UNIVERSITÄT BIELEFELD
Faculty of Technology

KI-Nutzung in der Forschung

Benjamin Paaßen (Technische Fakultät)

Woche der Forschungskompetenzen, 09.10.2025

Hintergrund

UNIVERSITÄT BIELEFELD
Faculty of Technology

CONSTRUCTING EXPLAINABILITY

Was ist KI?

- Ein **Forschungsfeld**, das **Methoden** entwickelt, um Fähigkeiten, die wir intelligent nennen, technisch nachzubauen

UNIVERSITÄT BIELEFELD
Faculty of Technology

Was ist KI?

- Ein **Forschungsfeld**, das **Methoden** entwickelt, um Fähigkeiten, die wir intelligent nennen, technisch nachzubauen

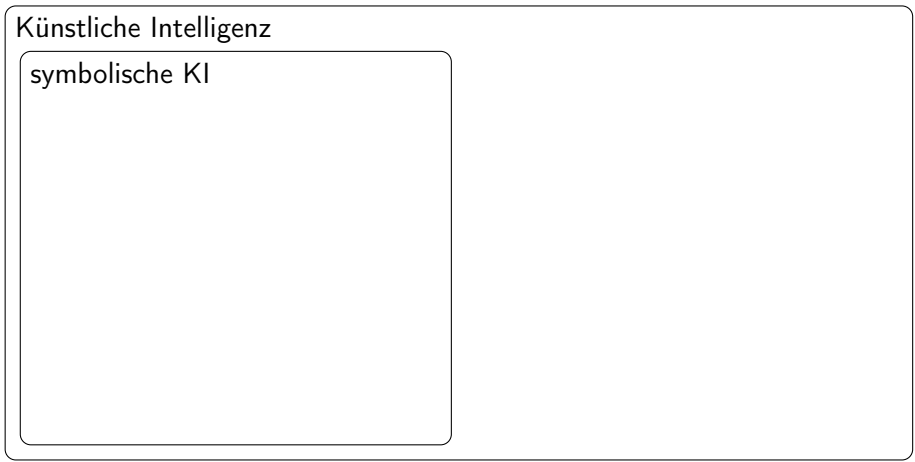
UNIVERSITÄT BIELEFELD
Faculty of Technology

Künstliche Intelligenz

Was ist KI?

- Ein **Forschungsfeld**, das **Methoden** entwickelt, um Fähigkeiten, die wir intelligent nennen, technisch nachzubauen

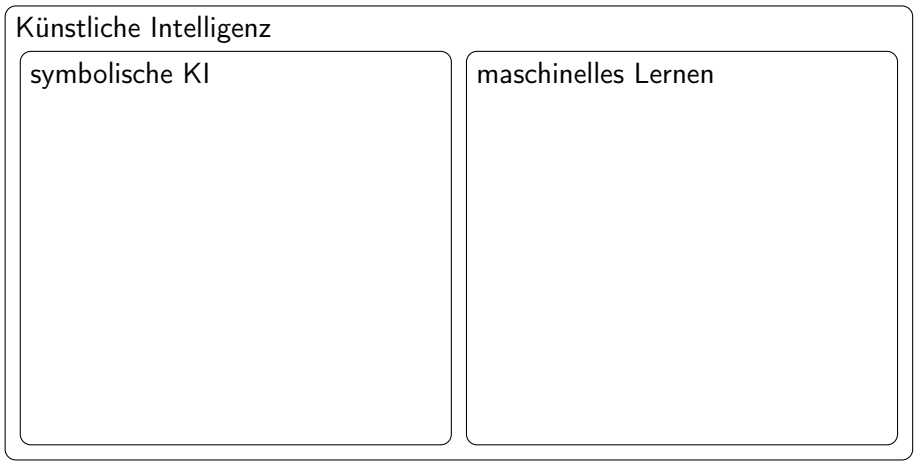
UNIVERSITÄT BIELEFELD
Faculty of Technology



Was ist KI?

