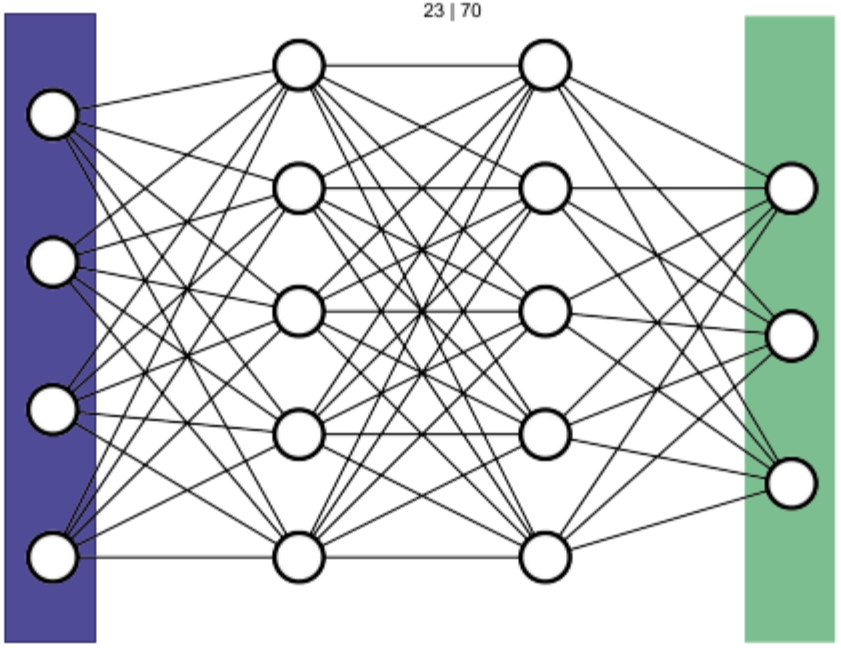
Lebensphasen eines Sprachmodells



aus eTeach-Netzwerk Thüringen • KI Grundlagen • Ludwig Lorenz

Eingabeschicht

Erhält alle relevanten Informationen (Kontext) als Liste von Zahlen (Aktivierung)



Ausgabeschicht

Sagt für jedes mögliche Ergebnis vorher, wie gut es zum Kontext passt

Versteckte Schichten

Leitet die Aktivierung von Schicht zu Schicht. Jedes Neuron darf wie bei einer **Mehrheitswahl** seine Stimmen auf die Neuronen in der nächsten Schicht verteilen.

aus eTeach-Netzwerk Thüringen • KI Grundlagen • Ludwig Lorenz

Schichten

- sind aufeinanderfolgende Verarbeitungsebenen in einem neuronalen Netz
- frühe Schichten erkennen einfache Muster, spätere komplexe Strukturen: *jede Schicht transformiert die Daten ein Stück weiter*
- bilden zusammen die *Architektur des Modells*

Training

- Anhand von Trainingsdaten versucht der Lernalgorithmus eine eigene *Hypothese* (Modell) zu entwickeln, die zu den Annotationen passt
- Läuft in mehreren Runden ab und ist **energieintensiv**

- die Neuronen in einer Schicht können parallel berechnet werden, weil ihr Ergebnis nur aus den Werten von der Schicht vor ihnen abhängt
- ähnliche parallele Berechnungen werden sehr effizient von Grafikkarten ausgeführt, die sich deswegen gut für KI-Modelle eignen
- Ähnlich wie der Arbeitspeicher der CPU (RAM), gibt es auch für die Grafikkarte einen Arbeitspeicher (VRAM)

Modell

- ist das **Ergebnis des Trainings**
- Kann Vorhersagen zu Objekten machen, die es davor noch nicht gesehen hat (*Inferenz*)

Kontext

- sind **alle Informationen**, die dem Modell für die eigene *Vorhersage* zur Verfügung stehen

Beispiel

- ist bei einem Chatbot nicht nur die letzte Nachricht, sondern die **gesamte Konversation**

