

Digital Souveräne Künstliche Intelligenz an Hochschulen

Gerd Kortemeyer, ETH Zürich



Digitale Souveränität bei Künstlicher Intelligenz

US



OpenAI



Google

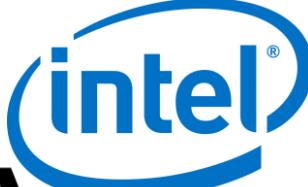
Microsoft



ANTHROPIC



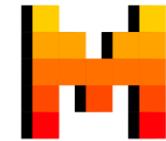
NVIDIA®



Meta

AMD

EUCHGB



MISTRAL
AI_

APERTVS



DeepL



black forest labs.

Digitale Souveränität bei Künstlicher Intelligenz

Der Zug ist
abgefahren?



Aus Hochschulperspektive



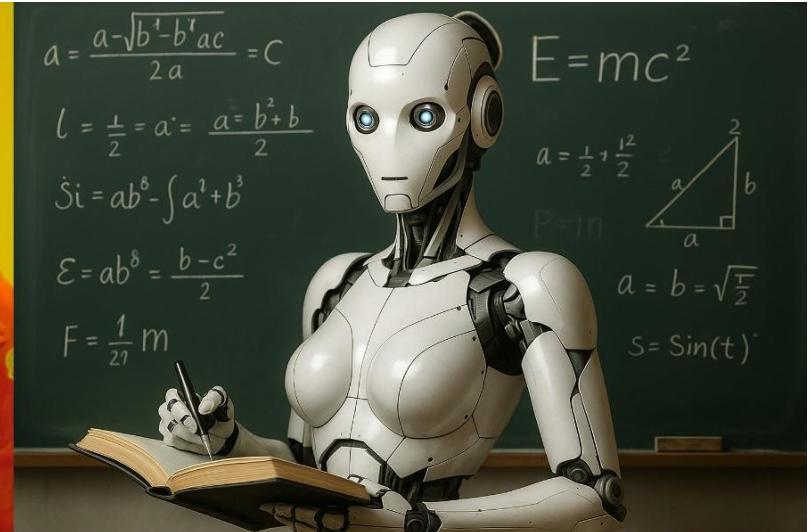
Perspektive

- Lehren und Lernen an einer technischen Hochschule
- Fokus auf MINT



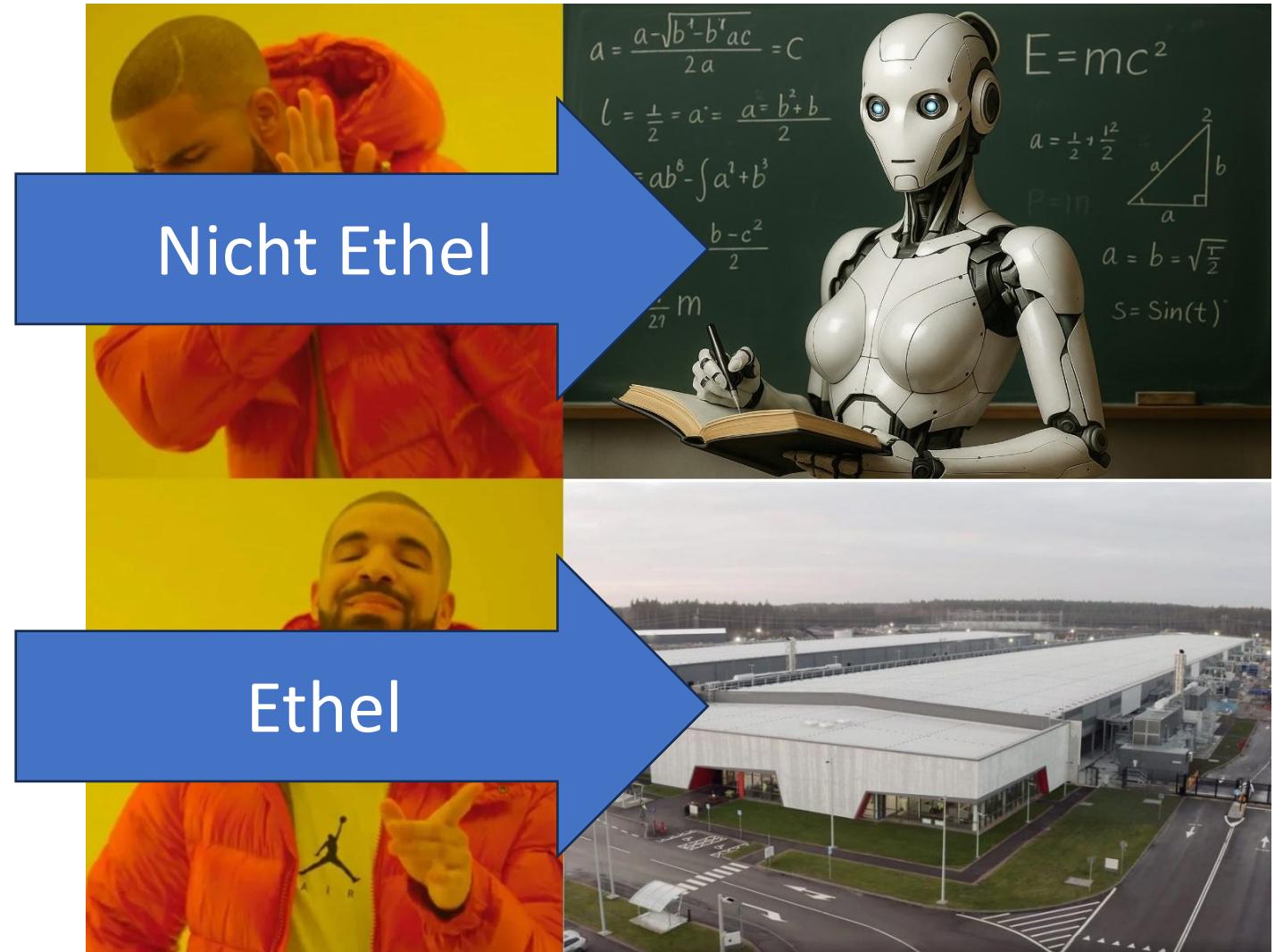
Rolle von KI an der ETH

- Vermenschlichte „Tutoren“ nicht im Vordergrund
- KI im Hintergrund – als Infrastruktur



Speziell: Ethel

- Bei uns an der ETH als zentrales KI-System: open-source System Ethel
- Kosten zentral getragen
- Seit über vier Semestern im Einsatz



Speziell: Ethel

- Kurs-spezifische Chatbots
 - derzeit > 20 Kurse
 - $\cong 4000$ Studierende
- Basierend auf den jeweiligen Kursunterlagen:
 - Skripte, Hausübungen, alte Prüfungen, Powerpoints, ...
- Retrieval-Augmented Generation
 - weniger Halluzinationen



1. Die Ausbreitungsgeschwindigkeit nimmt mit der Z ist, desto schneller kehren die Massenelemente in
2. Die Ausbreitungsgeschwindigkeit nimmt mit der Z ist, desto langsamer kehren die Massenelemente in

Explain that example about waves on a rope; I don't get what that "dm" is about.

The term "dm" in the context of the wave on a rope example represents a small mass element of the rope. When waves propagate in a medium like a rope, the energy of the wave is distributed among small pieces or elements of that medium.

In the given example, they're considering a small segment of the rope with mass "dm". This segment is small enough that it can be treated as a point mass for the purpose of calculations, but large enough that it contains a very large number of atoms so macroscopic quantities like mass can still be defined.