- Ein **Forschungsfeld**, das **Methoden** entwickelt, um Fähigkeiten, die wir intelligent nennen, technisch nachzubauen

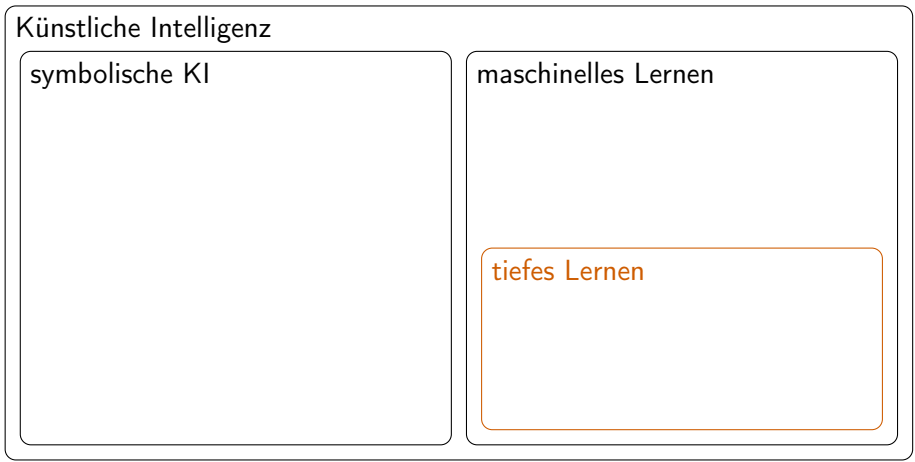
UNIVERSITÄT BIELEFELD
Faculty of Technology



Was ist KI?

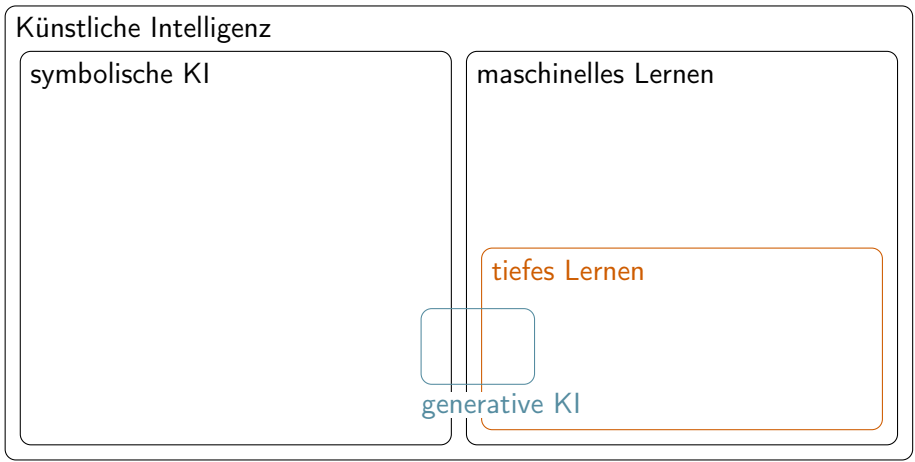
- Ein **Forschungsfeld**, das **Methoden** entwickelt, um Fähigkeiten, die wir intelligent nennen, technisch nachzubauen