Gewichte

- sind *numerische Werte*, die in einem Modell gelernt werden
- **steuern den Fluss der Aktivierung** im neuronalen Netzwerk und verändern sich nur während des Trainings, danach stehen sie fest
- speichern „*Wissen*“ des Modells in verdichteter Form

⚠ Achtung

KI-Modelle lernen nach dem Training nicht mehr weiter. Ihre grundsätzlichen Fähigkeiten und ihr gespeichertes *Wissen* bleiben für immer auf dem Stand des letzten Trainings **eingefroren**.

Knowledge Cutoff

- Das Datum, ab dem keine aktuellen Informationen mehr in den Trainingsdaten vorkommen

Was sind offene Modelle?

Erkenntnis

Wenn die alle Parameter wie z.b. die *Gewichte* eines fertig trainierten Modells veröffentlicht werden, sprechen wir von einem **Open Weight Model**.

-> erlaubt es, das Modell auf eigener Hardware auszuführen

Erkenntnis

Wenn zusätzlich auch die *Trainingsdaten* und *Validierungsdaten* veröffentlicht werden, sprechen wir von einem **Open Source Model**.

-> erlaubt es, das Training des Modells nachzuvollziehen

Was braucht es für lokale KI-Modelle?

Modelle

- Grundlage aller KI-Anwendungen
- Training durch große Firmen auf teurer Hardware

z.B. DeepSeek, Gemma, Qwen

Modell-Runner

- Verwaltet Modell-Ausführung & Datentransformation
- Orchestriert Hardware-Ressourcen

z.B. Ollama, vLLM

Modell-Interfaces

- Leitet Anfragen an die Runner-API weiter
- Stellt Antworten lesbar dar

z.B. OpenWebUI, HAWKI

Parameter

- sind alle Gewichte und andere Werte die beim Training des Modells verändert bzw. *erlernt* werden
- diese Anzahl bestimmt maßgeblich die Modellgröße und den benötigten Speicherplatz

HINTERGRUND-WEBSITE

https://apxml.com/models/gemma-3-4b

Open

Token

- ist die kleinste verarbeitbare Einheit von Text für ein Sprachmodell, kann ein Wort, Wortteil oder Zeichen sein
- **Kosten, Kontextlänge und Geschwindigkeit werden oft in Tokens gemessen**, diese Messdaten sind relevant bei der Auswahl des geeigneten Modells

HINTERGRUND-WEBSITE

https://platform.openai.com/tokenizer

Open

Reasoning

- beschreibt die Fähigkeit eines Modells, mehrschrittig zu denken
- umfasst logisches Schlussfolgern, Planen und Kombinieren von Informationen
- entsteht nicht explizit als Regelwerk, sondern aus Trainingsmustern
- variiert stark je nach Modellgröße und Training