The kinetic energy of this mass element "dm" is given by the equation

$$dT = \frac{1}{2}v^2dm,$$

where $v = \frac{\partial \xi(x,t)}{\partial t}$ is the speed of the mass element, and $\xi(x, t)$ is the displacement of the rope at position x and time t .

The kinetic energy density (kinetic energy per unit volume) is then obtained by dividing this kinetic energy by the volume of the mass element, which is dV . With $\rho = \frac{dm}{dV}$ as the mass density of the rope, the kinetic energy density becomes

$$\frac{dT}{dV} = \frac{1}{2}\rho\left(\frac{\partial \xi}{\partial t}\right)^2.$$

I hope this clarifies the role of "dm" in the energy calculations for a wave on a rope. Let me know if you have any further questions!

The "v" in your answer seems different from the "v" we got out of the partial differential equation for eta(x,t) - is that the velocity of the wave?

Yes, the "v" in the context of the equation for $\xi(x, t)$ is indeed the velocity of the wave, more specifically, the phase velocity. This is the speed at which a particular phase of the wave (for instance, the crest) propagates in space. It can be represented by the equation $v = \frac{\omega}{k}$, where ω is the angular frequency of the wave and k is the wave number.

In the kinetic energy equation, v represents the speed of the small mass element "dm" of the rope, which is a function of time and position on the rope and is related to the displacement of the rope.

These two "v"s, while represented by the same letter, refer to different concepts. The context should make it clear which one is referred to.

$$\frac{dT}{dV} = \frac{1}{2}\rho\left(\frac{\partial \xi}{\partial t}\right)^2$$

Wir werden nun die elastische Energiedichte dE_{el}/dV spannung $S = F/A$ wirkt, ist die in einem um die Länge λ periodisch potentielle Energie gegeben durch $dE_{el} =$

Speziell: Ethel

Hausübungsfeedback

- Studierende laden ihre Lösungen in Ethel und bekommen dazu Feedback

4 (Betrachte v stets grösser 0, falls $v < 0$, wechsle Vorzeichen von β)
Bez. Erde: Raumschiff schickt Signale mit freq $f_0 = 1/\lambda_0$
Auf der Erde erhält er diese mit der Freq: $f_0 = \sqrt{1-\beta^2}/\lambda$

$$f_1 = \sqrt{\frac{1-\beta}{1+\beta}} f_0 = (1-\beta) f_0$$

Mit der Summe $S_1 = f_1 \cdot \frac{\lambda_0}{v} = (1-\beta) \cdot \frac{f_0 \cdot \lambda_0}{v}$

Auf der Rückreise:

$$f_1' = \sqrt{\frac{1+\beta}{1-\beta}} f_0 = (1+\beta) f_0$$

2.4 Problem 4

The student's solution approaches the exercise correctly, but there are several discrepancies and wrong calculations.

a) The student correctly identifies the frequencies of the signals detected from brother A in both parts of the trip (when moving away and towards Earth). The Doppler shift formula is used accurately addressing both cases, however, the equalities

$$f_1 = (1 - \beta) f_0 \quad \text{and} \quad f_1' = (1 + \beta) f_0$$

are incorrect since the Doppler shift factor should be precisely

$$\sqrt{\frac{1-\beta}{1+\beta}} \quad \text{and} \quad \sqrt{\frac{1+\beta}{1-\beta}}$$

respectively, as correctly stated in the sample solution's equations (26) and (27).



Speziell: Ethel

Prototyp: Assistenz bei Prüfungsbenotung

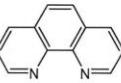
- Bewertung handschriftlicher Prüfungen
- Rubriken
- Human-in-the-loop



Aufgabe 6-C

(5.0 Punkte) _____

Betrachten Sie das pinkfarbene Komplex-Ion $[\text{Ni}(\text{phen})_3]^{2+}$. Der phen-Ligand hat die folgende Struktur:



- a) Ergänzen Sie die folgende Tabelle:

(1.0) _____

	$[\text{Ni}(\text{phen})_3]^{2+}$
Oxidationszahl des Metalls	2+
Anzahl der d-Elektronen des Metallions	8
Koordinationszahl des Metallions	6
Anzahl der Liganden	3

- b) Skizzieren Sie die Verteilung der d-Elektronen des Komplex-Ions $[\text{Ni}(\text{phen})_3]^{2+}$ in dem Kristalfeld seiner Liganden.



(1.0) _____

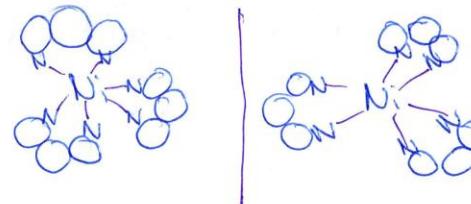
- c) Bestimmen Sie die magnetischen Eigenschaften von $[\text{Ni}(\text{phen})_3]^{2+}$.

(0.5) _____

paramagnetisch

- d) Ist der Komplex chiral? Begründen Sie Ihre Aussage mit einer geeigneten Skizze.

(1.5) _____



Nein, die beiden Spiegelbilder können durch ~~die~~ 14 drehen ineinander verschoben werden und mit berücksichtigung der Doppelbindungen

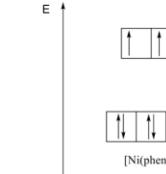
Bewertung 6-C-a

+ 0.25 Punkte für jede gleiche Antwort wie in der Tabelle unten

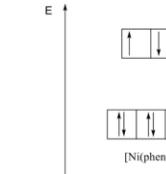
+II
8
6
3

Bewertung 6-C-b

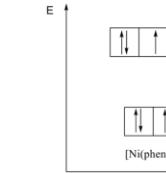
+ 1 Punkt für die folgende Skizze (die Pfeile – Anzahl, Verteilung und Orientierung – sind das Wichtigste)