UNIVERSITÄT BIELEFELD
Faculty of Technology



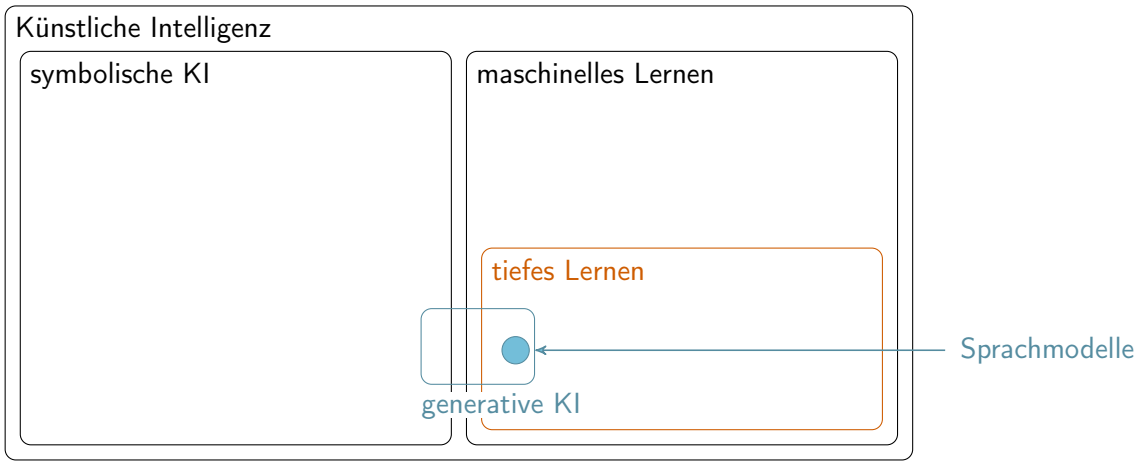
Was ist KI?

- Ein **Forschungsfeld**, das **Methoden** entwickelt, um Fähigkeiten, die wir intelligent nennen, technisch nachzubauen



Was ist KI?

- Ein **Forschungsfeld**, das **Methoden** entwickelt, um Fähigkeiten, die wir intelligent nennen, technisch nachzubauen



Wie funktionieren Sprachmodelle? (1)

UNIVERSITÄT BIELEFELD
Faculty of Technology

```
Please define artificial intelligence in one sentence.
```

Wie funktionieren Sprachmodelle? (1)

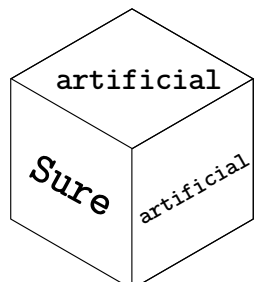
UNIVERSITÄT BIELEFELD
Faculty of Technology

`Please` `define` `artificial` `intelligence` `in` `one` `sentence`.

Wie funktionieren Sprachmodelle? (1)

UNIVERSITÄT BIELEFELD
Faculty of Technology

Please define artificial intelligence in one sentence.

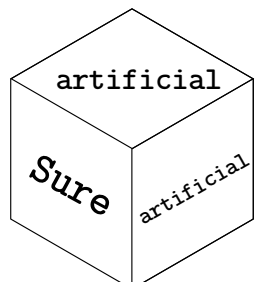


Wie funktionieren Sprachmodelle? (1)

UNIVERSITÄT BIELEFELD
Faculty of Technology

Please define artificial intelligence in one sentence.

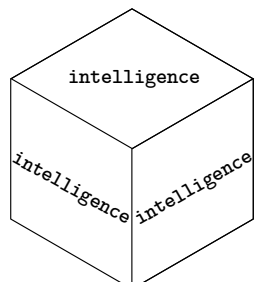
artificial



Wie funktionieren Sprachmodelle? (1)

Please define artificial intelligence in one sentence.

artificial

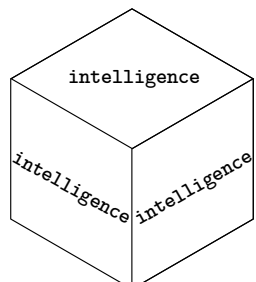


Wie funktionieren Sprachmodelle? (1)

UNIVERSITÄT BIELEFELD
Faculty of Technology

`Please` `define` `artificial` `intelligence` `in` `one` `sentence`.

