

# What is it like to be a bot?<sup>1</sup>

---

## Large Language Models und die gute wissenschaftliche Praxis

# Gliederung

1. Wo beginnen? – Academia est omnis divisa in partes tres ...
2. Text in progress ... – Wie erzeugen LLM eigentlich Texte und was „wissen“ sie?
3. „Come, give us a taste of your quality.“ – ChatGPT und die Kunst der Konfabulation und fehlerhafter Schlüsse
4. ChatGwP – LLMs und gute wissenschaftliche Praxis

# Academia est omnis divisa in partes tres ...

- Enthusiasmus – Erdrutschartige Veränderung wissenschaftlicher Arbeit (Zeitersparnis, Zugang zu Wissen, personalisierte Lernszenarien usw.)
  - Skepsis – Entwertung schriftlicher Prüfungsleistungen (Betrug, Autorschaftsfragen usw.)
  - Pragmatismus – Die Technologie ist nun da, also müssen wir damit umgehen
- Die Debatte bleibt leider oft auf ChatGPT verkürzt (obwohl es andere LLMs gibt und neue absehbar sind)
- Ein zentrales Augenmerk liegt auf KI-generiertem Text (obwohl es mehr Aspekte gibt, die Anlass zu Sorge oder Enthusiasmus geben könnten)

# Text in progress ...

Wie erzeugen LLM eigentlich Texte und was „wissen“ sie?

# Input, Output und etwas „magischer Feenstaub“<sup>2</sup>

- ChatGPT ist wie alle modernen Large Language Models ein Transformermodell\*, das von einer Texteingabe ausgehend Textoperationen durchführt
  - Transformermodelle basieren auf einer neuronalen Netzwerkstruktur<sup>+</sup>
  - 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)
  - Das auf eine konkrete Anfrage (Prompt) erwartbare Output wird durch das Prompt begrenzt (→ Stichwort: Promptingstrategien)

<sup>2</sup> | Debora Weber-Wulff über texterzeugende Machine-Learning-Tools auf dem Ombudssymposium 2023.

\* | 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<sup>3</sup>

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<sup>4</sup>
  - Webseiten
  - Bücher und Artikel
  - Inhalte aus Sozialen Medien, Blogs, Foren, Wikipedia usw.

3 | 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>

4 | 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, LMMs)
- Die Trainingsdaten sind idR bereinigt, um problematische Inhalte wie Gewalt, Vorurteile, Hate Speech etc. auszuschließen<sup>5</sup>
  - 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



## Limitations

May occasionally generate incorrect information

May occasionally produce harmful instructions or biased content

Limited knowledge of world and events after 2021

5 | Perrigo, Billy. 2023. "The \$2 per Hour Workers who made ChatGPT safer". *Time*, 18.01.2023; <https://time.com/6247678/openai-chatgpt-kenya-workers/>

**“Come, give us a taste of  
your quality.”**

ChatGPT und die Kunst der Konfabulation und fehlerhafter Schlüsse



AR

If Alice is not born in Germany can she be a German citizen?



No, if Alice is not born in Germany, she would typically not be a German citizen by birth.



German citizenship is primarily determined by jus sanguinis (right of blood), which means that one is usually a German citizen if at least one of their parents is a German citizen at the time of their birth. Being born in Germany to non-German parents can also grant one citizenship in certain circumstances.

If Alice is not born in Germany, she would need to go through the naturalization process to become a German citizen, which usually involves meeting specific residency requirements, demonstrating proficiency in the German language, and fulfilling other legal criteria set by the German government.

 Regenerate

# 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.”<sup>6</sup>

- 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 (, weil es ,menschelt‘)

# ChatGwP

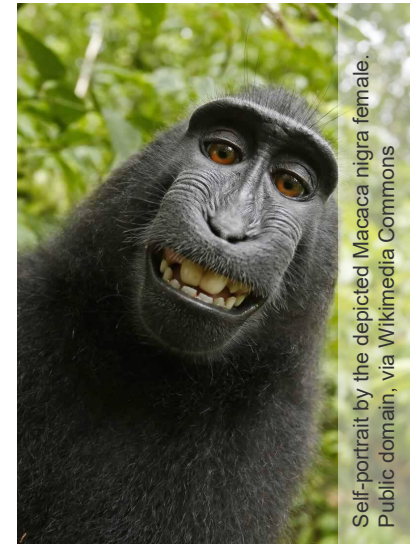
LLMs, gute wissenschaftliche Praxis und andere Problemfelder

# ChatGPT, 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.“<sup>7</sup>

- 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.