Bei falscher Skizze: +0.5 Punkte für



Bei falscher Skizze: +0.5 Punkte für



Bewertung 6-C-c

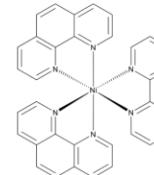
+ 0.5 Punkte für paramagnetisch

Bewertung 6-C-d

+ 0.5 Punkte für chiral
+ 0.5 Punkte für die folgende Skizze

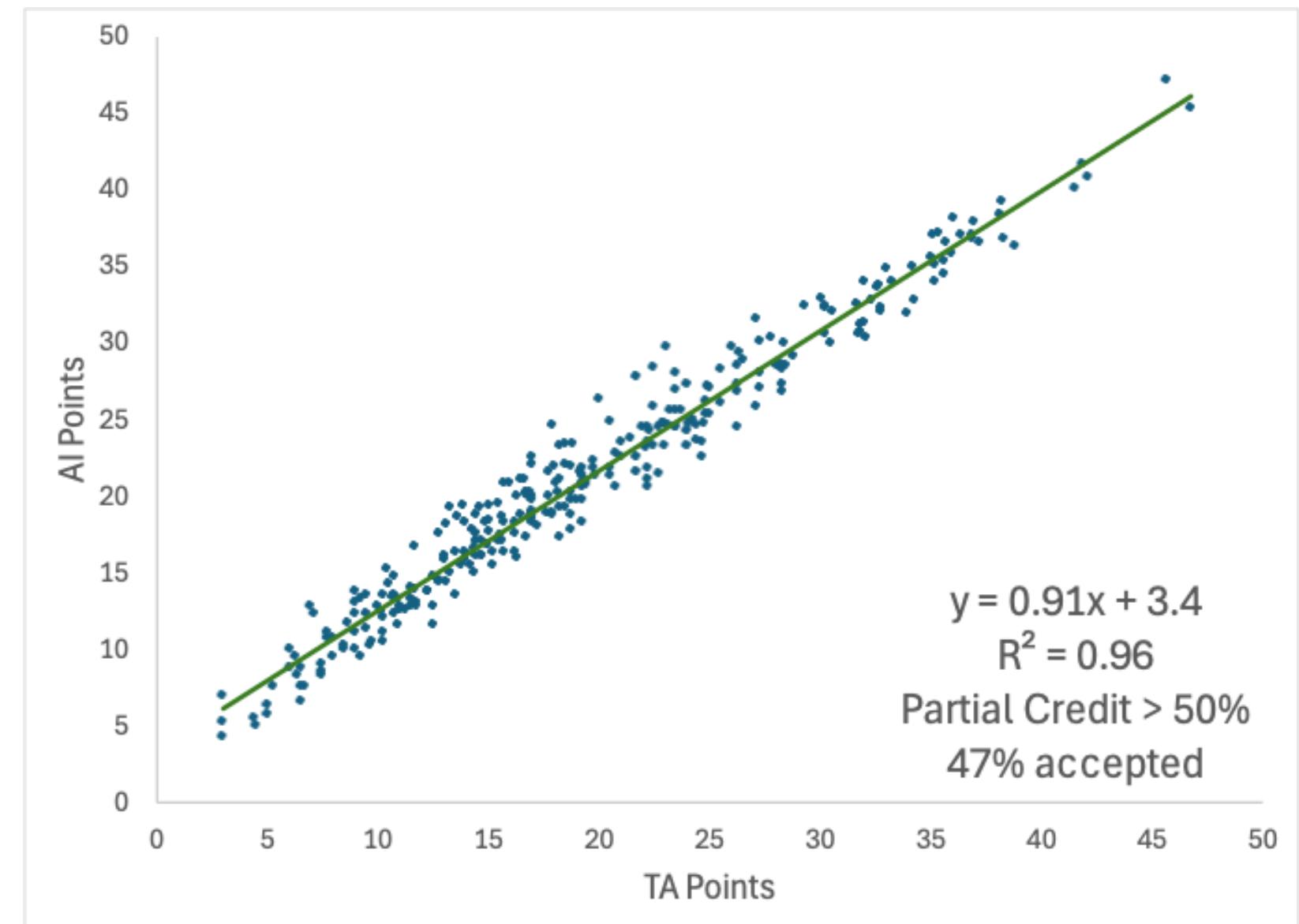


+0.5 Punkte für ein Spiegelbild der oberen Struktur



Speziell: Ethel

Human-in-the-Loop
Experiment



Speziell: Ethel

- Proof-of-concept:
Interaktive
Übungsprobleme
eingebettet in Chat

Give me more problems like the one with the outer wall of the apartment building.

Response 1

A house is constructed with two layers of insulation: the inner layer is made of 10 cm thick wood with a thermal conductivity of 0.12 W/(m·K), and the outer layer is 5 cm thick polystyrene with a thermal conductivity of 0.035 W/(m·K). The interior temperature of the house is maintained at 22°C, while the outside temperature is 2°C. Calculate the total heat loss through the wall per square meter due to conduction. Use the formula for heat transfer through a wall: $Q = (T_1 - T_2) / (R_1 + R_2)$, where R is the thermal resistance calculated as $R = d / (k * A)$.

W/m²

Check Answer

Give me more problems like the one with the outer wall of the apartment building.

Response 2

A house has two outer walls: the first wall consists of 10 cm thick brick, while the second wall is made of a 20 cm thick insulating material. The temperature inside the house is maintained at 22°C, and the outside temperature is -5°C. Given that the thermal conductivity (k-value) of brick is 0.7 W/(m·K) and that of the insulating material is 0.04 W/(m·K), calculate the total heat loss (in watts) through both walls if the total area of the walls is 50 m².

Response 1

A house is constructed with two layers of insulation: the inner layer is made of 10 cm thick wood with a thermal conductivity of 0.12 W/(m·K), and the outer layer is 5 cm thick polystyrene with a thermal conductivity of 0.035 W/(m·K). The interior temperature of the house is maintained at 22°C, while the outside temperature is 2°C. Calculate the total heat loss through the wall per square meter due to conduction. Use the formula for heat transfer through a wall: $Q = (T_1 - T_2) / (R_1 + R_2)$, where R is the thermal resistance calculated as $R = d / (k * A)$.

10 W/m²

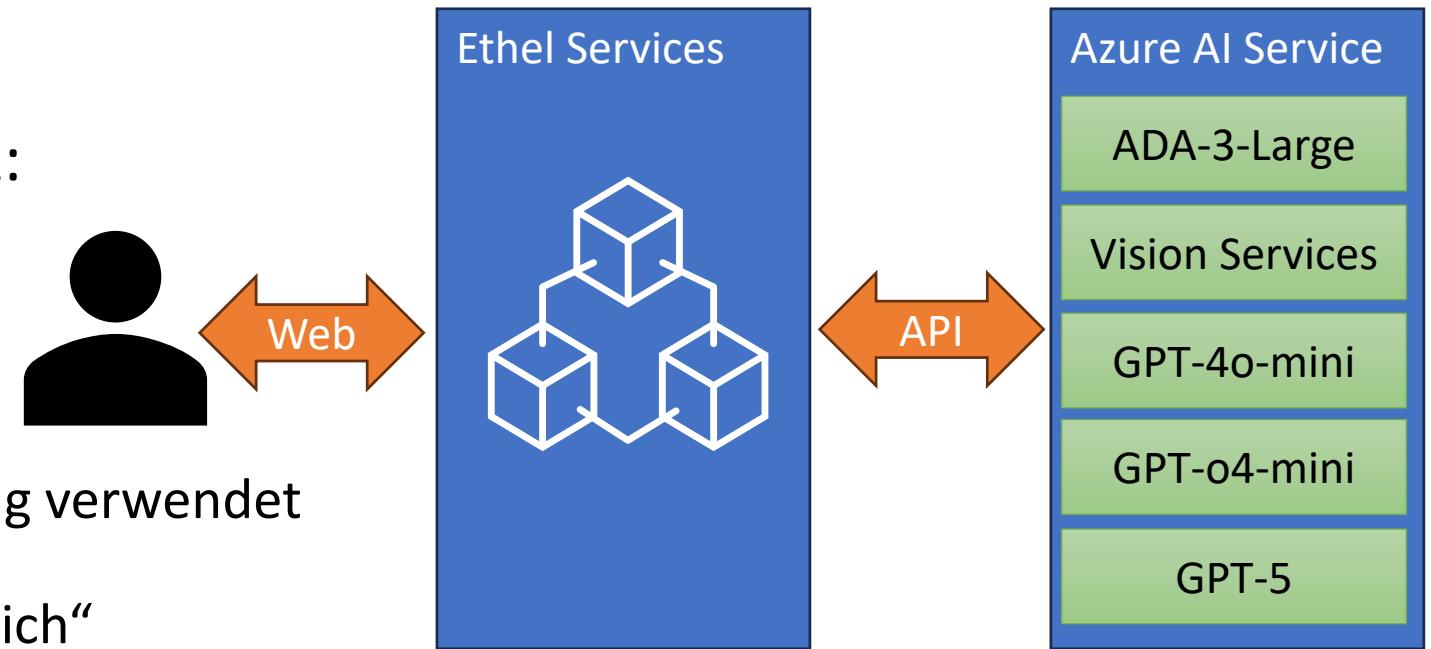
Check Answer

X Incorrect.



