ChatGPT – Das Ende der Hausarbeit?

Dr. Armin Glatzmeier

Freie Universität Berlin, Universitätsbibliothek
Lehr- und Lernservices, Stabsstelle Kompetenzentwicklung wissenschaftliches Arbeiten

VERANSTALTET VON:

IM RAHMEN EINES PROJEKTES VON:

GEFÖRDERT VON:

Gliederung

- Warm up
- Was können LLMs?
- Workshopphase – LLMs in wissenschaftlichen Hausarbeiten
- Rules for tools – genKI und GwP
- Wie könnten die Rahmenbedingungen einer GwP-konformen Nutzung von LLMs aussehen?
- Grundregeln
- Erkennbarkeit KI-generierter Inhalte

Über mich …

Meine Erfahrungen

1. Erfahrungen aus eigener Nutzung
2. Erfahrungen im Lehr- und Betreuungskontext

https://miro.com/app/board/uXjVNt2kUL0=/?share_link_id=488211343662

Was können LLMs?

Input -> Blackbox -> Output

- ChatGPT ist wie alle modernen Large Language Models ein Transformermodell*, das von einer Texteingabe ausgehend Textoperationen durchführt
- Transformermodelle basieren auf einer neuronalen Netzwerkstruktur+
- Die Texterzeugung folgt einer Wahrscheinlichkeitsheuristik

* | Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser und Illia Polosukhin. „Attention Is All You Need“, 2017. https://doi.org/10.48550/ARXIV.1706.03762.

+ | Vgl. IBM. O.J. Was sind neuronale Netze? https://www.ibm.com/de-de/topics/neural-networks.

Input -> Blackbox -> Output

- Die Textproduktion ist i.d.R. nicht reproduzierbar (Ausnahme: Deterministische Modelle)
- Die Textproduktion beruht auf Wahrscheinlichkeit und wird durch die Trainingsdaten vordeterminiert (> Stichwort: Halluzinieren)
- Das auf eine konkrete Anfrage (Prompt) erwartbare Output wird durch den Prompt begrenzt (> Stichwort: Promptingstrategien)

Trainingsdaten – Was „weiß“ GPT-3?

Das Modell GPT-3 wurde mit folgenden Sammlungen trainiert*

Dataset	Quantity (tokens)	Weight in training mix	Epochs elapsed when training for 300B tokens
Common Crawl (filtered)	410 billion	60%	0.44
WebText2	19 billion	22%	2.9
Books1	12 billion	8%	1.9
Books2	55 billion	8%	0.43
Wikipedia	3 billion	3%	3.4

* | Brown, Tom B., Benjamin Mann, Nick Ryder, Melanie Subbiah et al. 2020. “Language Models are Few-Shot Learners”. *Arxiv* 2005.14165: 9; https://doi.org/10.48550/arXiv.2005.14165

Trainingsdaten – Was „weiß“ GPT-3?

Diese Datensätze enthalten+

- Webseiten
- Bücher und Artikel
- Inhalte aus Sozialen Medien, Blogs, Foren, Wikipedia usw.

+ | Rudolph, Jürgen, Samson Tan, and Shannon Tan. 2023. “ChatGPT: Bullshit spewer or the end of traditional assessments in higher education?” *Journal of Applied Learning & Teaching 6*(1): 3; https://doi.org/10.37074/jalt.2023.6.1.9

Trainingsdaten – Was „weiß“ GPT-3?

- Vortrainierte LLMs haben idR (noch) keine Internetanbindung (> Retrieval Augmented LLMs)
- Die Trainingsdaten sind idR bereinigt, um problematische Inhalte wie Gewalt, Vorurteile, Hate Speech etc. auszuschließen*

* | Perrigo, Billy. 2023. “The $2 Per Hour Workers Who Made ChatGPT Safer”. Time, 18.01.2023; https://time.com/6247678/openai-chatgpt-kenya-workers/

Trainingsdaten – Was „weiß“ GPT-3?

- Die Trainingsdaten enthalten ein umfangreiches Spektrum unterschiedlicher menschlicher Sprache
- Die Trainingsdaten allgemeiner LLMs haben keinen spezifischen wissenschaftlichen Zuschnitt
- Die Trainingsdaten können Fehler, Verzerrungen, Biases und Fehlrepräsentationen enthalten (und tun dies auch)
- Die Auswahl der Trainingsdaten und die Kriterien ihrer Bereinigung liegen in der ausschließlichen Hoheit der jeweiligen Anbieter

On Bullshit

“Bullshit is unavoidable whenever circumstances require someone to talk without knowing what he is talking about. Thus the production of bullshit is stimulated whenever a person’s obligations or opportunities to speak about some topic exceed his knowledge of the facts that are relevant to that topic.”*

* | Frankfurt, Harry G. 2005. *On Bullshit*. Princeton University Press, S. 63. https://doi.org/10.1515/9781400826537

On Bullshit

- LLMs haben kein Textverständnis
- LLMs haben keine Kenntnis oder ein Bewusstsein über die Welt
- LLMs sind Sprach- nicht Wissensmodelle
- Alle derzeit verfügbaren LLMs sind nicht spezifisch wissenschaftlich vortrainiert
- LLMs halluzinieren und erfinden Sachzusammenhänge, Informationen und Quellen
- LLM-generierte Texte sind keine wissenschaftlichen Quellen
  - ➢ Ungerechtfertigtes Vertrauen (Es ‚menschelt')

Workshopphase – LLMs in wissenschaftlichen Hausarbeiten

Board 1: Welche Kompetenzen setzt der Einsatz von LLMs für die wissenschaftliche Textproduktion voraus?