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

# ChatGPT, Autorschaft und Verantwortung

- USA: LLM-generierte Texte sind weder die Texte einer dritten Person noch als eigenes Werk urheberrechtlich geschützt (so die US Copyright Authority im Fall *Zarya of the Dawn*; diskutierbar)

## Verantwortung übernehmen

- Große Wissenschaftsverlage erstellten zeitnah Richtlinien zum Umgang und zur Verwendung KI-generierter Texte<sup>8, 9</sup>
  - Die Nutzung von LLMs oder ähnlichen Technologien muss ausgewiesen werden
  - Nutzen Autor\*innen LLM-generierte Inhalte tragen sie selbst die Verantwortung für die Korrektheit der darin enthaltenen Informationen und Aussagen

8 | Flanagin, Annette, Kirsten Bibbins-Domingo, Michael Berkwits, and Stacy L. Christiansen. "Nonhuman 'Authors' and Implications for the Integrity of Scientific Publication and Medical Knowledge". *JAMA*, 31. Januar 2023. <https://doi.org/10.1001/jama.2023.1344>

9 | Editorial. "Tools Such as ChatGPT Threaten Transparent Science; Here Are Our Ground Rules for Their Use". *Nature* 613, Nr. 7945 (26th January 2023): 612. <https://doi.org/10.1038/d41586-023-00191-1>

# Wissenschaftliches Fehlverhalten (DFG)

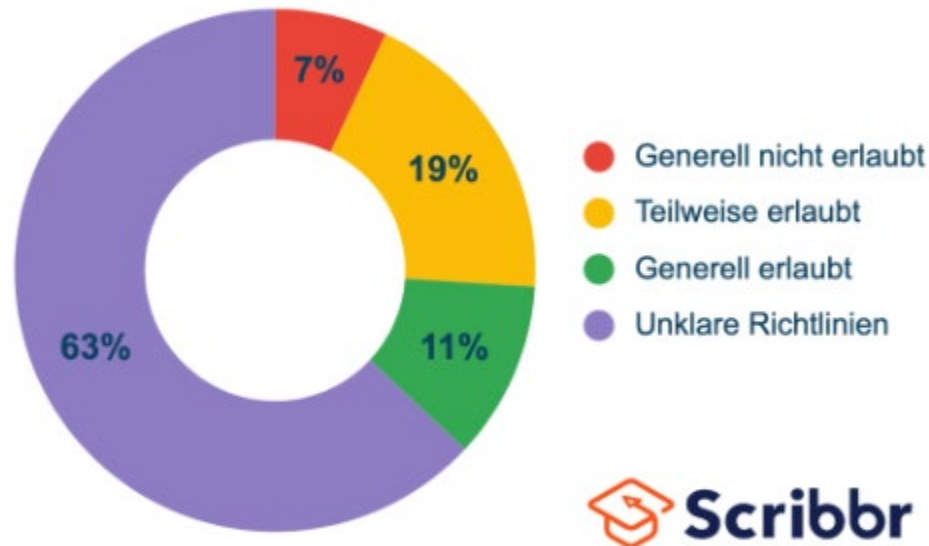
- Wissenschaftliches Fehlverhalten setzt einen vorsätzlichen oder grob fahrlässigen Verstoß gegen die Grundsätze der guten wissenschaftlichen Praxis voraus<sup>6</sup>
- In den DFG-Leitlinien werden explizit drei Formen wissenschaftlichen Fehlverhaltens genannt<sup>10</sup>
  - Erfinden von Daten
  - 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
- Generieren LLMs Fehlinformationen, Falschangaben oder (in seltenen Fällen) wörtliche Textplagiate liegt die Verantwortung bei der Person, die diese Texte verwendet

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

# Nutzung von KI-Tools an deutschen Hochschulen<sup>11</sup>

## 100 deutsche Hochschulen zu ChatGPT

Stand: 19. Juni 2023



[Zur Analyse – Google Spreadsheet](#)

# Zur Situation an der Freien Universität Berlin

- Eckpunkte zum Umgang mit KI-basierten Systemen und Tools in Studium und Lehre vom 10.05.2023
    - Über die grds. Zulässigkeit der Verwendung als „zugelassenes Hilfsmittel“ entscheidet der jeweilige Prüfungsausschuss
      - Daraus resultiert, dass die Verwendung solcher Hilfsmittel a) unter Vorbehalt steht und b) zwingend offengelegt werden muss
    - In diesem Rahmen liegt die Entscheidung darüber, ob und wenn ja, welche Tools verwendet werden dürfen bei der Person, die die Prüfungsleistung abnimmt
- Können wir überhaupt erkennen, ob ein Text mit KI-generiert wurde?