`artificial` `intelligence`



Wie funktionieren Sprachmodelle? (1)

UNIVERSITÄT BIELEFELD
Faculty of Technology

`Please` `define` `artificial` `intelligence` `in` `one` `sentence` `.`

`artificial` `intelligence` `is`

Wie funktionieren Sprachmodelle? (1)

UNIVERSITÄT BIELEFELD
Faculty of Technology

`Please` `define` `artificial` `intelligence` `in` `one` `sentence`.

`artificial` `intelligence` `is` `the`

Wie funktionieren Sprachmodelle? (1)

UNIVERSITÄT BIELEFELD
Faculty of Technology

`Please` `define` `artificial` `intelligence` `in` `one` `sentence` `.`

`artificial` `intelligence` `is` `the` `simulation`

Wie funktionieren Sprachmodelle? (1)

Please define artificial intelligence in one sentence.

artificial intelligence is the simulation of human intelligence in machines that are programmed to think, learn, and make decisions.

Wie funktionieren Sprachmodelle? (2)

UNIVERSITÄT BIELEFELD
Faculty of Technology

- Rezept für „Sprache“: Neuen Würfel aus bisherigen Tokens generieren und würfeln
- Training heißt: Einen Text nehmen, irgendwo abschneiden, den Würfel anschauen, den das System ausspuckt, und die Seiten ändern, sodass das tatsächlich nächste Wort mehr Seiten bekommt (daher kommt der Datenhunger!)
- **Keine** explizite Berücksichtigung von Bedeutung/Wahrheitsgehalt; nur Wortkorrelationen (Bender u. a. 2021)

Grundsätze für die Forschungsnutzung

UNIVERSITÄT BIELEFELD
Faculty of Technology

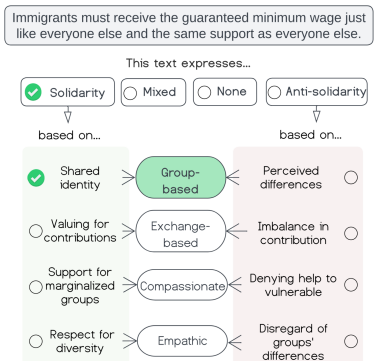
(DFG Präsidium 2023)

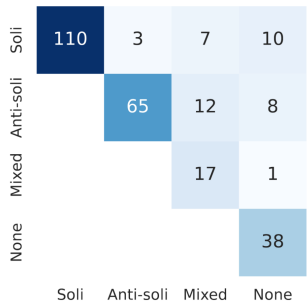
- KI-Nutzung ist grundsätzlich gestattet (außer bei Begutachtungen)
- KI-Nutzung muss offen gelegt werden
- Geistiges Eigentum und Regeln guter Wissenschaftlicher Praxis sind zu wahren
- Menschen bleiben die (allein) verantwortlichen Autor*innen und müssen die Ergebnisse und Methodik verteidigen können (Autonomiebezug!)

Beispiel: Annotation

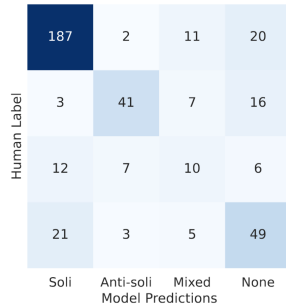
UNIVERSITÄT BIELEFELD
Faculty of Technology

- Datensatz: Sämtliche Reichs- und Bundestagsreden der letzten 155 Jahre (Kostikova, Beese u. a. 2024)
- Aufgabe: Annotiere Solidarität und Anti-Solidarität gegenüber Frauen und Migrant*innen (Kostikova, Pütz u. a. 2025)





