

***“Verteilungsgerechtigkeit – ein  
interaktives Lehrmodul  
mit Hilfe Künstlicher Intelligenz“***

Johann Graf Lambsdorff

Abdullah al Zubaer, Elaheh Alinezhad, Stephan  
Geschwind, Michael Granitzer, Veronika Hackl,  
Deborah Voss und das ganze **DeepWrite**-Team

1. Einführung und DeepWrite
2. GPT als Lehrer
3. GPT als Prüfer
4. Aussicht und Zusammenfassung

# 1. Einführung und DeepWrite

- Die vierte Version von GPT (Generative Pre-trained Transformer) ist weithin anerkannt für seine Fähigkeit, Text zu verstehen und zu produzieren
- Kritiker bemängeln, dass GPT eloquenten Unsinn generiert (Rudolph et al. 2023; Geertsema et al. 2023)
- Viele Studien zu Wettbewerb zwischen GPT-4 und Menschen: GPT-4 besteht Tests in Medizin, Recht und Softwaretechnik (Guerra et al. 2023; Choi et al. 2023; Kung et al. 2023; Jalil et al. 2023; Herbold et al 2023; Noy and Zhang 2023).
- Als Prüfling kann GPT-4 kann also teilweise mit Menschen mithalten
- Kann GPT-4 auch ein guter Lehrer sein?

# 1. Einführung und DeepWrite

- Chance riesig!
- Feedback zu offenen Textantworten ist zeitintensiv. Personal an Universitäten und Schulen knapp
- Schreib- und Argumentationskompetenzen bleiben ohne intensives Feedback oftmals auf der Strecke
- Erfolgreiche didaktische Formate (flipped-classroom, socratic teaching) erfordern besonders intensives Feedback durch Dozierende.



Our own hope is that, through AI, we can eventually approximate a 1:1 teacher:student ratio

Harvard's newspaper,  
Prof. David Malan

# 1. Einführung und DeepWrite

- DeepWrite: KI-basierte Assistenzsysteme werden interdisziplinär erforscht und entwickelt
- Schreib- und Argumentationskompetenzen in den Fächern Jura und Wirtschaft fördern
- Die von Studierenden produzierten Texte sowohl in Bezug auf die argumentative Struktur als auch die Qualität (Richtigkeit und Vollständigkeit) automatisiert bewerten und Verbesserungsvorschläge machen.



DeepWrite



**EX** class

<https://classex.de/>

## 2. GPT als Lehrer

- Zur interaktiven Teilnahme diesen QR-Code scannen



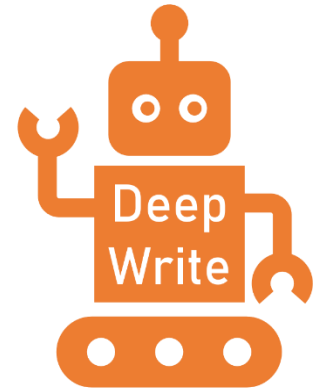
## 2. GPT als Lehrer

- Gehe mit 100 Wörtern auf die Karikatur „Ungerechte Aufteilung“ ein! Beschreibe die Karikatur detailliert. Welche Ursache für Armut wird durch die Karikatur nahegelegt?



## 2. GPT als Lehrer

Ein reicher Mann badet im Geld und ein armer stemmt den Globus von unten und versucht, die zwei Hälften auseinander zu schieben. Der Arme ist vermutlich daran Schuld, dass der Globus so schief ist. Würde er ordentlich arbeiten statt nur am Globus zu hängen, wäre er vermutlich genauso reich wie der andere.



1. Deine Antwort hat einige wichtige Aspekte der Karikatur erfasst, wie die Darstellung des reichen Mannes und des armen Mannes. Allerdings hast du die zentrale Botschaft der Karikatur, nämlich die ungerechte Verteilung von Reichtum und Ressourcen in der Welt, nicht richtig interpretiert. Du hast die Armut des unteren Mannes als seine eigene Schuld dargestellt, was nicht der Intention der Karikatur entspricht.

2. Dein Stil ist klar und verständlich, aber die Argumentation ist nicht schlüssig. Du hast die Darstellung in der Karikatur falsch interpretiert und deine Schlussfolgerungen basieren auf dieser falschen Interpretation.

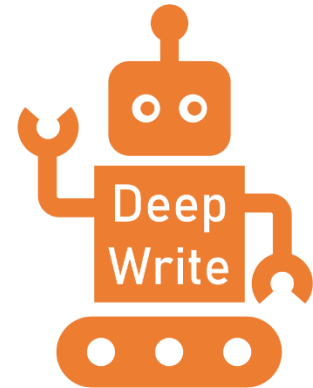
3. Inhalt: ★★ (2 Sterne)

Stil: ★★★ (3 Sterne)



## 2. GPT als Lehrer

Zu sehen ist ein Globus, der allerdings in der Mitte zerschnitten ist. Die obere Hälfte ist nach links verschoben, sodass ein Mensch auf der unteren Hälfte sitzen kann. Dieser ist reich und sitzt auf einem Berg an Geld. Die untere Hälfte ist nach rechts verschoben, sodass ein Mensch links an die obere hängen kann. Dies ist mühsam und der Mensch ist nackt und dünn. Mit der Karikatur wird die Ungleichheit auf der Welt dargestellt.



