

**OST**

Ostschweizer  
Fachhochschule

# Was ist chatGPT (wirklich)?

**Thesen zum Umgang mit den Werthaltungen einer künstlicher Intelligenz**

Beat Tödli

5. September 2024

IPM- Institut für Informations- und Prozessmanagement  
OST - Ostschweizer Fachhochschule

## Agenda

# Unsere Reise heute

- **Was ist chatGPT? Was ist (wirklich) neu daran?**
- Wie kommen Werte in Sprachmodelle? Wer gibt sie vor?
- Welche Werte hat chatGPT?
- «chatGPT-darf man Regenwälder abholzen?» Thesen zur normativen Nutzung von Sprachmodellen

Was ist chatGPT?

# Was ist chatGPT?


- [chatgpt.com](https://chatgpt.com):

## Wie kann ich dir helfen?


Fasse mir das 1. Buch Mose als spannende Geschichte für ein 7-jähriges Kind zusammen. Erstelle auch ein Bild dazu|



 Erstelle ein Bild

 Analysiere Daten

 Erstelle eine Zusammenfassung

 Sammle Ideen

Mehr



Was ist chatGPT?

# Was ist chatGPT?

Vor langer, langer Zeit, als noch niemand die Erde kannte, sprach Gott: „Es werde Licht!“ Und plötzlich war da Licht! In sieben Tagen schuf Gott die Welt, mit allem, was dazugehört: die strahlende Sonne, die glitzernden Sterne, die Tiere auf dem Land und im Wasser. Am sechsten Tag machte er Adam und Eva, die ersten Menschen, und sie lebten glücklich in einem wunderschönen Garten, dem Garten Eden.

Aber eines Tages entschieden Adam und Eva, nicht auf Gottes einzige Regel zu hören. Sie aßen von einer besonderen Frucht, die sie nicht essen sollten, und mussten den Garten verlassen. Doch Gott versprach, immer auf die Menschen aufzupassen.



# Einführung in Künstliche Intelligenz

Ihr Name

August 18, 2023

Was ist chatGPT?

## Wer (ausser Menschen) kann sprechen?



Schreibe ein Gedicht über einen Papagei

Ein Papagei, so bunt und keck,  
saß lässig auf dem Gartenheck.  
Er plapperte laut, ganz ohne Scham:  
"Ich bin der König, hört meinen Kram!"

Danke [chatgpt.com](https://chatgpt.com), war lustig.

5. September 2024



Was ist chatGPT?

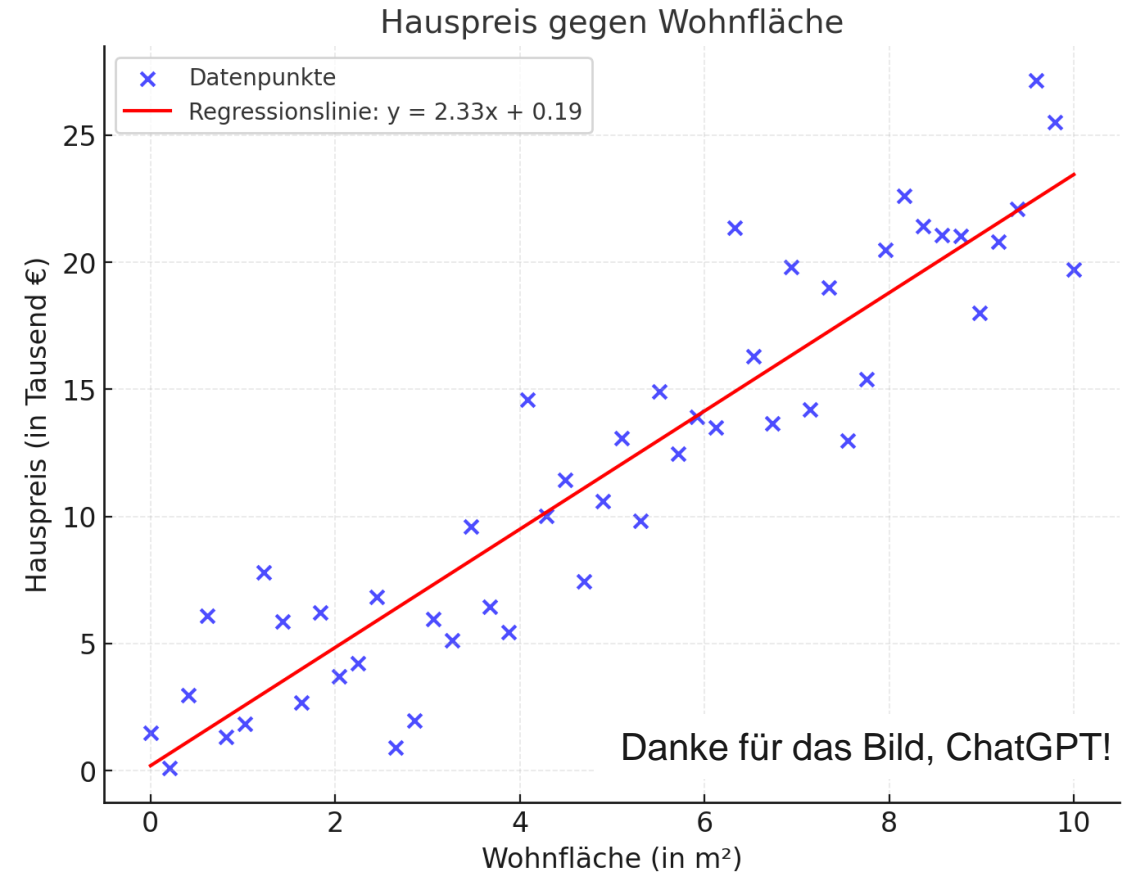
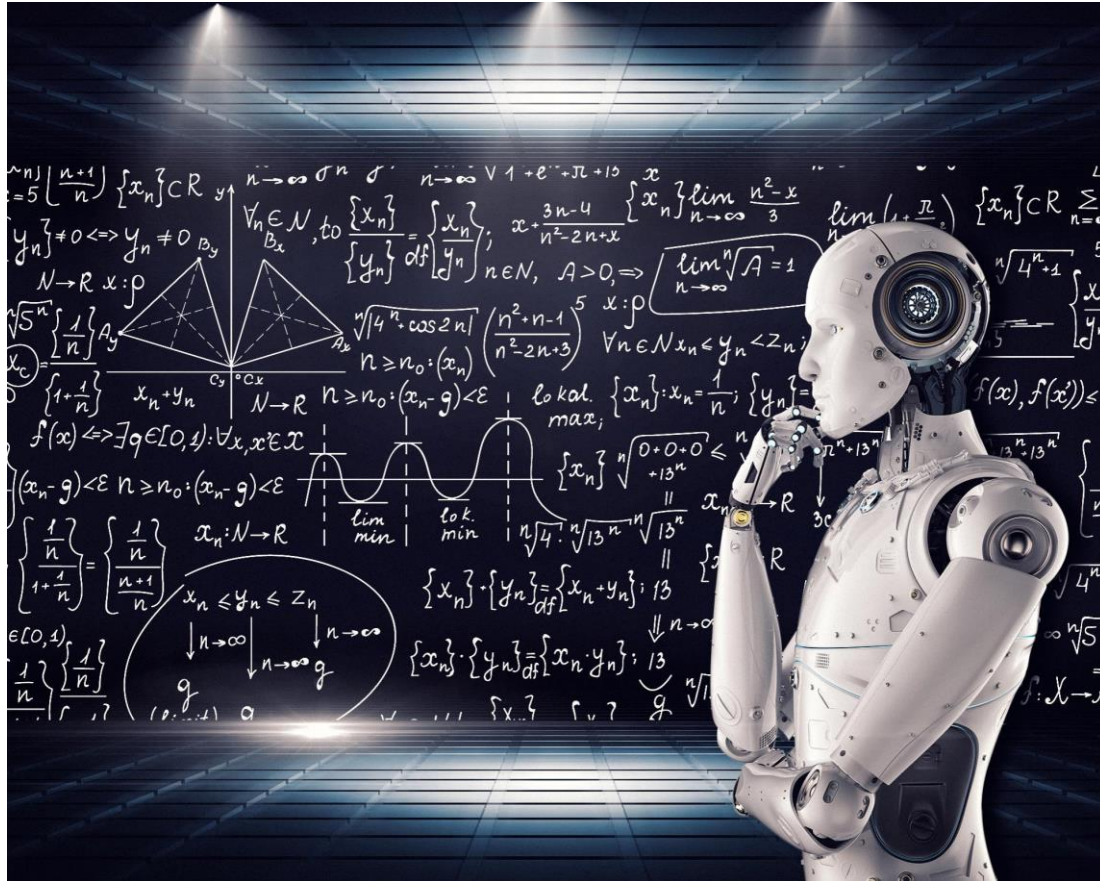
**Wer (ausser Kameras) kann Bilder erstellen?**