VRAM

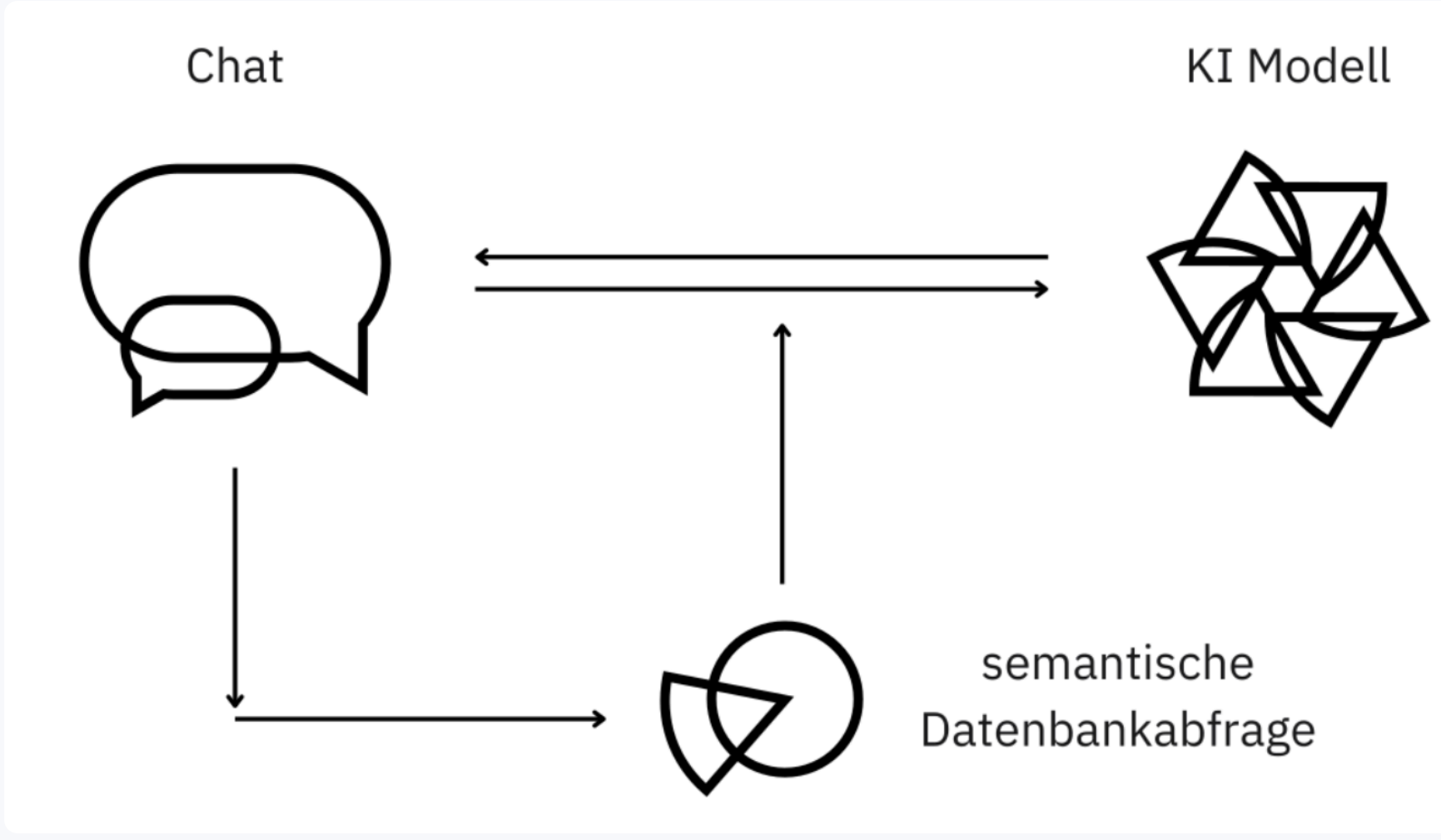
- ist der Speicher auf der Grafikkarte (GPU-RAM)
- wird für Modellgewichte, Zwischenergebnisse und Berechnungen genutzt
- begrenzt, wie große Modelle lokal ausgeführt werden können

Modell	*VRAM (ca.)*	*Anwendungsart*	*Kontextlänge*
Llama 3 8B	6–8 GB	Allround, Chat, RAG	8k
Llama 3 70B	40–48 GB	High-End Chat, Reasoning	8k
Mistral 7B	5–7 GB	Effizienter Allrounder	8k
DeepSeek LLM 7B	6–8 GB	Coding, Mathe	16k
DeepSeek V2 (MoE)	20–40 GB	Stark in Reasoning, effizient	128k
Whisper (large)	8–10 GB	Speech-to-Text	-
Moondream	2–4 GB	Vision-Language (Bild + Text)	~2k–4k
Stable Diffusion (SDXL)	8–12 GB	Bildgenerierung	-

Quantisation

- macht es möglich große Modelle mit **weniger Speicher** auszuführen, **auf Kosten der Genauigkeit**
- dafür wird die *numerische Genauigkeit* der Gewichte reduziert (z. B. von 32-bit Fließkommazahlen zu 8-bit oder 4-bit)

Wie ist eine KI-Anwendung aufgebaut?