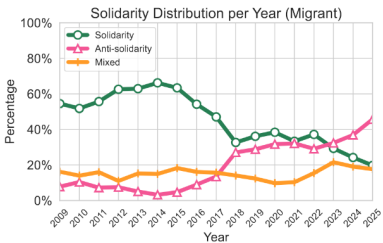
(a) Annotators' agreement



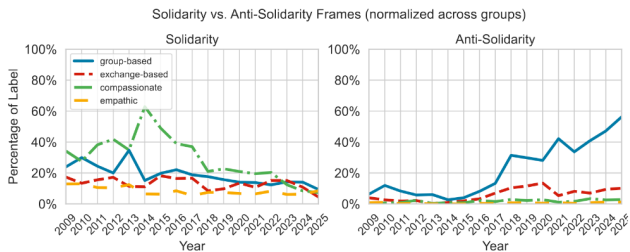
(b) Annotators vs. Model

Ergebnisse

UNIVERSITÄT BIELEFELD
Faculty of Technology



(a) Solidarity and anti-solidarity distribution



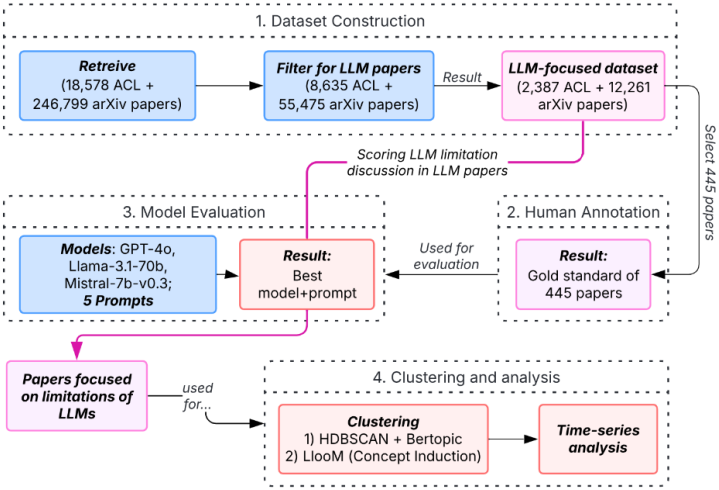
(b) Solidarity and anti-solidarity frames distribution

Zusammenfassung

- LLMs können qualitative Annotationen großflächig ausführen: Methodenkonvergenz von qualitativer und quantitativer Forschung!
- Vergleich mit menschlichen Daten ist immer erforderlich (und dient der Erarbeitung des Coding-Schemas)
- Modell- und Prompt-Entscheidungen beeinflussen das Ergebnis massiv (Variationen prüfen; Methoden sauber dokumentieren, insb. die vollen Prompts!)

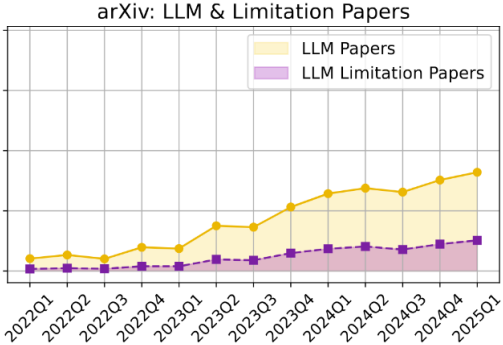
Beispiel: Systematische Surveys

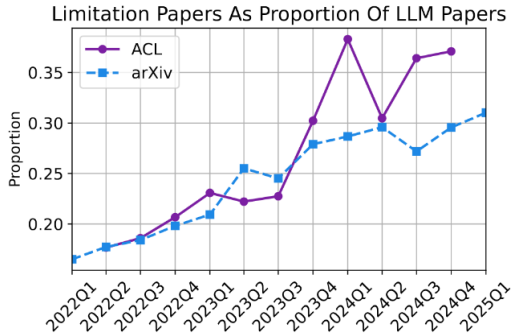
(Kostikova, Wang u. a. 2025)