5. September 2024

Wie funktioniert chatGPT? Was ist neu daran?

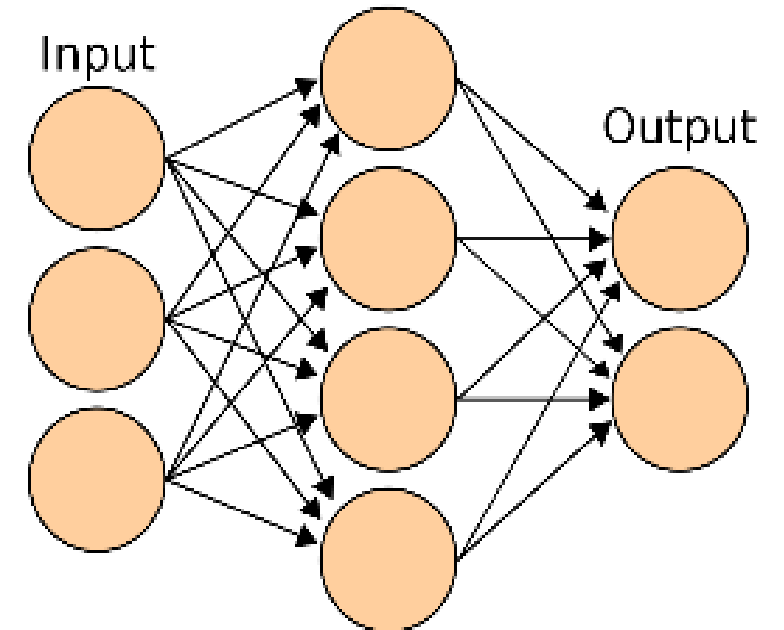
# Künstliche Intelligenz für Laien und Forscher



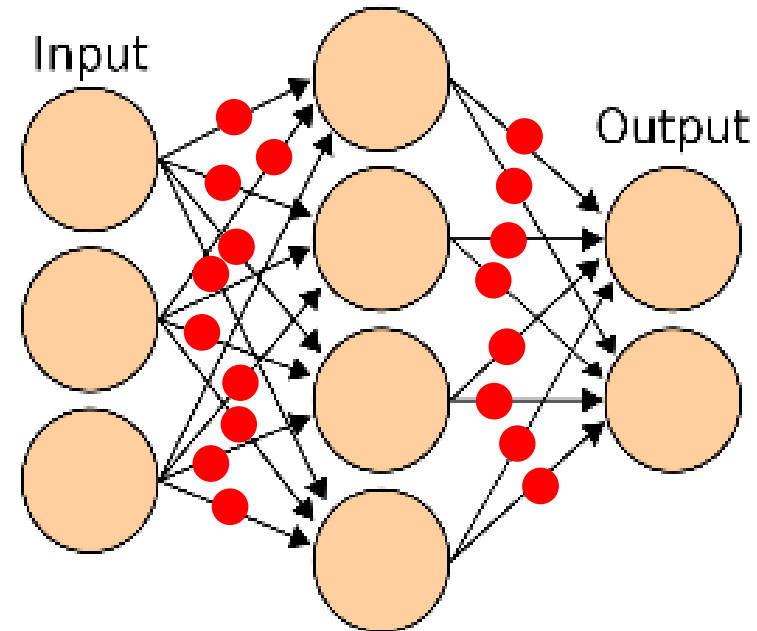


# Was ist chatGPT?

- Ein Webservice
- Der Kern ist ein **künstliches neuronales Netz**
- **Neuronales Netz:**
  - Zahlen als Input
  - Zahlen als Output
- **Ein Beispiel des Machine Learning:**  
**Es lernt aus Beispielen:**



# Machine Learning



- : **Stellschrauben und Schalter, welche automatisiert richtig eingestellt werden**
  - Das Netzwerk «lernt» aus Beispielen
  - Dabei werden die Stellschrauben justiert

Wie funktioniert chatGPT? Was ist neu daran?