Chat
KI Modell
semantische
Datenbankabfrage

Schematische Darstellung einer "Retrieval Augmented Generation"-Architektur

HINTERGRUND-VIDEO

https://ludattel.de/files/2026/demo_rag.mp4

Open

Demo einer KI-Anwendung mit Retrieval Augmented Generation zur Beratung bei der Kursauswahl

Open Model Lab

Projektvorstellung

HINTERGRUND-WEBSITE

https://ki.eteach-thueringen.de/

Open

Technische Infrastruktur

- Server betrieben vom eTeach-Netzwerk Thüringen
- unter ki.eteach-thüringen.de stehen offene Modelle über eine **API** oder zum direkten Ausprobieren im **Browser** zur Verfügung

Beratung

- Fallberatung durch die Fachstelle Medientechnologie
- didaktische und technische Betreuung bei der Umsetzung eigener Experimenteller Ki-Anwendungen

HINTERGRUND-WEBSITE

https://ki.eteach-thueringen.de/

Open

Vorteile

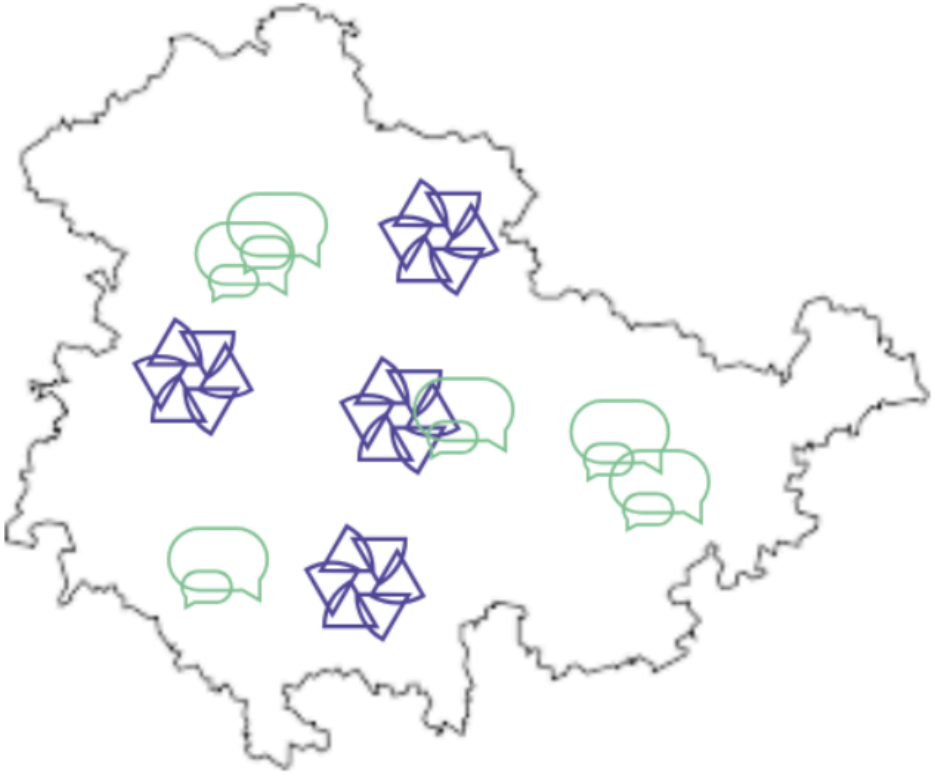
- ✓ datenschutzkonform
- ✓ kostenlose API-Schnittstelle
- ✓ Nutzerkontrolle

