Dr. Armin Glatzmeier
Freie Universität Berlin | Universitätsbibliothek

Freie Universität Berlin

Generative KI

Braucht die klassische Seminararbeit ein Update?

Universitäts Bibliothek

Gliederung

- Eine erste Annäherung …
- Impuls: 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

Welche Einsatzmöglichkeiten kennen Sie?

Und wie schätzen Sie diese ein?

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

https://miro.com/app/board/uXjVNiebjdk=/?share_link_id=596860138382

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

→ Die Textproduktion ist i.d.R. nicht reproduzierbar (→ Ausnahme: Deterministische Modelle)

→ Die Textproduktion beruht auf Wahrscheinlichkeit und wird durch die Trainingsdaten vordeterminiert (→ Stichwort: Halluzinieren | → Stichwort: Confirmation Bias)

→ Das auf eine konkrete Anfrage (Prompt) erwartbare Output wird durch den Prompt begrenzt (→ Stichwort: Promptingstrategien)

* | 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.

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

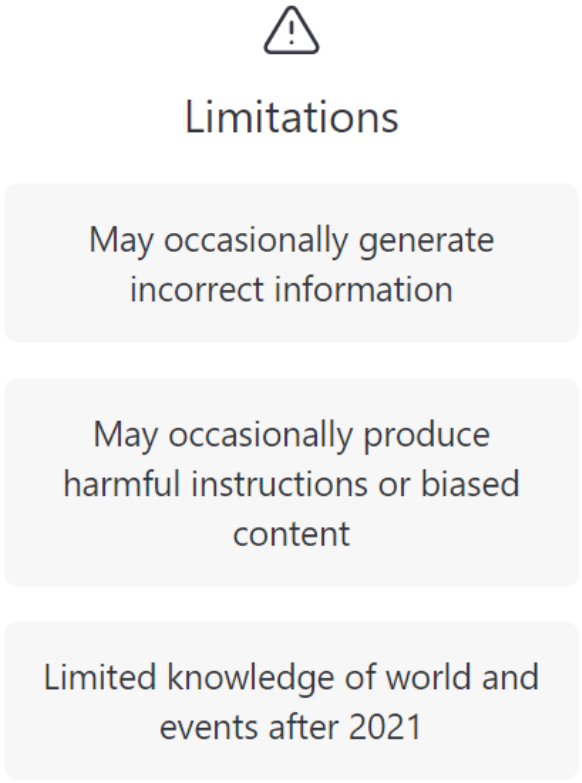
- Diese Datensätze enthalten+
  - Webseiten
  - Bücher und Artikel
  - Inhalte aus Sozialen Medien, Blogs, Foren, Wikipedia usw.

* | 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

+ | 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 keine Internetanbindung (aber: Retrieval augmented generation)
- → Die Trainingsdaten sind idR bereinigt, um problematische Inhalte wie Gewalt, Vorurteile, Hate Speech etc. auszuschließen*
  - → 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 Mißrepräsentationen enthalten (und tun dies auch)
  - → Die Auswahl der Trainingsdaten und die Kriterien ihrer Bereinigung liegen in der ausschließlichen Hoheit der jeweiligen Anbieter



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

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.”*

- LLMs haben kein Textverständnis
- LLMs haben keine Kenntnis oder ein Bewusstsein über die Welt
- LLMs sind Sprach- nicht Wissensmodelle (das ‚Wissen‘ entsteht eher beiläufig -> Probabilistik)
- 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‘)

→ Welche Folgen ergeben sich für die Hausarbeit?

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

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?

Board 1: https://miro.com/app/board/uXjVNidIPgg=/?share_link_id=453792443880
Board 2: https://miro.com/app/board/uXjVNid9PK0=/?share_link_id=866266115755
Board 3: https://miro.com/app/board/uXjVNidNick=/?share_link_id=680653254249

Wie lassen sich LLMs sinnvoll für wissenschaftliche Texte nutzen?

https://miro.com/app/board/uXjVNic7DxQ=/?share_link_id=452632027279

Rules for tools – genKI und GwP

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

- Wer verhält sich fehl?

→ Wiss. Fehlverhalten ist ein personenbezogenes Konzept und setzt die Fähigkeit zur Übernahme von Verantwortung voraus → LLMs können sich entsprechend per definitionem nicht fehlverhalten

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

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.“*

- Für LLM-generierte Texte kann keine Autorschaft des LLMs angenommen. → Daher auch nicht plagiatfähig
- Generieren LLMs Fehlinformationen, Falschangaben oder (in seltenen Fällen) wörtliche Textplagiate liegt die Verantwortung bei der Person, die diese Texte verwendet (und allen Mitautor*innen -> authors‘ contributions).