Digitale Souveränität?

- Lehrende bei uns bewerten unsere Werkzeuge hauptsächlich aufgrund fachlicher Richtigkeit
 - Wenn das System falsche Antworten liefert, ist es sofort „unten durch“
- Im Hintergrund deshalb:
 - Große Modelle: >> 70B
 - Multimodal
 - Reasoning
- Momentan bei uns im Einsatz:
 - GPT-4o
 - GPT-o4-mini
 - GPT-5
- Azure AI Services
 - Vertrag: Daten nicht für Training verwendet
 - Europäische Rechenzentren
 - Freigegeben bis Stufe „vertraulich“





Da hab' ich
was Eigenes!

Digitale Souveränität
- nur eine Illusion?

Souverän unterwegs ...

- “Normale” IT-Systeme brauchen CPUs
- Generative AI-Systeme brauchen GPUs
 - zum Training, Fine-Tuning und
 - **zum Laufen-Lassen („Inferenz“)**
 - letzteres wird immer wieder vernachlässigt/vergessen/verdrängt
 - **Nicht das gleiche!**



Souverän unterwegs ...



Abbildung kann vom Original abweichen



NVIDIA Tesla H200 NVL FH 141GB PCIe HBM3e 900-21010-0040-000

Nvidia H200 NVL-Passive PCIe 141G AI Ent

29.438,90 €

Energie-
effizienz

inkl. 19% MwSt.

zzgl. Versand ab 8,95 €

1

Warenkorb

Vergleichspool Bestand abfragen

Art.-Nr. 39038768

Hst.-Art.-Nr. 900-21010-0040-000

EAN 8592978607388

Lieferzeit: 1-3 Werkstage

Souverän unterwegs ...

- “Open-weight“ kann man zwar einfach herunterladen und bei sich laufen lassen, aber: für eine Hochschule reicht nicht eine Karte!
 - Erfahrung aus Auslastung von Ethel an der ETH Zürich: ein mittleres Modell
 - 70B
 - non-reasoning
 - non-multimodal
- im **alltäglichen** Chatbetrieb mit akzeptablen Antwortzeiten zu betreiben **würde** eine Viertel-Million Euro in Infrastruktur benötigen
- ... plus Stromkosten zum Betrieb (Leistungsaufnahme und Kühlung)
 - ... und dann Elektroschrott nach fünf Jahren



Souverän unterwegs ...

- Eigenbetrieb eines mittleren Modells ($\cong 70B$) ginge ja noch, aber:
 - nicht geeignet für didaktische Szenarien,
 - wo in einem Hörsaal Hunderte von Studierenden auf einmal auf das System zugreifen
 - oder wenn es um MINT geht (derzeitige mittlere Modelle nicht “intelligent” genug)
- Ein großes Modell ($>>70B$) auf Hochschulebene zu betreiben macht keinen Sinn
 - Besser auf Landes- oder Bundesebene
 - Dann können auch Spitzenlasten besser verteilt werden
- Billig wird das aber auch nicht
 - **VORSICHT:**
Nachfolgend kommt eine Hochrechnung von einer Hochschule (ETH) auf Gesamtdeutschland
 - Das hat natürlich große Unsicherheiten



Souverän unterwegs

Hochschulbetrieb für ganz Deutschland (GPUs: fünf Jahre)

	Eigenhosting	Cloud (Microsoft)	
	Mittleres Modell	Großes Modell	Großes Modell
Anschaffung (ohne CPUs)	60 MEUR	300 MEUR	-
Betrieb fünf Jahre (ohne Personal)	15 MEUR	75 MEUR	120 MEUR
Totale Kosten fünf Jahre	75 MEUR	375 MEUR	120 MEUR



Souverän unterwegs ...

- Hyper-Scalers sind weniger *umweltfeindlich*, weil die Auslastung der Karten mit mehr Kunden besser balanciert werden kann
 - Rechenleistung versus Leistungsaufnahme nicht-linear
- Hyper-Scalers können gebaut werden,
 - wo es mehr verlässliche erneuerbare Energien
 - und es über das Jahr hinweg einfachere Kühlung gibt



Ethels Inferenz
in Schweden



Souverän unterwegs ...



Ein Hochleistungsrechner wie dieser hilft da gar nichts, denn ...