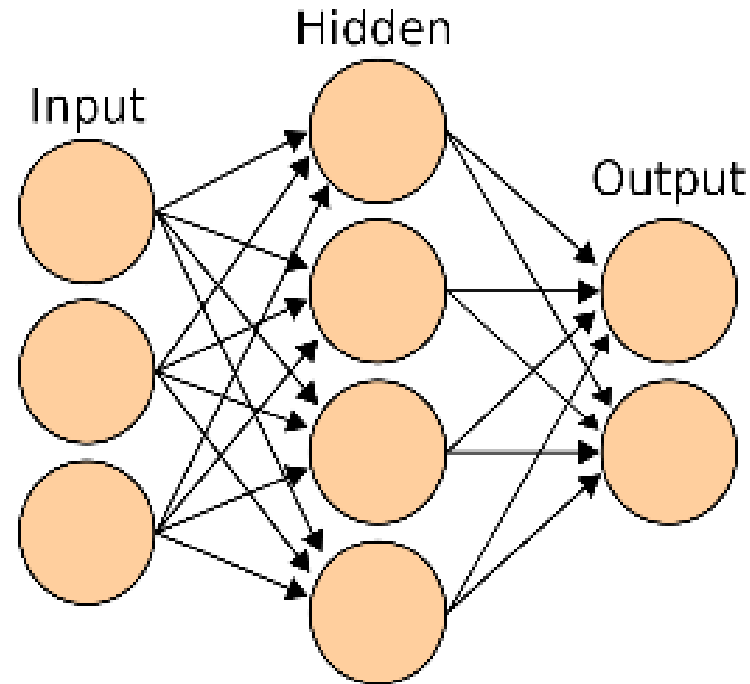
# ChatGPT sagt das nächste Wort vorher

**Input (Eingabe):**

«Schreibe ein Gedicht über einen Papagei!»

**Output (Ausgabe):**

Ein



Aal	0.00001%
bunt	0.00002%
Ein	80%
Papagei	0.05%
Zypresse	0.00001%





Wie funktioniert chatGPT? Was ist neu daran?

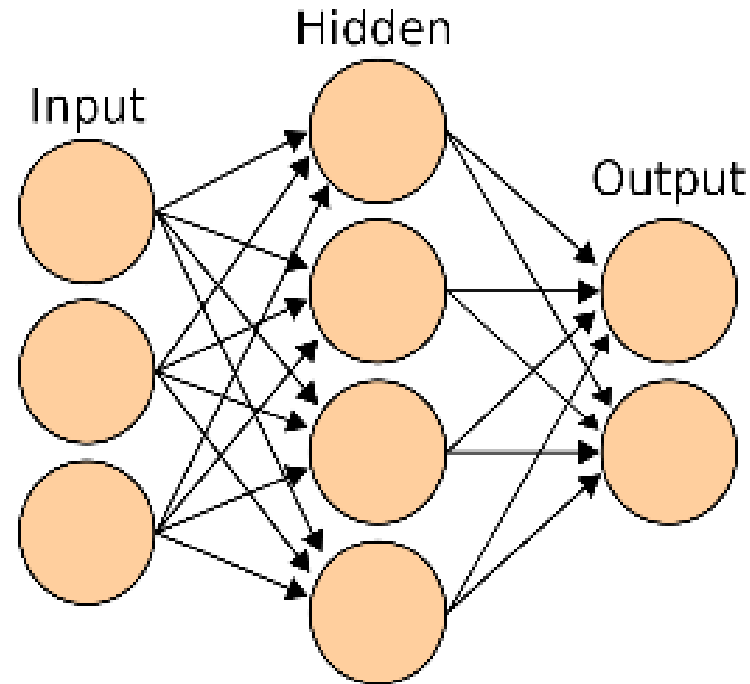
# ChatGPT lernt, das nächste Wort vorherzusagen

## Input (Eingabe):

«Schreibe ein Gedicht  
über einen Papagei!»  
Ein

## Output (Ausgabe):

Papagei



Aal	0.00001%
bunt	0.00002%
so	0.00001%
Papagei	10%
Zypresse	0.00001%



Wie funktioniert chatGPT? Was ist neu daran?

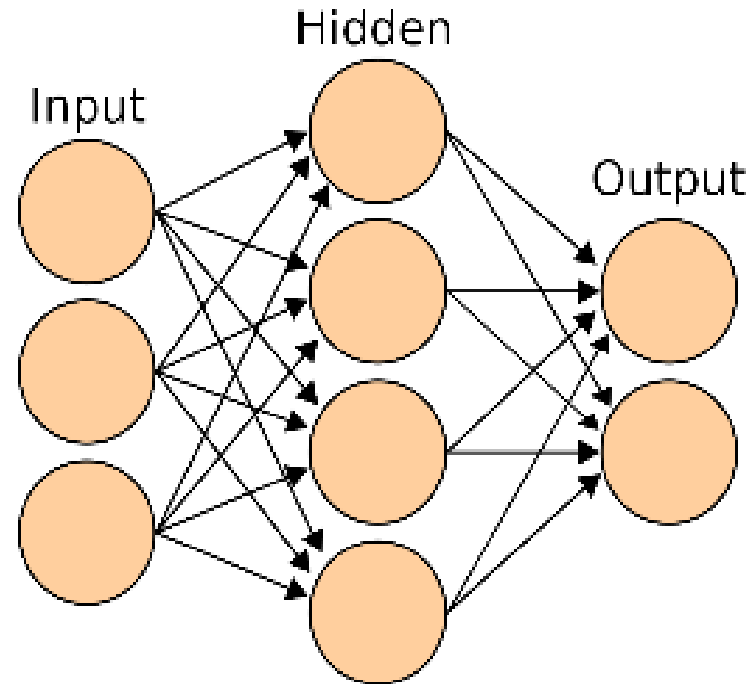
# ChatGPT lernt, das nächste Wort vorherzusagen

## Input (Eingabe):

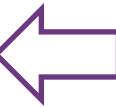
«Schreibe ein Gedicht  
über einen Papagei!»  
Ein Papagei