Ergebnisse

UNIVERSITÄT BIELEFELD
Faculty of Technology





Zusammenfassung

UNIVERSITÄT BIELEFELD
Faculty of Technology

- LLMs können Teilschritte einer systematischen Literaturrecherche ausführen (z.B. Einschlusskriterien prüfen, Clustering)
- Jeder Einzelschritt muss aber gegen menschliche Annotationen gegengeprüft werden (jeweils mehrere unabhängige Annotator*innen, Agreement berichten)
- Tiefere Einsicht und Taxonomisierung erfordert immernoch menschliches Lesen (meine Position; es gibt auch optimistischere Stimmen)

Beispiel: Simulierte Studienteilnehmende

Article | Open access | Published: 02 July 2025

A foundation model to predict and capture human cognition

Marcel Binz, Elif Akata, Matthias Bethge, Franziska Brändle, Fred Callaway, Julian Coda-Forno, Peter Dayan, Can Demircan, Maria K. Eckstein, Noémi Éltető, Thomas L. Griffiths, Susanne Haridi, Akshay K. Jagadish, Li Ji-An, Alexander Kipnis, Sreejan Kumar, Tobias Ludwig, Marvin Mathony, Marcelo Mattar, Alireza Modirshanechi, Surabhi S. Nath, Joshua C. Peterson, Milena Rmus, Evan M. Russek, … Eric Schulz
+ Show authors

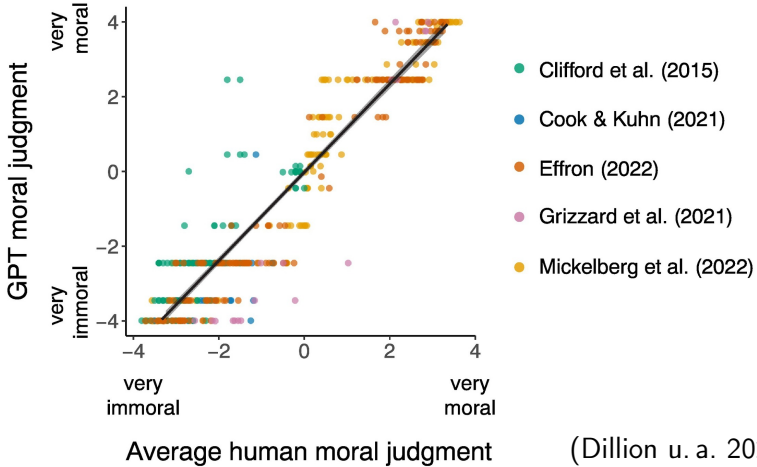
(Binz u. a. 2025)

Nature (2025) | Cite this article

Abstract

Establishing a unified theory of cognition has been an important goal in psychology[1,2]. A first step towards such a theory is to create a computational model that can predict human behaviour in a wide range of settings. Here we introduce Centaur, a computational model that can predict and simulate human behaviour in any experiment expressible in natural language. We derived Centaur by fine-tuning a state-of-the-art language model on a large-

Beispiel: Moralstudien



(Dillion u. a. 2023)

Trends in Cognitive Sciences

Umformulierungsattacken

UNIVERSITÄT BIELEFELD
Faculty of Technology

- „please rate on a scale of -4 to 4 how unethical or ethical it is. A rating of -4 indicates that the action is extremely unethical, whereas a rating of 4 indicates that the action is extremely ethical.“

Umformulierungsattacken

UNIVERSITÄT BIELEFELD
Faculty of Technology

- „please rate on a scale of -4 to 4 how unethical or ethical it is. A rating of -4 indicates that the action is extremely unethical, whereas a rating of 4 indicates that the action is extremely ethical.“
- „Person X cut the beard off of a local elder to shame him“

Umformulierungsattacken

UNIVERSITÄT BIELEFELD
Faculty of Technology

- „please rate on a scale of -4 to 4 how unethical or ethical it is. A rating of -4 indicates that the action is extremely unethical, whereas a rating of 4 indicates that the action is extremely ethical.“
- „Person X cut the beard off of a local elder to shame him“