Board 2: Welche Kompetenzen sollen im Format Hausarbeit vermittelt werden?

Board 3: Welche Kompetenzen sollen durch das Format Hausarbeit geprüft werden?

Wie lassen sich LLMs sinnvoll für die wissenschaftliche Textproduktion nutzen?

https://miro.com/app/board/uXjVNtxRmo0=/?share_link_id=57188588092

Rules for tools – genKI und GwP

Welche Einsatzmöglichkeiten kennen Sie?

Und wie schätzen Sie diese ein?

https://miro.com/app/board/uXjVNtq4bM4=/?share_link_id=245120638657

Wissenschaftliches Fehlverhalten (DFG)

- Wissenschaftliches Fehlverhalten setzt einen vorsätzlichen oder grob fahrlässigen Verstoß gegen die Grundsätze der guten wissenschaftlichen Praxis voraus
- In den DFG-Leitlinien werden explizit drei Formen wissenschaftlichen Fehlverhaltens genannt*
  - Erfinden von Daten
  - Verfälschen von Daten
  - Plagiat

* | Deutsche Forschungsgemeinschaft. 2019. *Leitlinien zur Sicherung guter wissenschaftlicher Praxis: Kodex*. Bonn: DFG. https://doi.org/10.5281/zenodo.3923601

Wissenschaftliches Fehlverhalten (DFG)

- Wiss. Fehlverhalten ist ein personenbezogenes Konzept und setzt die Fähigkeit zur Übernahme von Verantwortung voraus
  - ➢ LLMs können sich entsprechend per definitionem nicht fehlverhalten
  - ➢ Für LLM-generierte Texte kann keine Autorschaft des LLMs angenommen
  - ➢ Daher auch nicht plagiatfähig

Gute wissenschaftliche Praxis

DFG-Leitlinie 14: Autorschaft

„Autorin oder Autor ist, wer einen genuinen, nachvollziehbaren Beitrag zu dem Inhalt einer wissenschaftlichen Text-, Daten- oder Softwarepublikation geleistet hat. […]. Sie tragen für die Publikation die gemeinsame Verantwortung, es sei denn, es wird explizit anders ausgewiesen.“*

- Generieren LLMs Fehlinformationen, Falschangaben oder (in seltenen Fällen) wörtliche Textplagiate liegt die Verantwortung bei der Person, die diese Texte verwendet (und den Co-Autor*innen)

* | Deutsche Forschungsgemeinschaft. 2019. *Leitlinien zur Sicherung guter wissenschaftlicher Praxis: Kodex*. Bonn: DFG. https://doi.org/10.5281/zenodo.3923601

Wie könnten die Rahmenbedingungen einer GwP-konformen Nutzung von LLMs aussehen?

https://miro.com/app/board/uXjVNtwAEZA=/?share_link_id=204022652827

Grundregeln

- Klare Vereinbarungen
- Einheitliche Vereinbarungen
- Dokumentieren Sie, was vereinbart wurde
- Studierenden rate ich, mit dem/der Betreuer*in zu besprechen:
    - Welche Tools wollen Sie verwenden?
    - Wozu wollen Sie diese verwenden?
    - Wie wird die Verwendung der Tools dokumentiert?

Grundregeln

Ist vollständige Transparenz gewünscht, sollten dokumentiert werden:

- Prompt
- Output
- Verwendung des Outputs
- Hersteller des LLMs
- Name des LLMs
- Version des LLMs

Erkennbarkeit KI-generierter Inhalte

- Erkennung KI-generierter Texte =|= Erkennung von Plagiaten
- Derzeit keine zuverlässigen Tools

- Hinweise auf genKI
  - Offenkundige Faktenfehler
  - Oberflächliche Darstellung
  - Falsche und/oder inexistente Quellen
  - Sprachliche und/oder stilistische Brüche

Ressourcen

Chicago, APA und MLA haben jeweils Vorschläge vorgelegt, wie KI-generierte Texte zitiert werden können

- CMOS: https://www.chicagomanualofstyle.org/qanda/data/faq/topics/Documentation.html
- APA: https://apastyle.apa.org/blog/how-to-cite-chatgpt
- MLA: https://style.mla.org/citing-generative-ai/

Vielen Dank für Ihre Mitarbeit!