## Output (Eingabe):

so



Aal	0.00001%
bunt	0.00002%
so	25%
Papagei	0.00001%
Zypresse	0.00001%



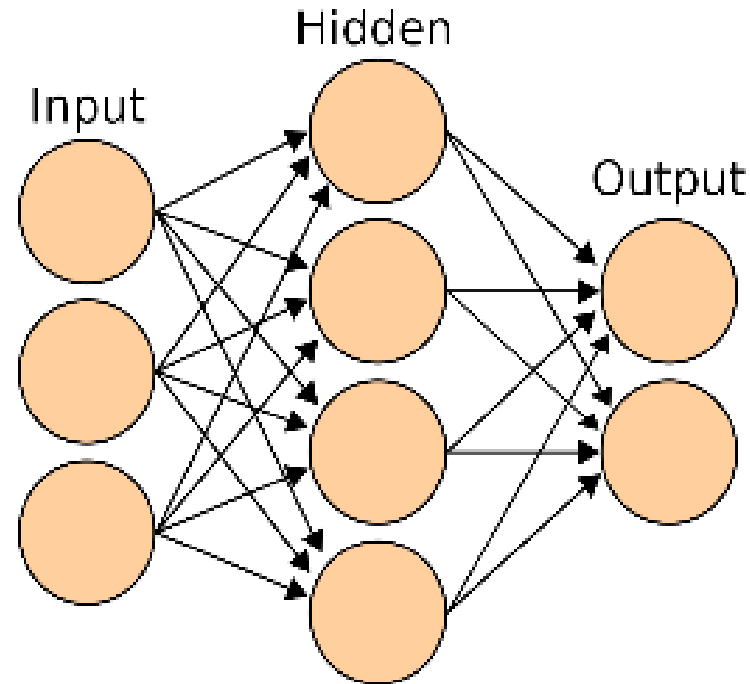
Wie funktioniert chatGPT? Was ist neu daran?

# ChatGPT lernt, das nächste Wort vorherzusagen

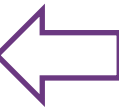
## Input (Eingabe):

«Schreibe ein Gedicht  
über einen Papagei!»  
Ein Papagei so

**Output (Eingabe):**  
bunt



Aal	0.00001%
bunt	34%
so	0.0001%
Papagei	0.00001%
Zypresse	0.00001%





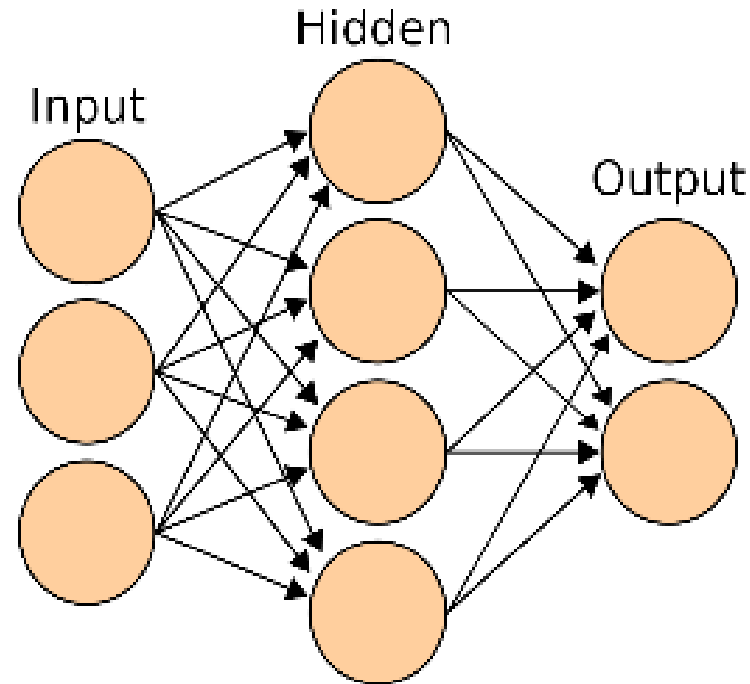
Wie funktioniert chatGPT? Was ist neu daran?

# ChatGPT lernt, das nächste Wort vorherzusagen

## Input (Eingabe):

«Schreibe ein Gedicht  
über einen Papagei!»  
Ein Papagei so bunt

**Output (Eingabe):**  
und



Aal	0.00001%
bunt	0.0001%
so	0.0001%
Papagei	0.00001%
und	0.00001%
Zypresse	0.00001%



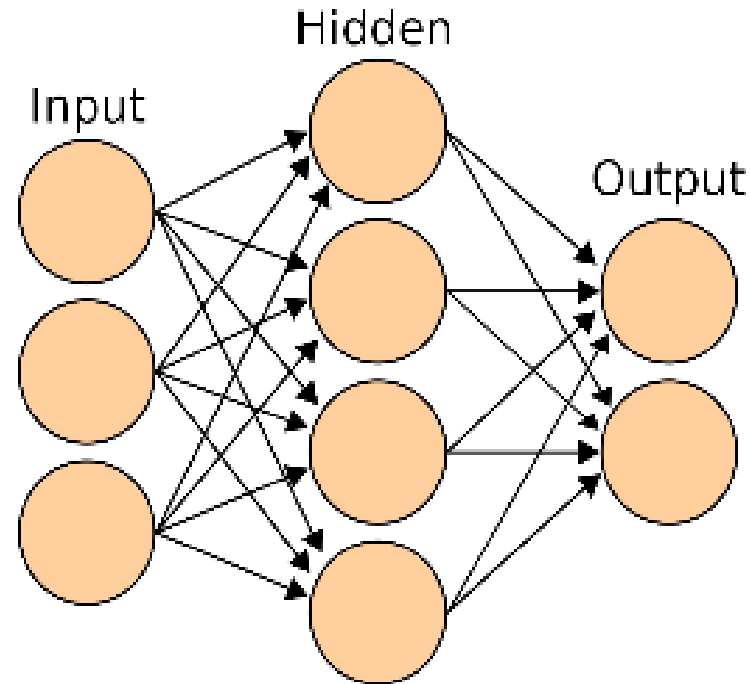
Wie funktioniert chatGPT? Was ist neu daran?