# 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.<sup>12</sup>

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<sup>13</sup>
- Das Tool wurde zum 20.07.2023 deaktiviert, weil sich die Erkennungsleistung nicht verbessern ließ.

12 | 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>

13 | <https://platform.openai.com/ai-text-classifier>

# Weitere Problemfelder

- DSGVO-Konformität (Pers. Daten, Sitzungsdaten)
- Nutzungs- bzw. Lizenzrechte die durch Prompting, Upload übertragen werden
- Trainingsdaten
  - Zusammensetzung (Biases, Fehlinformationen, Lücken usw.)
  - Herkunft (Urheberrecht, Nutzungsrechte)
    - Was bedeutet das z.B. für Open Access / Open Science?

This site uses cookies. By continuing to use our website, you are agreeing to our [privacy policy](#). [Accept](#)  
No content on this site may be used to train artificial intelligence systems without permission in writing from the MIT Press.



MIT Press Direct

- Bereinigung (Nach welchen Kriterien? Wer definiert diese?)
  - Überwiegend proprietäre Anbieter
  - Auslagerung der Klickarbeit (Neokolonialismus?)
- Exklusionfaktoren (Sprache, Zugang, Kosten, Marktmacht ...)
- Nachhaltigkeit (Energiebedarf, Hardware)

# 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.”

# Zitiervorschläge

# Eine Frage des Stils

- 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/>

Was soll dokumentiert werden?

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

# Sources in chronological order

- Nagel, Thomas. 1974. „What is it like to be a Bat?“ *The Philosophical Review* 83(4): 435-450.
- 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>
- Perrigo, Billy. 2023. “The \$2 Per Hour Workers Who Made ChatGPT Safer”. *Time*, 18.01.2023; <https://time.com/6247678/openai-chatgpt-kenya-workers/>
- Deutsche Forschungsgemeinschaft. 2019. *Leitlinien zur Sicherung guter wissenschaftlicher Praxis: Kodex*. Bonn: DFG. <https://doi.org/10.5281/zenodo.3923601>
- University of Oxford. 2023. *Plagiarism*, <https://www.ox.ac.uk/students/academic/guidance/skills/plagiarism>
- OpenAI. 2023. *ChatGPT General FAQ*, <https://help.openai.com/en/articles/6783457-chatgpt-general-faq>
- US Copyright Office. 2023. *Re: Zarya of the Dawn (Registration # VAU001480196)*, <https://copyright.gov/docs/zarya-of-the-dawn.pdf>
- OpenAI. 2022. *Documentation. Text Completion*, <https://platform.openai.com/docs/guides/completion/introduction>
- Flanagin, Annette, Kirsten Bibbins-Domingo, Michael Berkwitz, and Stacy L. Christiansen. “Nonhuman ‘Authors’ and Implications for the Integrity of Scientific Publication and Medical Knowledge”. *JAMA*, 31. Januar 2023. <https://doi.org/10.1001/jama.2023.1344>
- Editorial. “Tools Such as ChatGPT Threaten Transparent Science; Here Are Our Ground Rules for Their Use”. *Nature* 613, Nr. 7945 (26th January 2023): 612. <https://doi.org/10.1038/d41586-023-00191-1>
- 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>
- Haven, Tamarinde, Joeri Tjink, Brian Martinson, Lex Bouter, And Frans Oort. 2021. “Explaining Variance in Perceived Research Misbehavior: Results from a Survey Among Academic Researchers in Amsterdam”. *Research Integrity and Peer Review* 6(1): 7. <https://doi.org/10.1186/s41073-021-00110-w>
- Carl Sagan. 1997. *The demon-haunted World. Science as a Candle in the Dark*. New York: Random House
- Sattler, Sebastian und Martin Diewald. 2013. *FAIRUSE - Fehlverhalten und Betrug bei der Erbringung von Studienleistungen: Individuelle und organisatorisch-strukturelle Bedingungen*. <https://www.doi.org/10.2314/GBV:773897283>
- Frankfurt, Harry G. 2005. *On Bullshit*. Princeton University Press. <https://doi.org/10.1515/9781400826537>