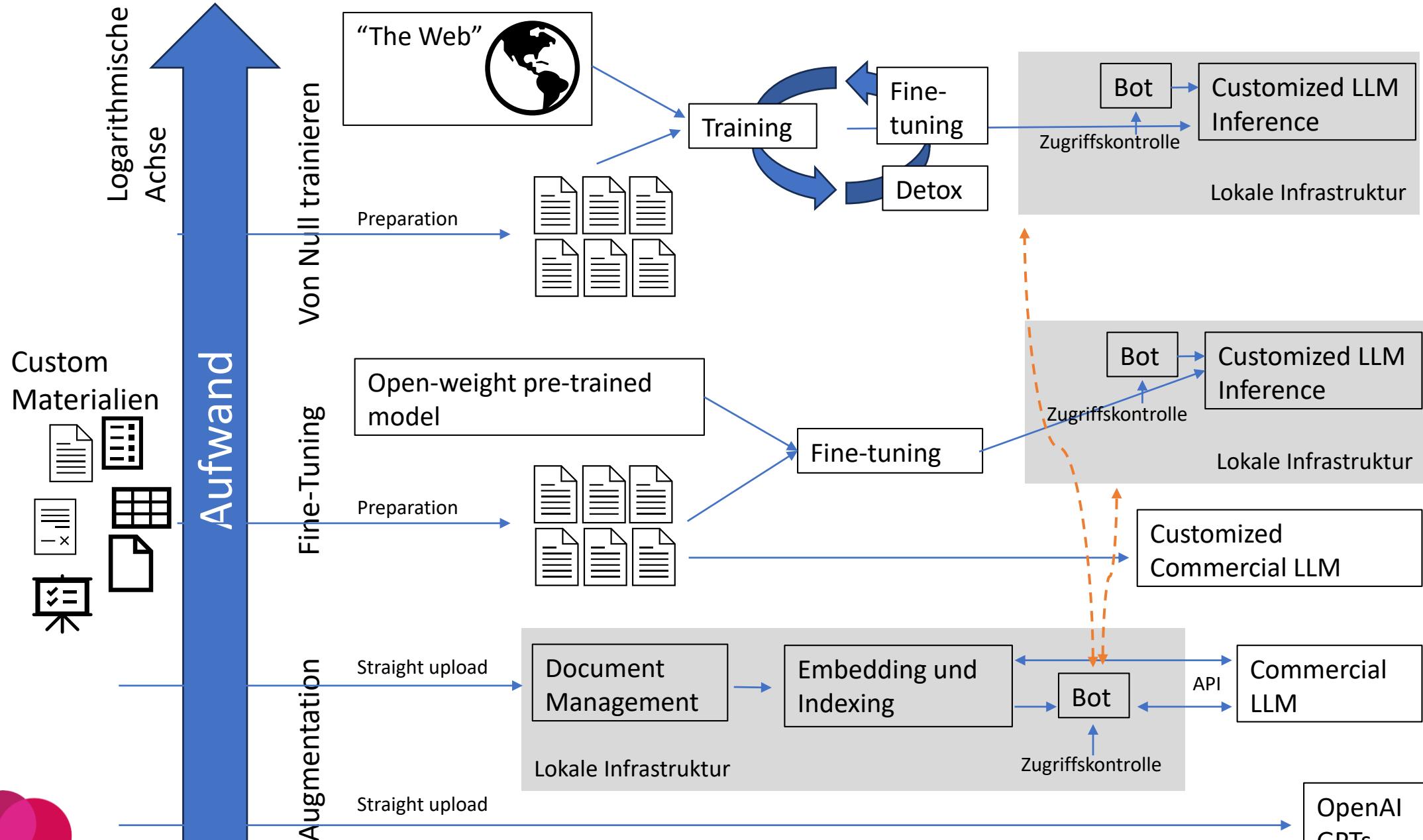


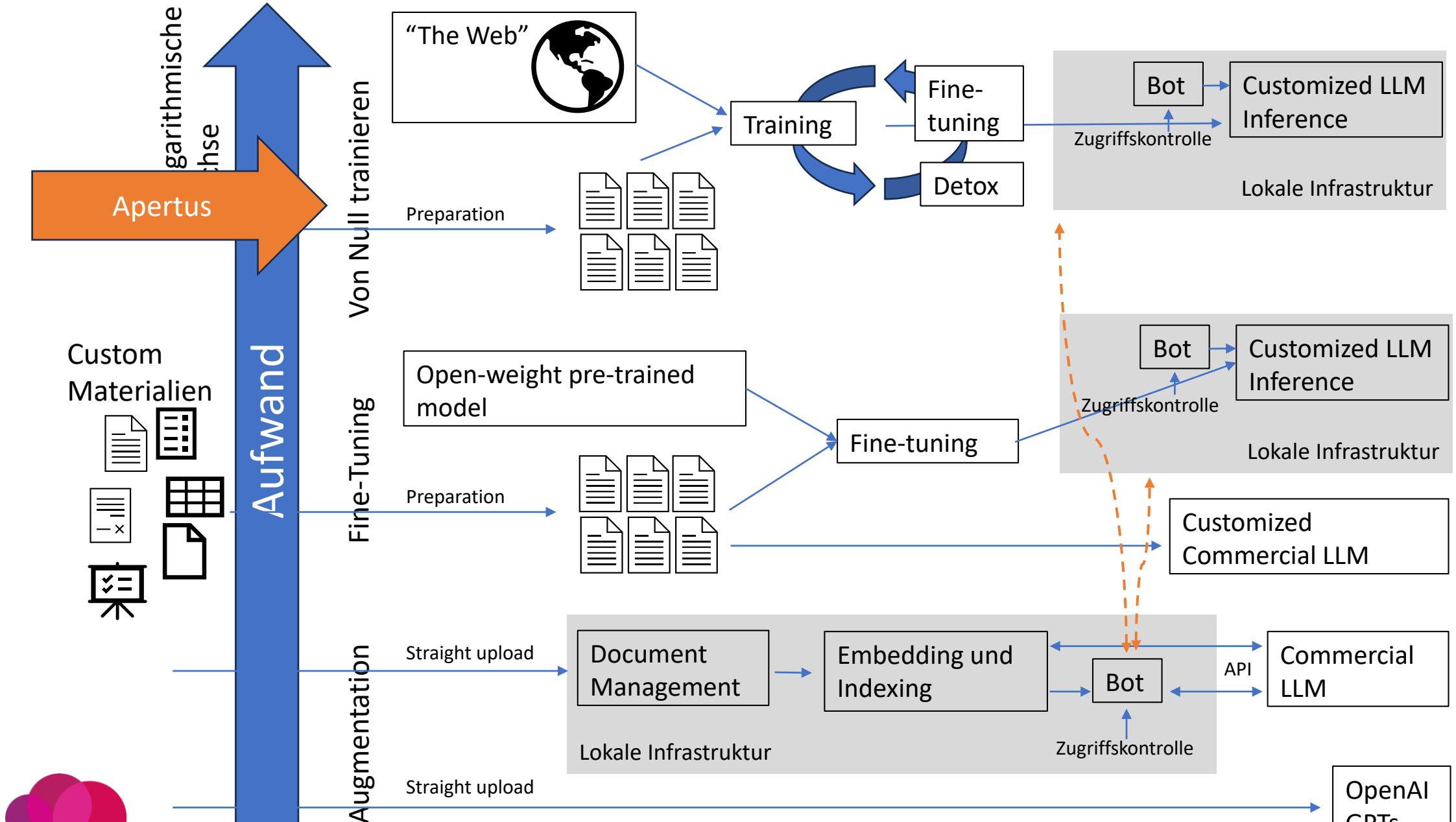
Souverän unterwegs ...

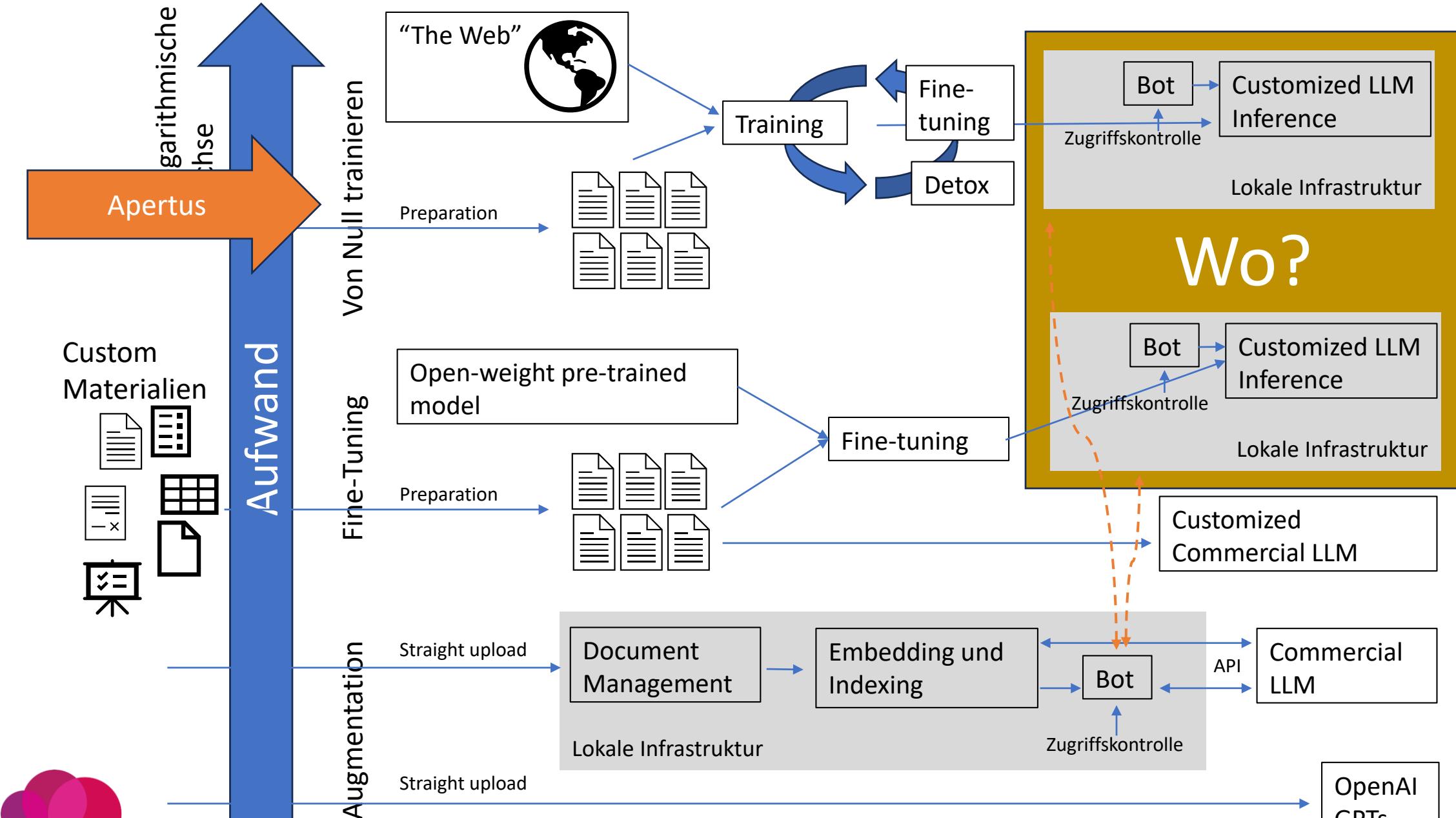
- ... diese Hochleistungsrechner sind für die **Forschung**
 - ausgelegt für Batch-Betrieb (erledigt asynchron Rechenjobs)
 - hat häufig System- oder andere Ausfallzeiten
- Man kann nicht permanent, 24/7 Hunderte von GPUs für den Betrieb eines Sprachmodells abstellen