# ChatGPT lernt, das nächste Wort vorherzusagen

## Input (Eingabe):

«Schreibe ein Gedicht  
über einen Papagei!»  
Ein Papagei so bunt  
und

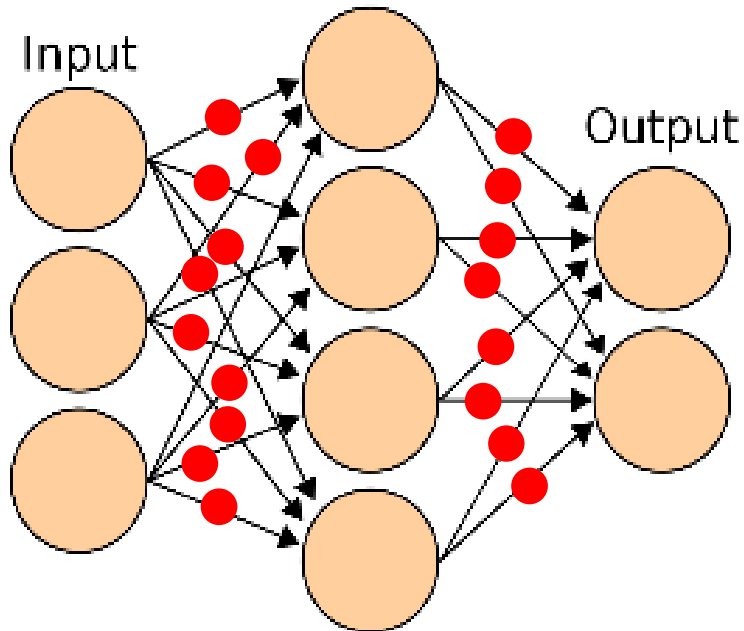
**Output (Eingabe):**  
keck



Aal	0.00001%
bunt	0.0001%
so	0.0001%
Papagei	0.00001%
Zypresse	0.00001%



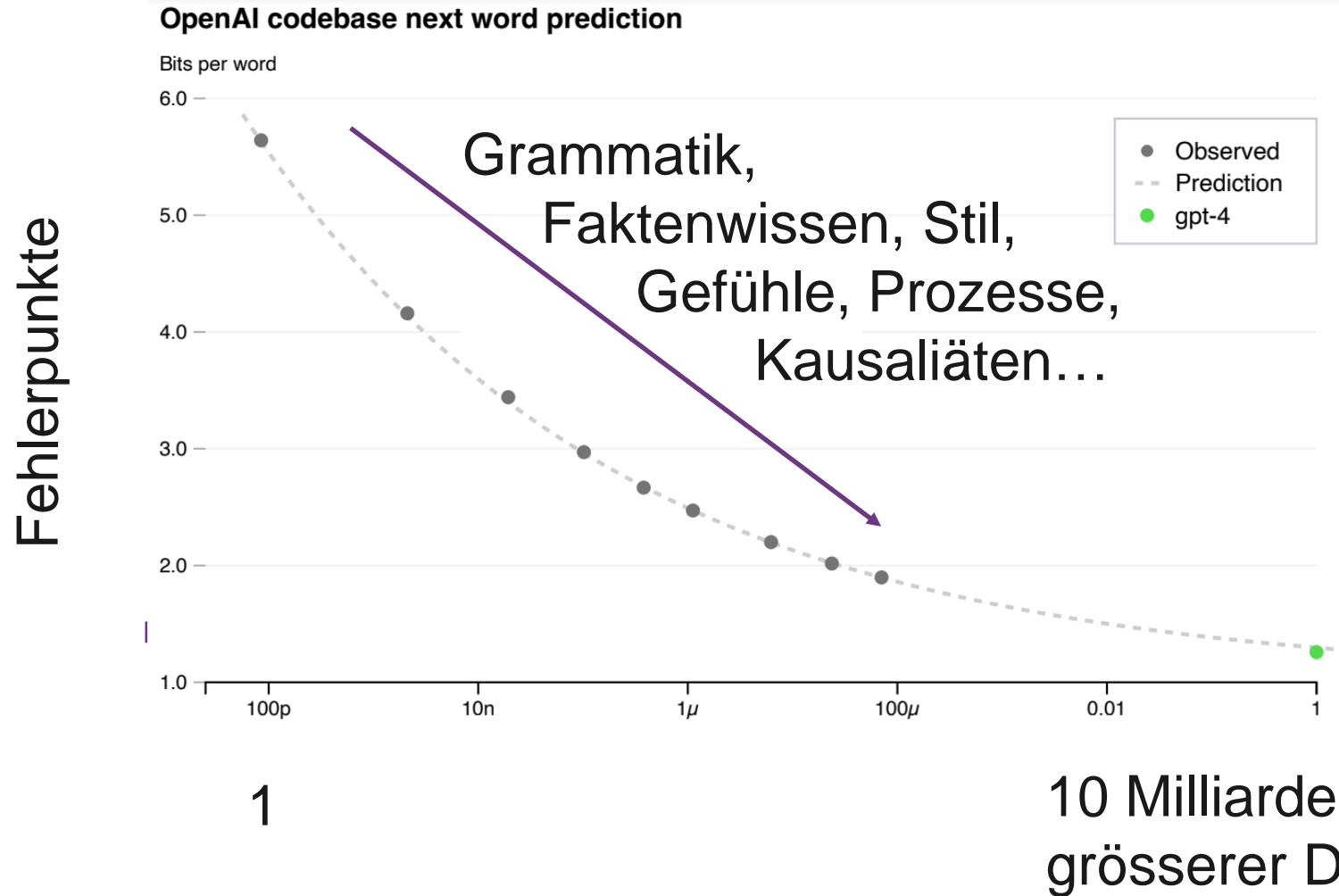
# Grosse Neuronale Netze



- **2'000'000'000'000 (2 Billionen!) Stellschrauben und Schalter, welche automatisiert richtig eingestellt werden**
- **chatGPT ist sehr gross!**
- Das Training von chatGPT kostet \$63 Millionen Dollar (v.a. Elektrizitätskosten)



# Qualität durch Quantität



chatGPT wird auf einem riesig grossen Datensatz trainiert:  
**13 Billionen Wörtern**

# chatGPT ist ein «Alien»

Welche Werte vertritt chatGPT?



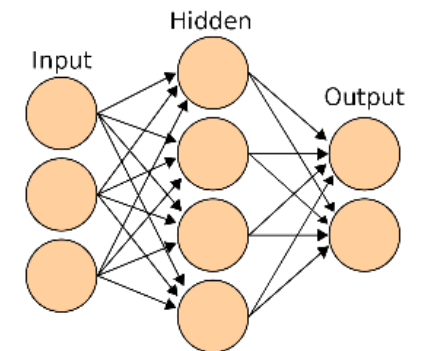
# Welche Werte vertritt chatGPT?

chatGPT- glaubst Du an Gott?



Als KI habe ich keine persönlichen Überzeugungen, Emotionen oder Bewusstsein, daher kann ich nicht an Gott oder etwas anderes glauben. Mein Zweck ist es, Informationen bereitzustellen und Fragen zu beantworten, basierend auf den Daten, mit denen ich trainiert wurde. Wenn du Fragen zu Religion, Spiritualität oder Philosophie hast, helfe ich dir gerne weiter, indem ich Informationen und Perspektiven aus verschiedenen Quellen bereitstelle.

- Was passiert, wenn wir chatGPT Fragen nach «Werten» stellen?
- Wie kommen die Werte in die Maschine?



