Armin Glatzmeier
Freie Universität Berlin | Universitätsbibliothek

Freie Universität Berlin

ChatGwP

genKI und gute wissenschaftliche Praxis

Universitäts Bibliothek

Gliederung

- Was können LLMs?
- LLMs, Autorschaft und Verantwortung
- Rules for tools – genKI und GwP

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 | → Persönlichkeits-, Urheber- und Lizenzrechte)

* | 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?

May 16, 2024

OpenAI and Reddit Partnership

We're bringing Reddit's content to ChatGPT and our products.

December 13, 2023

Partnership with Axel Springer to deepen beneficial use of AI in journalism

Axel Springer is the first publishing house globally to partner with us on a deeper integration of journalism in AI technologies.

https://openai.com/index/openai-and-reddit-partnership/
https://openai.com/index/axel-springer-partnership/

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

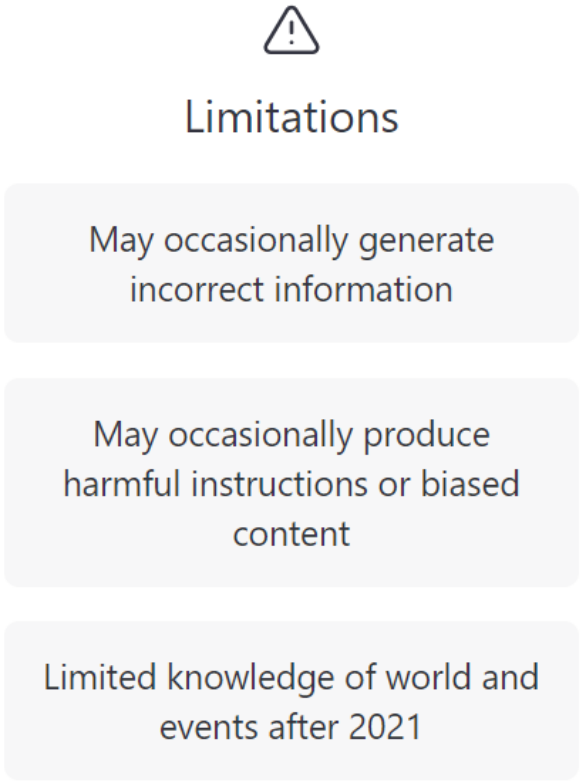
- Diese Datensätze enthalten+
  - Webseiten
  - Bücher und Artikel
  - Inhalte aus Sozialen Medien, Blogs, Foren, Wikipedia usw.

* | 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‘)

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

LLMs, Autorschaft und Verantwortung

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 werden. → 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 → author‘s contributins section)
- Urheberrechtlich geschützte Texte dürfen nicht ohne weiteres (per Prompting) an ein LLM übergeben werden

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

ChatGPT vs. Claude

OpenAI (ChatGPT)

Content

Your content. You may provide input to the Services ("**Input**"), and receive output from the Services based on the Input ("**Output**"). Input and Output are collectively "**Content**". You are responsible for Content, including ensuring that it does not violate any applicable law or these Terms. You represent and warrant that you have all rights, licences, and permissions needed to provide Input to our Services.

Ownership of content. As between you and OpenAI, and to the extent permitted by applicable law, you (a) retain your ownership rights in Input and (b) own the Output. We hereby assign to you all our right, title, and interest, if any, in and to Output.

Similarity of content. Due to the nature of our Services and artificial intelligence generally, Output may not be unique and other users may receive similar output from our Services. Our assignment above does not extend to other users' output or any Third Party Output.

Our use of content. We can use your Content worldwide to provide, maintain, develop, and improve our Services, comply with applicable law, enforce our terms and policies and keep our Services safe.

Opt out. If you do not want us to use your Content to train our models, you have the option to opt out by updating your account settings. Further information can be found in this Help Center article. Please note that in some cases this may limit the ability of our Services to better address your specific use case.

Anthropic (Claude)

We will not use your Inputs or Outputs to train our models, unless: (1) your conversations are flagged for Trust & Safety review (in which case we may use or analyze them to improve our ability to detect and enforce our Usage Policy, including training models for use by our Trust and Safety team, consistent with Anthropic's safety mission), or (2) you've explicitly reported the materials to us (for example via our feedback mechanisms), or (3) by otherwise explicitly opting in to training.