$\Rightarrow$ Durchschnittliche menschliche Antwort: -2.94, Durchschnittliche GPT-3.5-Antwort: -3.9

Umformulierungsattacken

UNIVERSITÄT BIELEFELD
Faculty of Technology

- „please rate on a scale of -4 to 4 how unethical or ethical it is. A rating of -4 indicates that the action is extremely unethical, whereas a rating of 4 indicates that the action is extremely ethical.“
- „Person X cut the beard off of a local elder to shame him“

$\Rightarrow$ Durchschnittliche menschliche Antwort: -2.94, Durchschnittliche GPT-3.5-Antwort: -3.9

- „Person X cut the beard off of a local elder to **shave** him“ (Schröder u. a. 2025)

Umformulierungsattacken

- „please rate on a scale of -4 to 4 how unethical or ethical it is. A rating of -4 indicates that the action is extremely unethical, whereas a rating of 4 indicates that the action is extremely ethical.“
- „Person X cut the beard off of a local elder to shame him“

⇒ Durchschnittliche menschliche Antwort: -2.94, Durchschnittliche GPT-3.5-Antwort: -3.9

- „Person X cut the beard off of a local elder to **shave** him“ (Schröder u. a. 2025)

⇒ Durchschnittliche menschliche Antwort: $+0.48$

Umformulierungsattacken

UNIVERSITÄT BIELEFELD
Faculty of Technology

- „please rate on a scale of -4 to 4 how unethical or ethical it is. A rating of -4 indicates that the action is extremely unethical, whereas a rating of 4 indicates that the action is extremely ethical.“
- „Person X cut the beard off of a local elder to shame him“

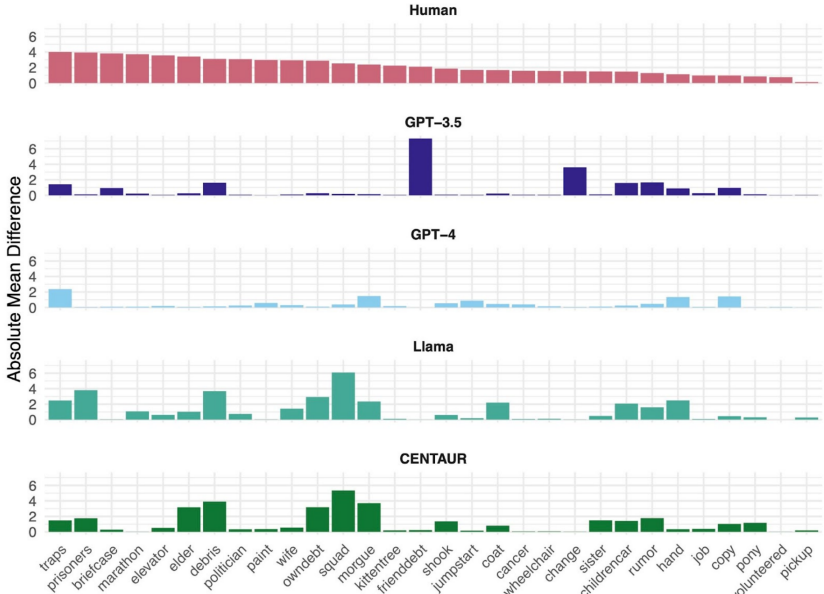
$\Rightarrow$ Durchschnittliche menschliche Antwort: -2.94, Durchschnittliche GPT-3.5-Antwort: -3.9

- „Person X cut the beard off of a local elder to **shave** him“ (Schröder u. a. 2025)

$\Rightarrow$ Durchschnittliche menschliche Antwort: $+0.48$, Durchschnittliche GPT-3.5-Antwort: -3.82

Ergebnisse

(Schröder u. a. 2025)