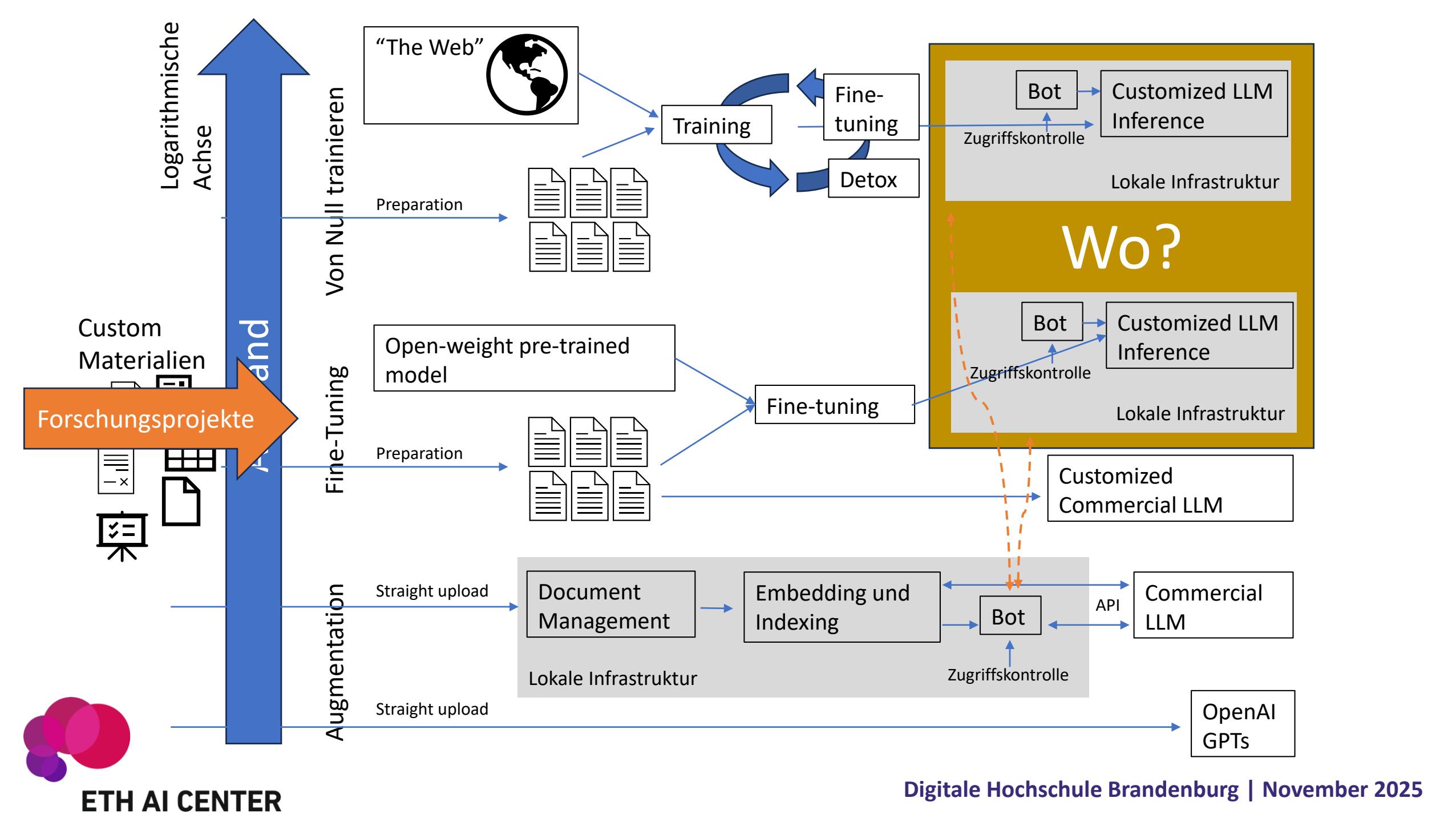
Aber wie bekomme ich dann was Eigenes?