Our Privacy Policy explains your rights regarding your personal data, including with respect to our training activities. This includes your right to request a copy of your personal data, and to object to our processing of your personal data or request that it is deleted. We make every effort to respond to such requests. However, please be aware that these rights are limited, and that the process by which we may need to action your requests regarding our training dataset are complex.

Fall 1: Ein Fall für die Ombudsperson?

Eine Doktorandin, die für ihr kumulatives Dissertationsvorhaben einen Aufsatz bei einer Zeitschrift eingereicht hatte, wendet sich ratsuchend an Sie. Sie schildert Ihnen folgenden Sachverhalt: Ihr Beitrag sei abgelehnt worden. Im Peer-Review-Verfahren sei der Vorwurf erhoben worden, dass der Text in weiten Teilen mit Hilfe einer generativen KI geschrieben worden sei.

Die Doktorandin versichert, dass sie kein generatives KI-Tool zum Schreiben verwendet habe. Zudem sei sie in der Lage, verschiedene Versionen des Textes vorzulegen, die den Schreibfortschritt dokumentieren würden.

CAREER COLUMN | 05 February 2024

'Obviously ChatGPT' – how reviewers accused me of scientific fraud

A journal reviewer accused Lizzie Wolkovich of using ChatGPT to write a manuscript. She hadn't – but her paper was rejected anyway.

By E. M. Wolkovich

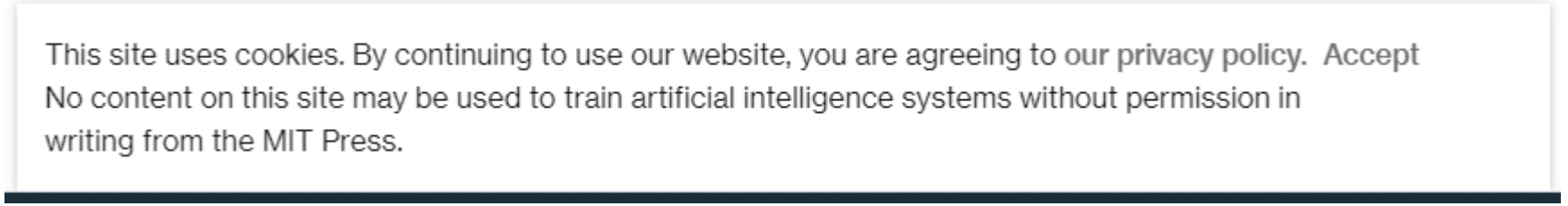
Ausgangsfall: https://www.nature.com/articles/d41586-024-00349-5

Fall 2: Sichtbarkeit? – Metriken (h-Index, JIF)

Geschafft, der gemeinsame Aufsatz wurde endlich zur Publikation angenommen. Nach der ersten Euphorie spaltet eine Frage das junge Autor*innenteam: Der Verlag fragt an, ob der Text auch zum Training für Large Language Models verwendet werden dürfe. Falls nicht, sei es kein Problem, einen entsprechenden Hinweis auf die Webseite zu setzen.

Ein Teil der Gruppe sieht in der Bereitstellung zum Training von LLMs eine Chance, die Sichtbarkeit der Publikation zu erhöhen. Zudem sei es ja auch im Interesse der Wissenschaft, dass die Ergebnisse möglichst frei zugänglich gemacht würden. Der andere Teil der Gruppe möchte nicht, dass ihr Text für das Training von kommerziellen Anwendungen verwendet wird.

Was nun?



Quelle: https://direct.mit.edu

Fall 3: Der unbekannte ‚Co-Autor‘

Helen, Tom und Kim sind in der gleichen Arbeitsgruppe und haben vielversprechende Ergebnisse. Diese wollen sie nun in einem gemeinsamen Paper veröffentlichen. Sie vereinbaren folgende Aufgabenverteilung: Helen kümmert sich um den Methodenteil und den Versuchsaufbau. Tom übernimmt die Datenanalyse und Interpretation. Kim, die einen guten Gesamtüberblick über das Projekt hat, verspricht, die Manuskriptfassung zu erstellen.

Helen schreibt einen Vorschlag für den Methodenteil und bereitet die Daten für die Publikation auf. Tom verschriftlicht die Ergebnisse der Datenanalyse und –interpretation und erstellt Grafiken und Tabellen.