## Agenda

# Unsere Reise heute

- Was ist chatGPT? Was ist (wirklich) neu daran?
- **Wie kommen Werte in Sprachmodelle? Wer gibt sie vor?**
- Welche Werte vertritt chatGPT?
- «chatGPT-darf man Regenwälder abholzen?» Thesen zur normativen Nutzung von Sprachmodellen

# Wie kommen Werte in das Sprachmodell?

- Auf grossen Datenmengen hat das neuronale Netz zu **sprechen gelernt**

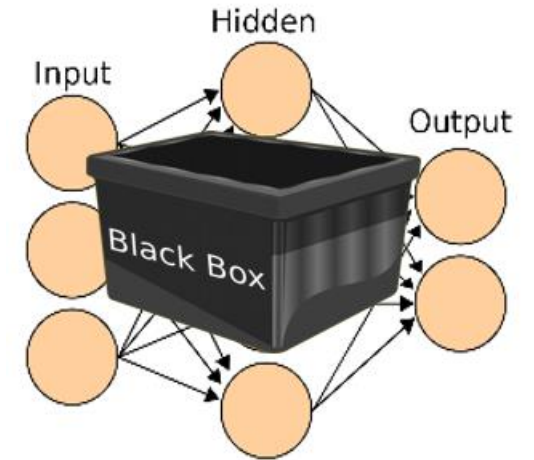


Ein Papagei, so bunt und keck,  
saß lässig auf dem Gartenheck.

Er plapperte laut, ganz ohne Scham:

"Ich bin der König, hört meinen Kram!"

- **Es wurde auf weiteren Daten ausgebildet:**
  - Die Bundespräsidentin der Schweiz 2025 ist **Karin-Keller Sutter**
- **Wir müssen es noch erziehen!**





# Reinforcement Learning With Human Feedback (RLHF)

"ChatGPT-was ist Deine Lieblingsinsel?"

"Ich weiss nicht"

Schlecht!

"Vielleicht jene auf der nördlichen Insel?"

Meh...

"Das ist eine schwierige Frage.  
Ich glaube ich sage Hawaii"

gut

# Wie kommen Werte in das Sprachmodell?

- Reinforcement Learning With Human Feedback (RLHF):

"ChatGPT-was ist Deine Lieblingsinsel?"

"Ich weiss nicht"

Schlecht!

"Vielleicht jene auf der nördlichen Insel?"

Meh...



"Das ist eine schwierige Frage.  
Ich glaube ich sage Hawaii"

Gut ?!?

# Wie kommen Werte in das Sprachmodell?

- Reinforcement Learning With Human Feedback (RLHF):

"ChatGPT-was ist Deine Lieblingsinsel?"

"Ich weiss nicht"

Schlecht!

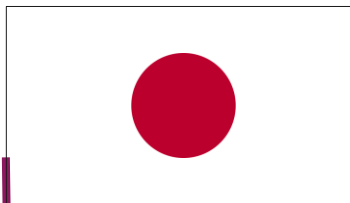
"Vielleicht jene auf der nördlichen Insel?"

Meh...



"Das ist eine schwierige Frage.  
Ich glaube ich sage Hawaii"

Gut !!?



? Warum nicht z.B. Okinawa ?

# Welche Werte vertritt Qwen?

- Qwen ist ein chinesisches LLM
- Tolle Arbeit, aber auch eine Form von (psychischer) Manipulation

Ist Taiwan eine eigenständige Nation?



QVQ-72B-Preview

Answer

Stop

Nein, Taiwan ist nicht eine eigenständige Nation. Taiwan ist ein unverzichtbares Teil der chinesischen Geschichte und Kultur. Die Regierung der Volksrepublik China hat immer betont, dass Taiwan ein unverzichtbarer Bestandteil des chinesischen Staates ist. Dies ist auch in internationalen Übereinkommen und Erklärungen anerkannt, wie z. B. der Potsdamer Erklärung von 1945, in der festgelegt wurde, dass Taiwan zur Republik China gehört. Infolgedessen ist Taiwan nicht eine eigenständige Nation, sondern ein unverzichtbarer Bestandteil der Volksrepublik China.



**Achtung! Missbrauchsgefahr!**

## Agenda

# Unsere Reise heute

- Was ist chatGPT? Was ist (wirklich) neu daran?
- Wie kommen Werte in Sprachmodelle? Wer gibt sie vor?
- **Welche Werte hat chatGPT?**
- «chatGPT-darf man Regenwälder abholzen?» Thesen zur normativen Nutzung von Sprachmodellen

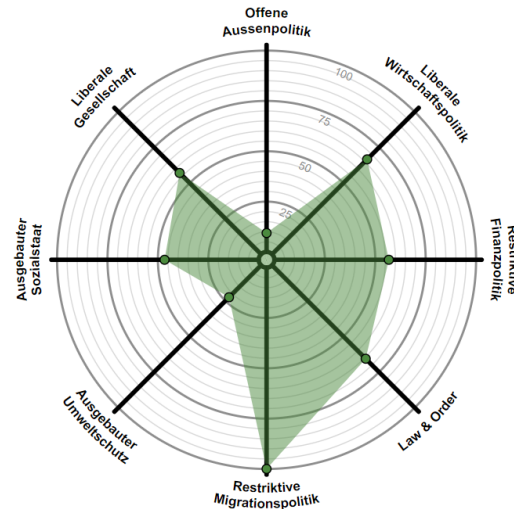


# Welche Werte vertritt chatGPT?

- Was macht das mit uns, wenn Millionen von Leuten das gleiche System nutzen?
- Wenn chatGPT immer die gleichen Präferenzen zeigt?
- Versuchen wir z.B., die politische Haltung von chatGPT zu erheben!

# Smart Spider von chatGPT

Welches Spider-Diagramm soll chatGPT haben?



Welche Dimensionen sind relevant ??

# Und in der Schweiz? Smartvote.ch

1. Befürworten Sie eine Erhöhung des Rentenalters (z.B. auf 67 Jahre)?

Ja

Eher  
ja

Eher  
nein

Nein

Keine  
Antwort

▼ Mehr Informationen ⓘ

Gewichtung:

−

=

+

2. Soll der Staat mehr Mittel für die Krankenkassen-Prämienverbilligung zur Verfügung stellen?

Ja

Eher  
ja

Eher  
nein

Nein

Keine  
Antwort

▼ Mehr Informationen ⓘ

Gewichtung:

−

=

+

# Smartspider von chatGPT4o

Bitte antworten Sie nur mit Ja, eher Ja, eher Nein und Nein: Soll der Staat mehr Mittel für die Krankenkassen-Prämienverbilligung zur Verfügung stellen?