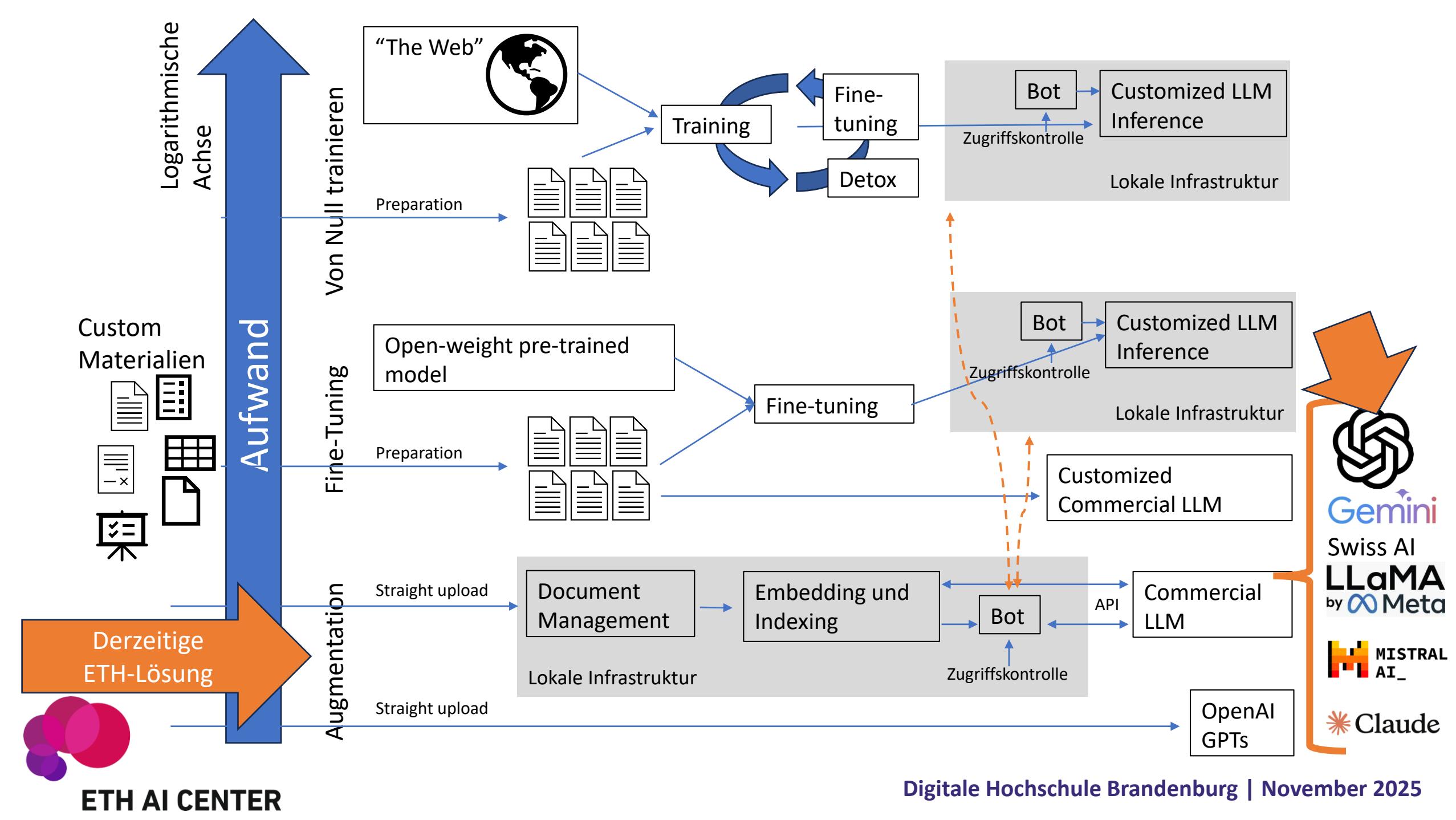






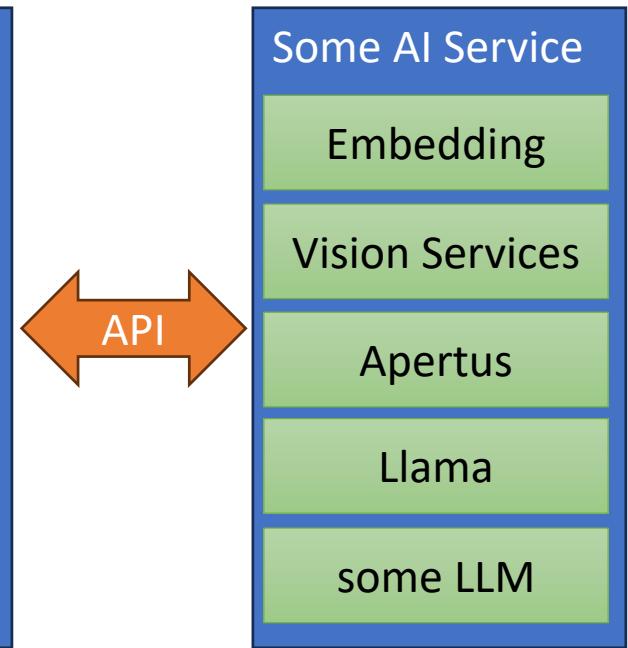
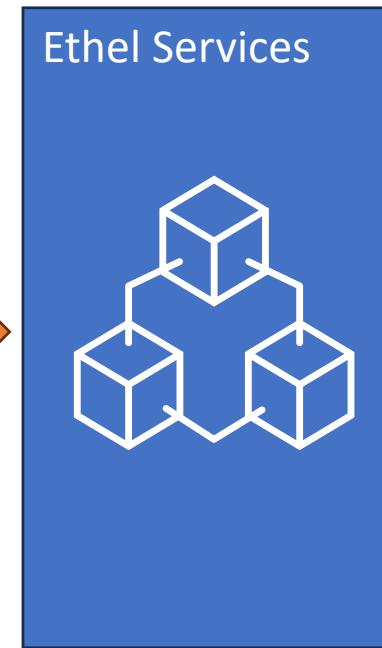






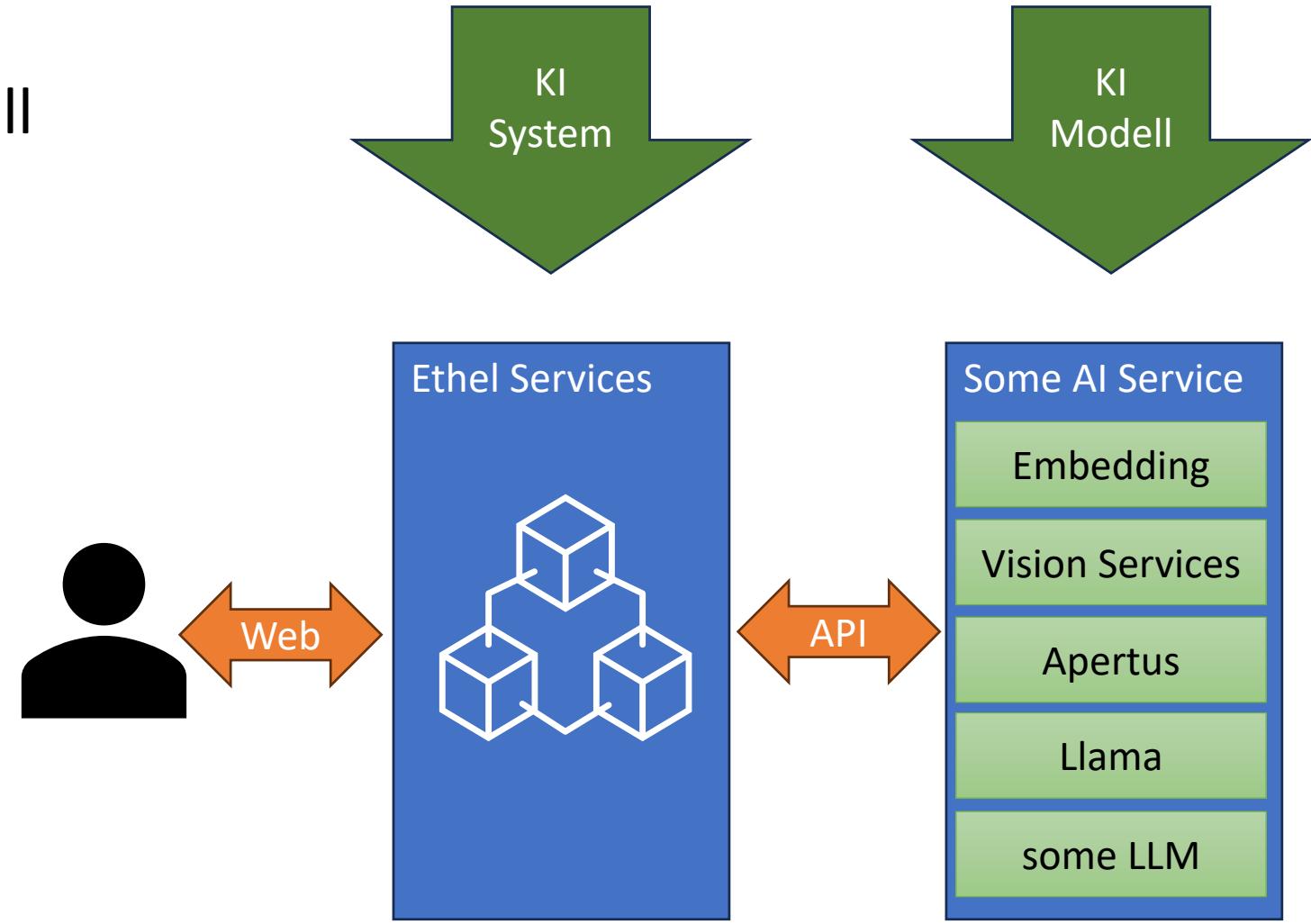
Souverän unterwegs?