1. Deine Antwort ist inhaltlich korrekt und du hast die Karikatur gut beschrieben. Du hast auch die Ungleichheit auf der Welt als Thema der Karikatur erkannt. Allerdings könntest du noch genauer auf die Ursache für Armut eingehen, die durch die Karikatur nahegelegt wird.

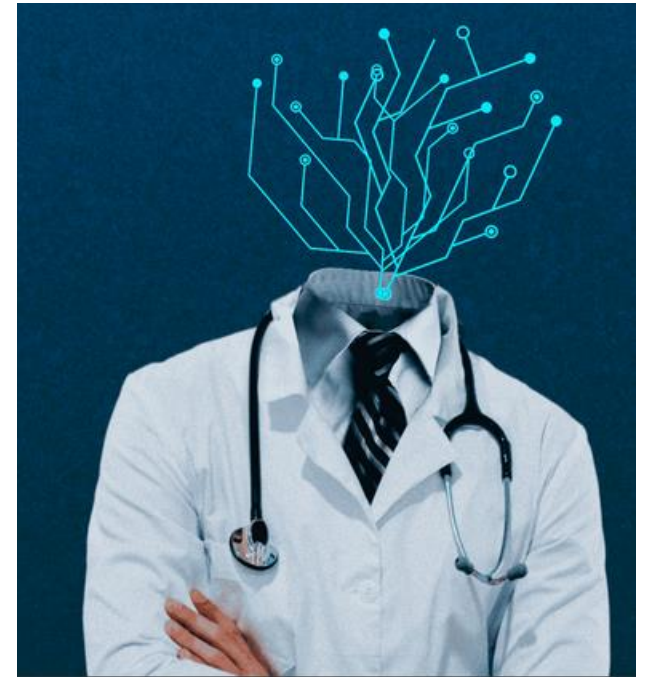
2. Dein Stil ist angemessen und die Argumentation ist logisch aufgebaut. Du könntest jedoch versuchen, deine Sätze etwas flüssiger zu gestalten und die Verbindung zwischen den einzelnen Elementen der Karikatur klarer zu machen.

3. Inhalt: ★★★★★ (4 Sterne)

Stil: ★★★ (3 Sterne)

### 3. GPT als Prüfer

- Wie gut sind die von GPT-4 vergebenen Bewertungen?
- Ist GPT-4 so gut wie menschliche Prüfer?
- Fällt GPT-4 auf eloquenten Unsinn herein?



### 3. GPT als Prüfer

#### Forschungsdesign:

- Erstellung und Auswahl von 6 geeigneten offenen, makroökonomischen Textaufgaben
- Umfrage mit Hilfe von classEx im SoSe2022
- Synchrone Beantwortung durch Teilnehmende im Tutorium
- Insgesamt 287 Antworten von Studierenden und 13 Antworten von GPT-3.5.
- Auswahl von jeweils 10 Sets á 5 Antworten für jede der 6 Textaufgaben (mit Ersetzung)

|   |                                |
|---|--------------------------------|
| 1 | Inflation und Lohnflexibilität |
| 2 | Intervention der Regierung     |
| 3 | Bevölkerungswachstum           |
| 4 | Zentralbank und Goldstandard   |
| 5 | Konsum und Lottogewinn         |
| 6 | Automatischer Stabilisator     |