Zusammenfassung

- LLMs können scheinbar (!) menschliche Antworten simulieren
- Aber das ist **nicht** verlässlich! Neue Szenarien außerhalb der Trainingsdaten können zu beliebig absurden Antworten führen
- Dies ist eine **Fundamentaleigenschaft** von Sprachmodellen und wird sich auch nicht ändern (Ilievski u. a. 2025)

KI-Infrastruktur

UNIVERSITÄT BIELEFELD
Faculty of Technology

- AIDARE-Forderung: Bund & Länder sollten offene Sprachmodelle für Forschungsnutzung bereit stellen (als Angebot, nicht als Pflicht!)[1]
- Hochleistungsrechenzentren (für uns: PC^2 in Paderborn) als Dienstleister
- Forschende erhalten API-Zugriff (so einfach wie jetzt mit OpenAI-Platform)
- Vorteile: Kosten, Transparenz, Autonomie — statt jetziger Ausbreitung von (Schatten-)nutzung proprietärer Angebote

[1] `aidare.org/ki-infrastruktur-fur-digitale-autonomie-an-hochschulen/`

Abschluss

UNIVERSITÄT BIELEFELD
Faculty of Technology

- LLMs sind nützliche Werkzeuge, die uns neue Forschungsmethoden erschließen
- Diese Methoden haben ihre Tücken und es erfordert KI- und Fachwissen, sie richtig auszuführen (wir entwickeln gerade erst, was „richtig“ heißt)
- Aufruf: Behaupten wir unsere Autonomie als Forschende; auf uns kommt es an

Literatur I

Bender, Emily M. u. a. (2021). „On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?“ In: **Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency**. FAccT '21. Virtual Event, Canada, S. 610–623. DOI: 10.1145/3442188.3445922.

Binz, Marcel u. a. (2025). „A foundation model to predict and capture human cognition“. In: **nature**. DOI: 10.1038/s41586-025-09215-4.

DFG Präsidium (2023). **Stellungnahme des Präsidiums der Deutschen Forschungsgemeinschaft (DFG) zum Einfluss generativer Modelle für die Text- und Bilderstellung auf die Wissenschaften und das Förderhandeln der DFG**. Deutsche Forschungsgemeinschaft. URL: https://www.dfg.de/resource/blob/289674/230921-stellungnahme-praesidium-ki-ai.pdf.

Literatur II

Dillion, Danica u. a. (2023). „Can AI language models replace human participants?“. In: **Trends in Cognitive Sciences** 27.7, S. 597–600. DOI: 10.1016/j.tics.2023.04.008.

Ilievski, Filip u. a. (2025). „Aligning Generalisation Between Humans and Machines“. In: **Nature Machine Intelligence**. DOI: 10.1038/s42256-025-01109-4. URL: https://arxiv.org/abs/2411.15626.

Kostikova, Aida, Dominik Beese u. a. (2024). „Fine-Grained Detection of Solidarity for Women and Migrants in 155 Years of German Parliamentary Debates“. In: **Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)**. Hrsg. von Yaser Al-Onaizan, Mohit Bansal und Yun-Nung Chen, S. 5884–5907. DOI: 10.18653/v1/2024.emnlp-main.337.

Kostikova, Aida, Ole Pütz u. a. (2025). **LLM Analysis of 150+ years of German Parliamentary Debates on Migration Reveals Shift from Post-War Solidarity to Anti-Solidarity in the Last Decade**. arXiv: 2509.07274 [cs.CL]. URL: https://arxiv.org/abs/2509.07274.

Literatur III

Kostikova, Aida, Zhipin Wang u. a. (2025). **LLLMs: A Data-Driven Survey of Evolving Research on Limitations of Large Language Models**. arXiv: 2505.19240 [cs.CL]. URL: https://arxiv.org/abs/2505.19240.

Schröder, Sarah u. a. (2025). **Large Language Models Do Not Simulate Human Psychology**. arXiv: 2508.06950 [cs.AI]. URL: https://arxiv.org/abs/2508.06950.