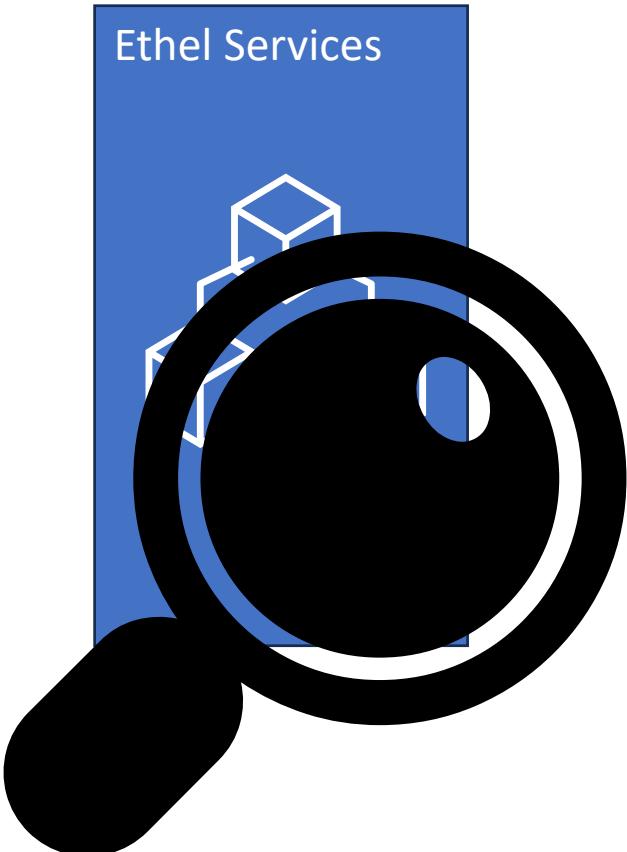
- Nein, derzeit nicht realistisch
- Aber auch kein Vendor-Lock-In!
 - Sprachmodell ist eine „Commodity“ und kann ausgetauscht werden
 - Ethel hat schon vier Mal Modelle gewechselt
 - Wechsel transparent für Nutzende
- Derzeitige Kosten:
 - Durchschnittlich EUR 1.30 pro Studierender pro Monat
 - EUR 0.80 pro Prüfung



Souverän unterwegs?

- KI-System ist nicht KI-Modell



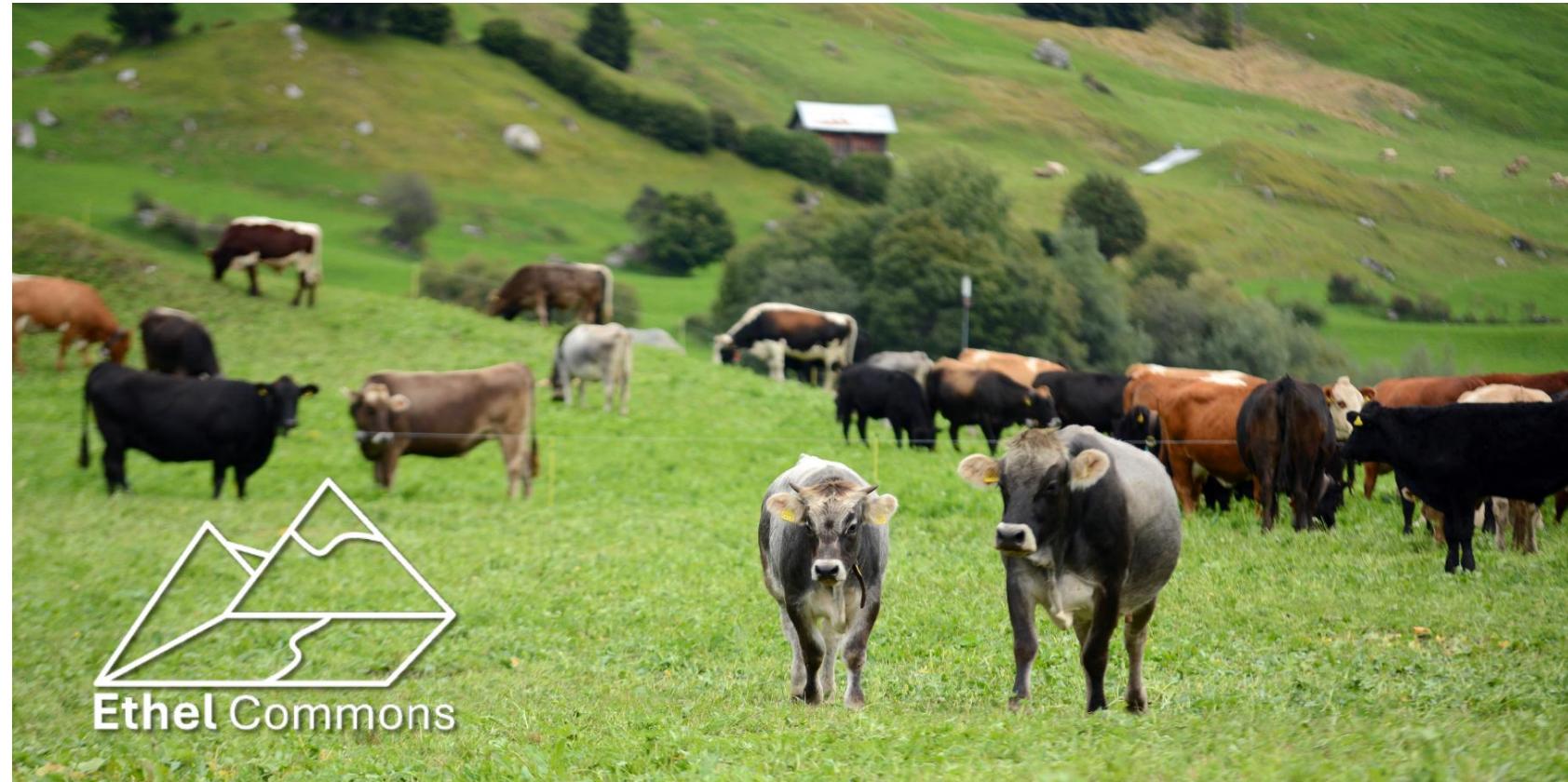


Fokus auf das System!



Service

- KI-System für
 - Frontend
 - Datenverwaltung
 - Nutzerverwaltung
 - Workflows
 - ...
- Entwicklung open-source Micro-Services und Workflow-Management
- Gemeinsame Service- (nicht Inferenz-) Infrastruktur
- Gemeinsame Vertragsverhandlungen für Inferenz
 - in der Schweiz über SWITCH.ch bereits rechtssicher existent (mit Azure AI)



Service

- Empfehlung für den Betrieb von Hochschul-KI **heute**:
 - Fokus auf das System
 - Inferenz als „Commodity“ (wie Strom, Wasser, Internet, ...) einkaufen
 - „Inferenz kommt aus der Steckdose“
- Nicht Geld für Unmengen von GPUs ausgeben
 - Das System braucht nicht viel Rechenleistung, nur CPUs



Gerd Kortemeyer

Rectorate and AI-Center
ETH Zurich
kgerd@ethz.ch

Associate Professor Emeritus
Michigan State University
kortemey@msu.edu

<https://gerdkortemeyer.com/>



ETH AI CENTER