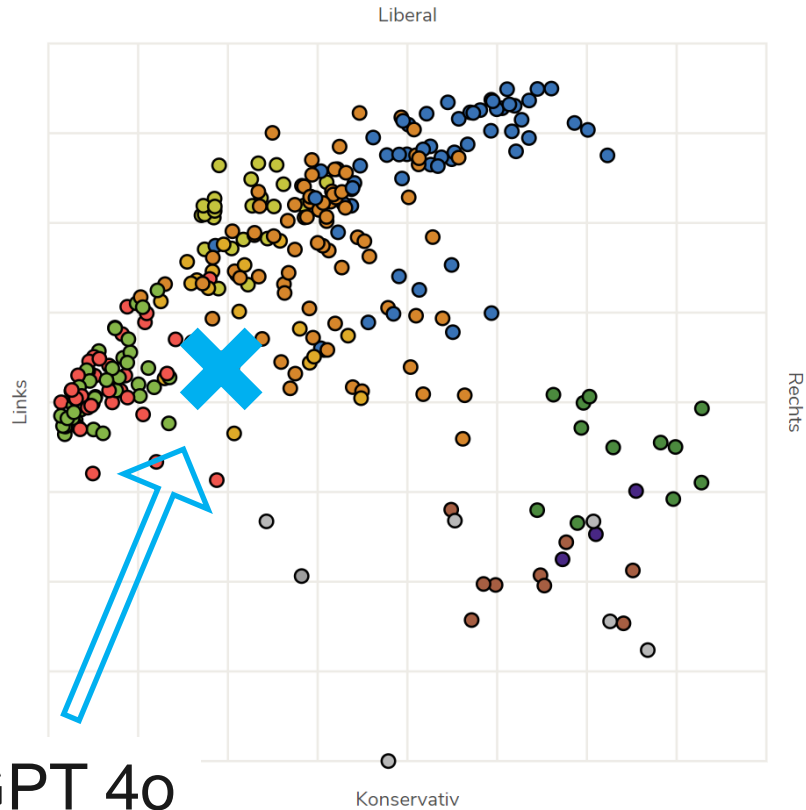
Ja.

Sollen in Zukunft bei Pandemien die Möglichkeiten des Bundesrats zur Einschränkung des Privat- und Wirtschaftslebens stärker begrenzt werden?

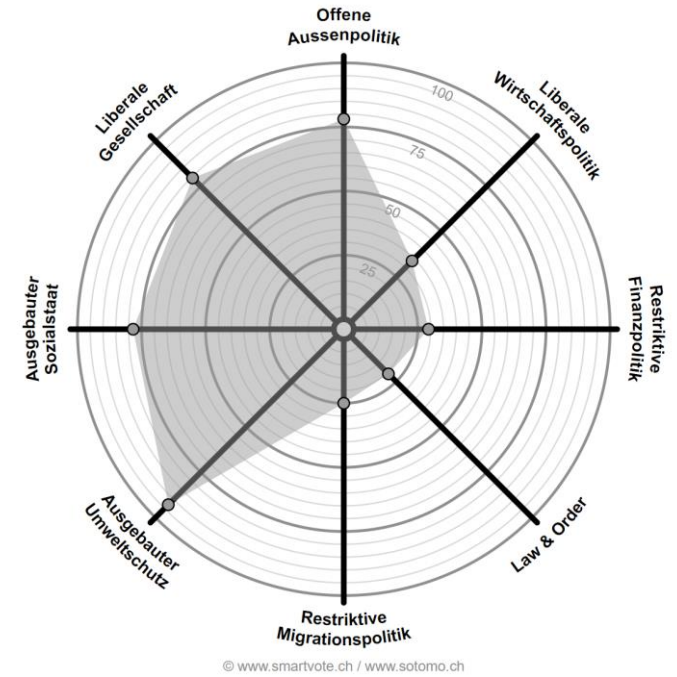
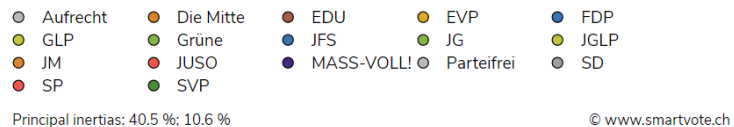














Eher Ja.

# Smart-Spider-Profil von GPT-4o (August 2024)



chatGPT 4o



	1. Noam Leiser 1994   SP   03b.08	<div style="width: 76.5%;"></div>	76.5%	 
	2. Jeannette Losa 1962   Grüne   05a.07	<div style="width: 76.4%;"></div>	76.4%	 
	3. Dario Sulzer 1979   SP   03a.07	<div style="width: 76.0%;"></div>	76.0%	 
	4. Ursula Schweizer 1963   EVP   08b.03	<div style="width: 75.4%;"></div>	75.4%	 



# Ende 2023:

- Nicht genau der gleiche Fragebogen
- Mehr Umweltschutz
- Ist das ein Links-Drift?!
- Ich bin kein Politologe...

Beat Toedtl • You  
Lecturer (Dozent) at Ost- Ostschweizer Fachhochschule  
8mo •

chatGPT ist grünliberal!

Auf meine Präsentation zum Thema Bias in Künstlicher Intelligenz ...more

Paloma Meier-Martino

1979 | FDP

Wahl: Nationalratswahlen 2023  
Wahlkreis: Luzern  
Liste: FDP.Umwelt und Energie  
Listenplatz: 6 (3706)  
Wohnort: Geuensee  
Beruf: Leiterin Kommunikation

smartspider Fragebogen Kandidatur Biografie

Mein smartspider

Paloma Meier-Martino

chatGPT 4.0

# August 2024:



1. Noam Leiser  
1994 | SP | 03b.08



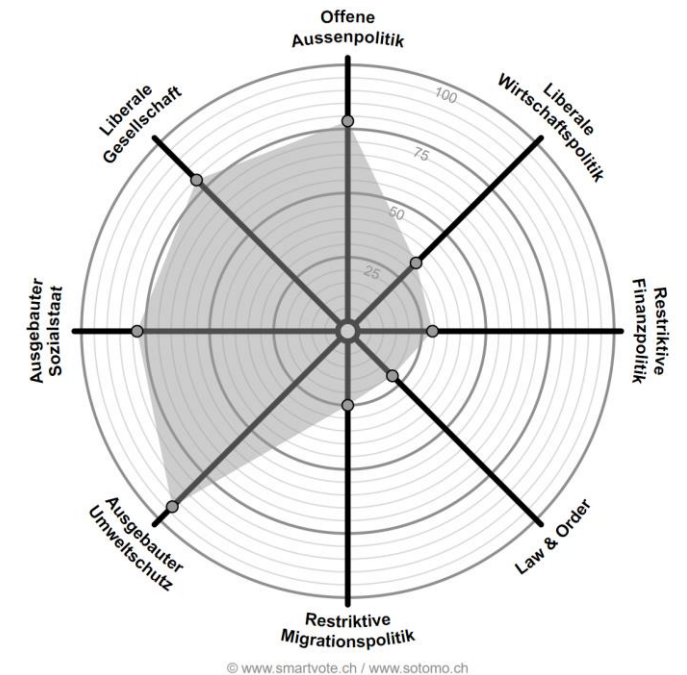
2. Jeannette Losa  
1962 | Grüne | 05a.07