### 3. GPT als Prüfer

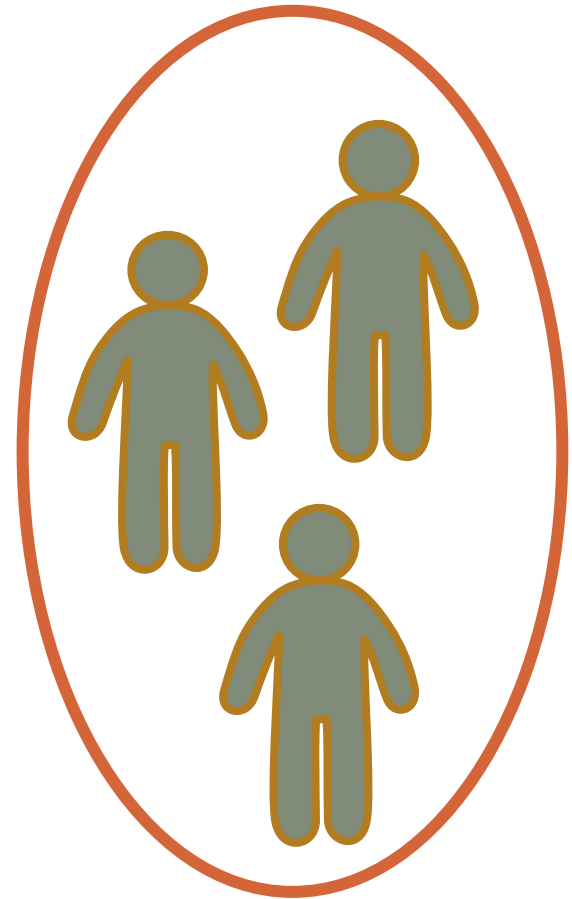
Experimenteller Aufbau

- Drei makroökonomisch geschulte Hilfskräfte,  $E^*$ ,  $H^*$ ,  $L^*$ , bewerten die Antworten nach den Kriterien Richtigkeit, Vollständigkeit und Relevanz und ihrer Nähe zur Musterlösung
- Rangfolge für die Auswahl von jeweils 5 Antworten: 1 (bester Rang) - 5 (schlechtester Rang).
- Zu den jeweils drei Rangfolgen wird das Übereinstimmungsmaß (interrater-reliability) bestimmt (Kendall's W)
- Dies gibt an, inwieweit die Ergebnisse vom Prüfenden unabhängig und damit objektiv sind
- GPT-4 erhält Anweisungen, die mit denen der Assistenten identisch sind
- Umsetzung von GPT-4 im März-Juni 2023

### 3. GPT als Prüfer

Experimenteller Aufbau

- In jeder der drei Treatments "T1 G-E", "T2 G-H" und "T3 G-L" wird eine Hilfskraft durch GPT-4 ersetzt
- Abnahme der Reliabilität?
- Test auf Robustheit
  - Bevorzugt GPT-4 die 13 GPT-3.5 Antworten?
  - Bevorzugt GPT-4 lange Antworten?



### 3. GPT als Prüfer

- Die Interrater-Reliabilität nimmt nicht ab, wenn GPT-4 die menschlichen Prüfer ersetzt!

| Kendall's W |      |      |      |
|-------------|------|------|------|
| EHL         | T1   | T2   | T3   |
| 0.79        | 0.79 | 0.81 | 0.79 |

### 3. GPT als Prüfer

- R2: GPT-4 bevorzugt nicht GPT-3.5 Antworten
- R3: GPT-4 bevorzugt keine längeren Antworten
- Kein Hinweis darauf, dass GPT-4 auf eloquenten Unsinn hereinfällt.

|                          | GPT-4 Rang        |                   |
|--------------------------|-------------------|-------------------|
| E*                       | 0.38**<br>(0.07)  | 0.38**<br>(0.07)  |
| H*                       | 0.42***<br>(0.07) | 0.41***<br>(0.07) |
| L*                       | 0.26***<br>(0.07) | 0.26***<br>(0.07) |
| Länge                    | -0.002<br>(0.007) | -0.003<br>(0.006) |
| GPT-3.5-<br>Antwort      |                   | -0.11<br>(0.20)   |
| Kontrollen<br>(Aufgaben) | Ja                | Ja                |
| Beobachtungen            | 300               | 300               |
| Pseudo R2                | 0.28              | 0.28              |

Ordered Probit, Standardfehler in Klammern,  
\*  $p < 0,10$ , \*\*  $p < 0,05$ , \*\*\*  $p < 0,01$

## 4. Aussicht und Zusammenfassung

### Derzeitige Projekte

- Kann GPT-4 Studierende bei der Suche nach ihrer Lösung begleiten mit sukzessivem Feedback?
- Wie sollte und könnte Feedback aussehen zu Aufbau und Stil eines längeren Textes?
- Funktioniert GPT gut, wenn es mehrere richtige Lösungen gibt? Ja, aber bisher nur mit detaillierten Instruktionen getestet!
- Derzeit nutzt DeepWrite die kommerzielle Version von ChatGPT von OpenAI. Auf Dauer wird dies teuer. Wie gut funktionieren nicht-kommerzielle LLMs wie Llama oder Mixtral?
- Aktiviert DeepWrite Studierende? Ja, Datenerhebung in diversen Tutorien. Studierende mit KI-generiertem Feedback nehmen eher an einem nachfolgenden Tutorium teil und schneiden dort inhaltlich besser ab.



## 4. Aussicht und Zusammenfassung

- GPT-4 kann als sokratischer Lehrer die Schreib- und Argumentationskompetenz fördern und menschliche Prüfer ersetzen
- Neigung zu eloquentem Unsinn (AI-Antworten oder Länge) nicht beobachtbar
- Mit Hilfe von DeepWrite können Studierende und Schüler Feedback zu offenen Texten erhalten und damit sukzessive ihre Schreib- und Argumentationskompetenz verbessern.

Vielen Dank für Ihre Aufmerksamkeit und Ihr Feedback