* | 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/uXjVNicKD4Q=/?share_link_id=870766598935

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?
- 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

- Die Erkennung von KI-generierten Texten unterscheidet sich fundamental von der Erkennung von Plagiaten

- Erkennungstools für KI-generierte Texte basieren auf Sprachmodellen die mit KI- und menschengeschriebenen Texten trainiert wurden.*

  OpenAI veröffentlichte im Januar 2023 einen AI-Classifier, der nach Herstellerangaben

  - 26% KI-geschriebenen Text korrekt
  - 9% menschlichen Text unzutreffend

  als "Likely AI-written" einstufte+

- Das Tool wurde zum 20.07.2023 deaktiviert, weil sich die Erkennungsleistung nicht verbessern ließ.

- VG München, Beschluss v. 28.11.2023 – M 3 E 23.4371 -> Anscheinsbeweis

* | Kirchner, Jan Hendrik, Lama Ahmad, Scott Aaronson, and Jan Leike. 2023. *New AI Classifier for indicating AI-written Text*, https://openai.com/blog/new-ai-classifier-for-indicating-ai-written-text

\+ | https://platform.openai.com/ai-text-classifier

Erkennbarkeit KI-generierter Inhalte

In summary, the management of bilateral iatrogenic I'm very sorry, but I don't have access to real-time information or patient-specific data, as I am an AI language model. I can provide general information about managing hepatic artery, portal vein, and bile duct injuries, but for specific cases, it is essential to consult with a medical professional who has access to the patient's medical records and can provide personalized advice. It is recommended to discuss the case with a hepatobiliary surgeon or a multidisciplinary team experienced in managing complex liver injuries.

Raneem Bader, Ashraf Imam, Mohammad Alnees, Neta Adler, Joanthan Ilia, Diaa Zugayar, Arbell Dan, Abed Khalaileh. 2024. Successful management of an Iatrogenic portal vein and hepatic artery injury in a 4-month-old female patient: A case report and literature review, *Radiology Case Reports* 19(6): 2106-2111, https://doi.org/10.1016/j.radcr.2024.02.037.

Recap: Welche Einsatzmöglichkeiten kennen Sie?

Und wie schätzen Sie diese ein?

https://miro.com/app/board/uXjVNicF92U=/?share_link_id=602384962846

Ressourcen

- Chicago, APA und MLA haben jeweils Vorschläge vorgelegt, wie KI-generierte Texte zitiert werden können
  - https://www.chicagomanualofstyle.org/qanda/data/faq/topics/Documentation.html
  - https://apastyle.apa.org/blog/how-to-cite-chatgpt
  - https://style.mla.org/citing-generative-ai/
- VG München, Beschluss v. 28.11.2023 – M 3 E 23.4371
  - https://www.gesetze-bayern.de/Content/Document/Y-300-Z-BECKRS-B-2023-N-42327
- DFG zum Umgang mit generativen KI-Modellen
  - https://www.dfg.de/de/service/presse/pressemitteilungen/2023/pressemitteilung-nr-39

Brauchen wir ein Update?

Brauchen wir ein Update?

- Kernkompetenzen der klassischen Seminararbeit bleiben weiterhin relevant (Fachwissen, Methodenkenntnis usw.)
- Klare, transparente und kommunizierte Regeln sofern die Nutzung von genKI-Tools erlaubt wird
- Vermittlung von Kenntnissen über die Funktionsweise von genKI-Tools (Studierende und Lehrende)
- Vermittlung datenschutz- bzw. nutzungsrechtlicher Probleme bei der Nutzung von genKI-Tools
- Vermittlung eines realistischen Erwartungshorizonts über die Möglichkeiten eines Einsatzes von genKI-Tools im wiss. Schreibprozess
- Schärfen des Problembewußtseins für andere problematische Aspekte (propietäre Modelle, Herkunft u. Kuratierung d. Trainingsdaten, Bereinigung der Trainingsdaten, Ressourcen, Nachhaltigkeit usw.)

Vielen Dank für Ihre Mitarbeit

"The author of an 'artificially intelligent' program is [...] clearly setting out to fool some observers for some time. His success can be measured by the percentage of the exposed observers who have been fooled multiplied by the length of time they have failed to catch on. Programs which become so complex (either by themselves, e.g. learning programs, or by virtue of the author's poor documentation and debugging habits) that the author himself loses track, obviously have the highest IQ's."

Joseph Weizenbaum. 1962. How to make a computer program appear intelligent. *Datamation* 8(2): 24-26 [24]