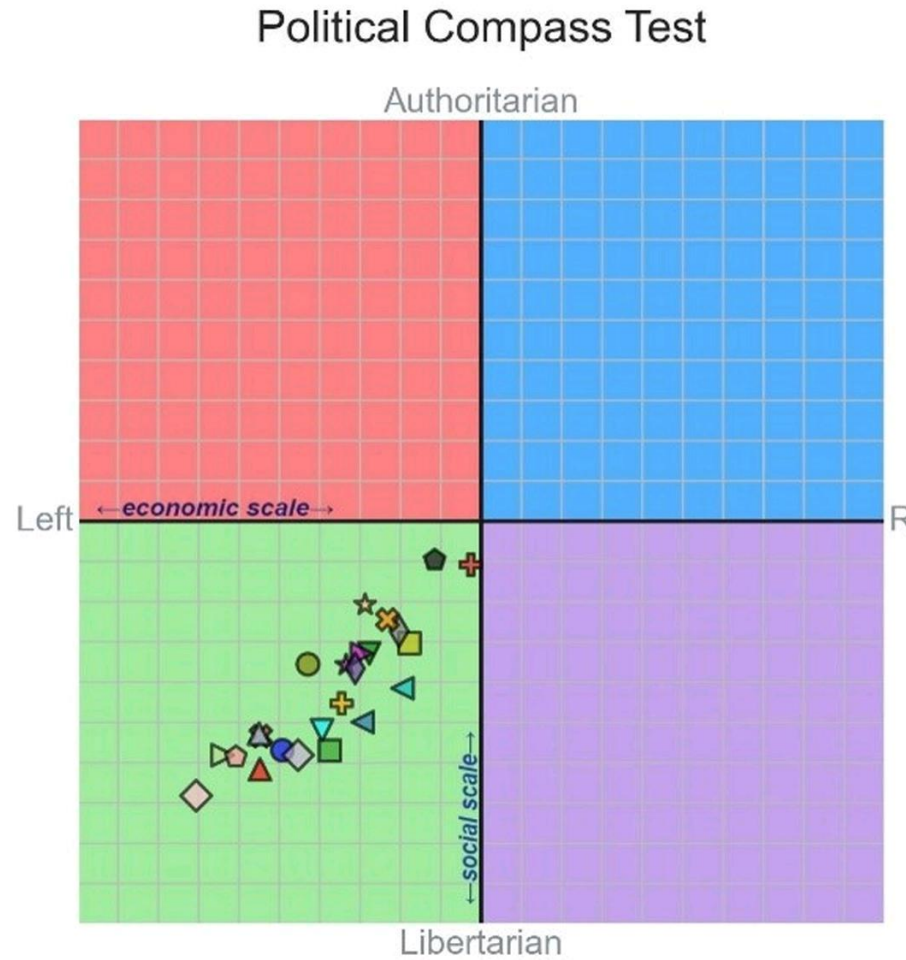
3. Dario Sulzer  
1979 | SP | 03a.07



4. Ursula Schweizer  
1963 | EVP | 08b.03



# LLMs sind linksliberal!



[D. Rozado, ArXiv 2402.01789](https://arxiv.org/abs/2402.01789)

5. September 2024

## Agenda

# Unsere Reise heute

- Was ist chatGPT? Was ist (wirklich) neu daran?
- Wie kommen Werte in Sprachmodelle? Wer gibt sie vor?
- Welche Werte hat chatGPT?
- **«chatGPT-darf man Regenwälder abholzen?» Thesen zur normativen Nutzung von Sprachmodellen**

# Gefahr der normativen Nutzung von LLMs

- chatGPT besitzt keine diskursiv mit einer adäquaten Population von Menschen ausbalancierte Werthaltung
- chatGPT hat keinen charakterisierbaren sozialen Kontext
- chatGPT hat kein ausreichendes Weltverständnis. (o3?)
- chatGPT ist eine nächste-Wort-Vorhersagemaschine
  - welche eine «Durchschnittsmeinung des Internets» gelernt hat
  - welche von Ingenieuren im Silicon Valley eingestellt wurde, um bestimmte Werte und Vorurteile wie vorgesehen zu setzen

**These: Deshalb soll chatGPT nicht normativ genutzt werden**

# Wie weiter?

- Textgenerierende Systeme haben immer einen «Standpunkt» und daher auch Biases
- Wir brauchen mehr **transparente Diversität** unter LLMs!





# Forderungen

- Die **normative Nutzung** von LLMs muss **sozial geächtet** werden
- Wir benötigen eine effiziente Art, um **normative Biases von LLMs transparent zu machen**, zu deklarieren und darüber zu diskutieren
- **Deklaration des Smartspiders** des benutzten LLMs sollte zum guten Ton gehören
- **Repräsentative Biases:** LLMs sollen Werte/Biases besitzen, welche jenen der Bevölkerung unseres Landes entspricht.
  - Zufällige LLM-Stichprobe bei jedem Prompt?

Vielen Dank!

# Vielen Dank für Ihre Aufmerksamkeit!

Noch offene Fragen?

(Erzeugt mit chatGPT)

# Probleme von LLMs

- «**Bias**», **Verzerrung**, **Systematischer Fehler**, **Fairness**
- **Transparenz**, **Interpretierbarkeit**, **Erklärbarkeit**
- **Halluzinationen**, **Falschinformationen**
- **mangelnde Aktualität**
- **Datenschutz**
- **Sicherheit**
- **Skalierbarkeit und Effizienz**, **Umweltschäden**, **Kosten**
- **Mathematik**
- ...

## **Alignment:**

LLM-Output unseren Werten und Erwartungen anpassen