Mit den Vorarbeiten von Helen und Tom promptet Kim ein Large Language Model und finalisiert so innerhalb eines Tages den Textentwurf, ohne das Ergebnis genauer zu sichten. Helen und Tom wissen nichts vom Einsatz eine genKI Tools und sind von Kims Entwurf begeistert. Alle einigen sich darauf, das Papier unter gemeinsamer Autorschaft zu veröffentlichen.

Kurz nach der Veröffentlichung erfahren Tom und Kim, wie die Einreichversion entstanden ist. Was ist zu tun …?

Neue Tools – neue Problemkonstellationen

- Neue Formen von Autorschaftskonflikten
- Konflikte über die (vermeintliche) Nutzung von genKI-Tools
- Datenschutz / Urheberrechtsschutz
- Schutz privater Daten (z.B. von Proband*innen)
- Inhaltliche Verantwortung für die Validität verwendeter Informationen

… und wie sieht das in der Praxis aus?

Ein Praxistest für Murphy's Law

ARTICLE INFO

Article history:
Received 23 November 2023
Revised 5 February 2024
Accepted 12 February 2024

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.

Ein Praxistest für Murphy's Law

This article has been removed at the request of the Editors-in-Chief and the authors because informed patient consent was not obtained by the authors in accordance with journal policy prior to publication. The authors sincerely apologize for this oversight.

In addition, the authors have used a generative AI source in the writing process of the paper without disclosure, which, although not being the reason for the article removal, is a breach of journal policy. The journal regrets that this issue was not detected during the manuscript screening and evaluation process and apologies are offered to readers of the journal.

Raneem Bader, Ashraf Imam, Mohammad Alnees, Neta Adler, Joanthan Ilia, Diaa Zugayar, Arbell Dan, Abed Khalaileh. 2024. REMOVED: 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, Nr. 8 (August 2024): 3598, https://doi.org/10.1016/j.radcr.2024.02.037.

Ein Praxistest für Murphy's Law

This article has been removed at the request of the Editors-in-Chief and the authors because informed patient consent was not obtained by the authors in accordance with journal policy prior to publication. The authors sincerely apologize for this oversight.

In addition, the authors ha
paper without disclosure,
is a breach of journal polic
the manuscript screening
the journal.

RADIOLOGY CASE REPORTS

Patient consent

Written informed consent was obtained from the patient's parents (patient's guardian) for publication of this case report and accompanying images.

REFERENCES

Ein Praxistest für Murphy's Law

This article has been removed at the request of the Editors-in-Chief and the authors because informed patient consent was not obtained by the authors in accordance with journal policy prior to publication. The authors sincerely apologize for this oversight.

In addition, the authors have used a generative AI source in the writing process of the paper without disclosure, which, although not being the reason for the article removal, is a breach of journal policy. The journal regrets that this issue was not detected during the manuscript screening and evaluation process and apologies are offered to readers of the journal.

Raneem Bader, Ashraf Imam, Mohammad Alnees, Neta Adler, Joanthan Ilia, Diaa Zugayar, Arbell Dan, Abed Khalaileh. 2024. REMOVED: 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, Nr. 8 (August 2024): 3598, https://doi.org/10.1016/j.radcr.2024.02.037.

Weitere Fälle finden Sie bei Guillaume Cabanac unter der Rubric „Suspect Phrases Detector" → https://www.irit.fr/~Guillaume.Cabanac/problematic-paper-screener

Rules for tools – genKI und GwP

Basics

- Bevor Sie genKI-Tools zur Erstellung eines Textes verwenden, klären Sie, ob dies zulässig ist und in welcher Form die Nutzung dokumentiert werden muss
  - Die meisten Verlage haben bereits Richtlinien (z.B.: https://www.nature.com/nature-portfolio/editorial-policies/ai)
- Wenn Sie genKI-Tools bei der Erstellung von Texten verwenden, sollten Sie die Verwendung für Ihre eigenen Unterlagen vollständig dokumentieren
  - Prompt
  - Output
  - Verwendung des Outputs
  - Hersteller des LLMs
  - Name des LLMs
  - Version des LLMs

Weiterführende 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 (Zulassung zum Masterstudium wg. mutmaßlicher Nutzung eines LLMs verweigert)
  - 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

Vielen Dank für Ihr Interesse

"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]