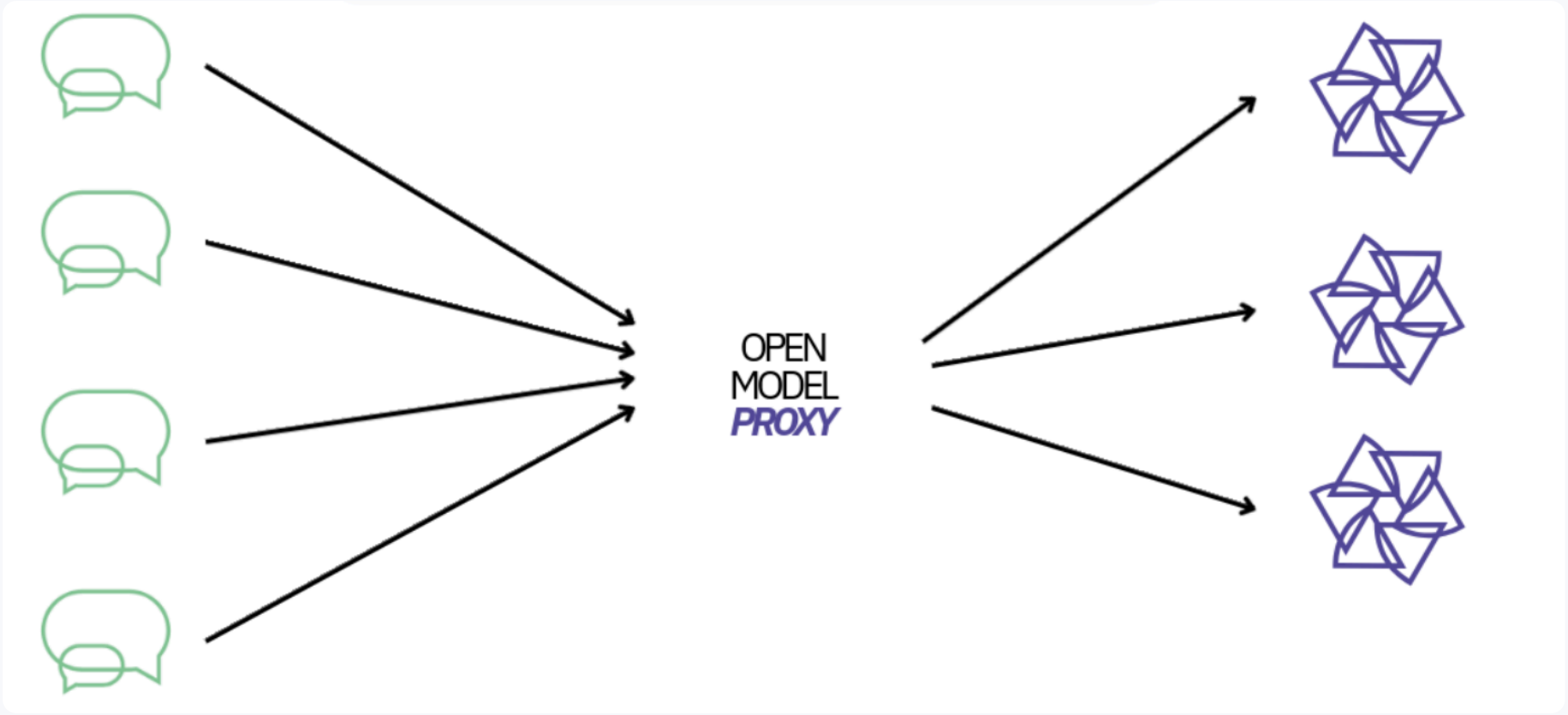
Herausforderungen

- ✗ zentraler Server
- ✗ technisches Verständnis
- ✗ eingeschränkte Leistung

Open Model Proxy

Zukunftspläne




OPEN
MODEL
PROXY

Open Model Lab

Hands-On

Ludwig Lorenz

Fachstelle Medientechnologie

linkedin.com/in/ludwig-lorenz/

ludwig.david.lorenz@uni-weimar.de