Contextualizing Large Language Models – What are the social and ethical implications?

7 May 2024

HIDA Lecture Series on AI and Large Language Models

Steffen Albrecht
Office of Technology Assessment at the German Bundestag (TAB)
and Institute for Technology Assessment and Systems Analysis (ITAS)

Sam Altman
@sama

today we launched ChatGPT. try talking with it here:

chat.openai.com

Tweet übersetzen

8:38 nachm. · 30. Nov. 2022

Contextualizing LLMs – Overview

with regard to:

1. **Infrastructural contexts**

2. **Information and its context(s)**

3. **Context of invention**

and potential implications for development, regulation and use of LLMs

Considering these contexts helps to more realistically assess the potentials and pitfalls of current AI systems and to identify ways to improve them.

1. Infrastructural context of LLMs

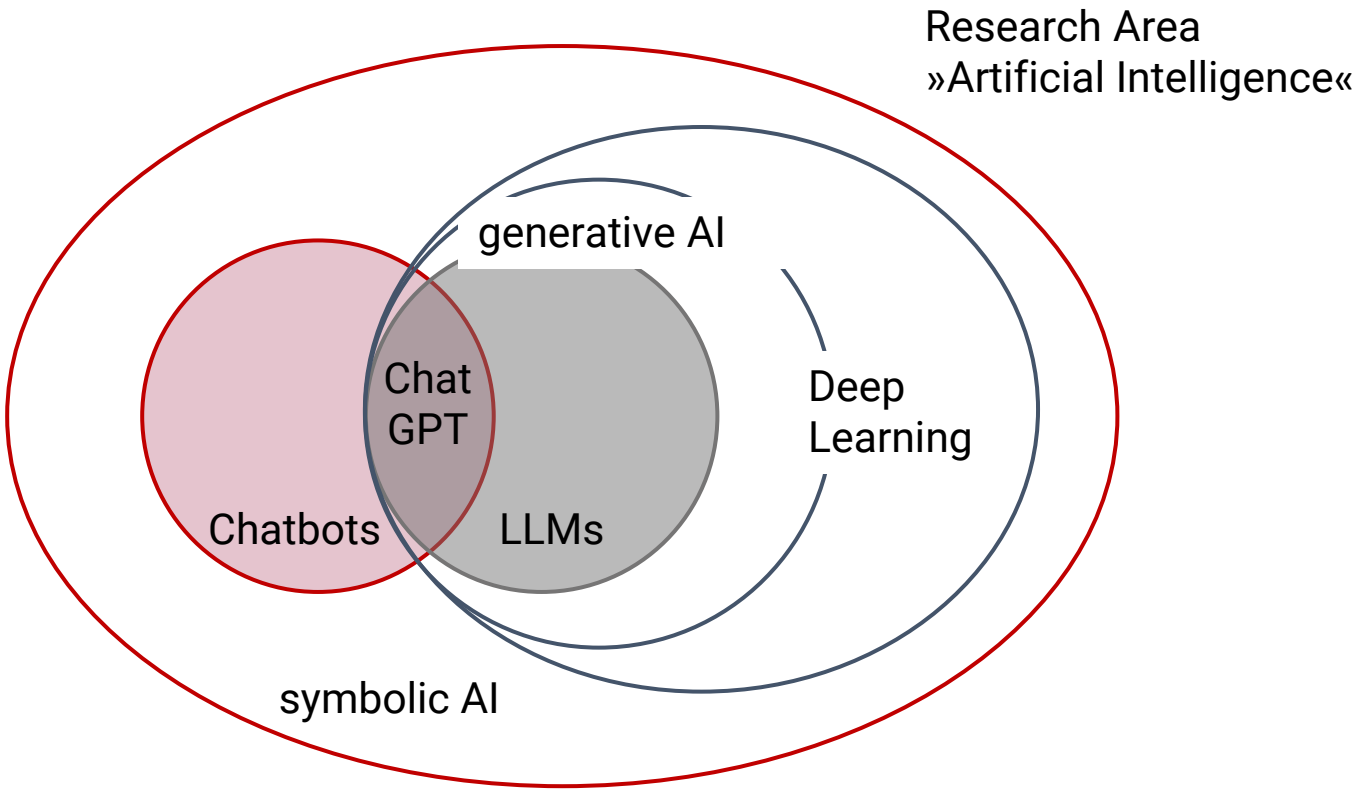
Infrastructural context of LLMs

- 14.03.2023 Release of GPT-4
- 30.11.2022 ChatGPT is published (100 Mio. active users in 2 months)

- 2018 First GPT model introduced
- 2017 Transformer architecture
- 2014 Attention mechanism
- 2012 Deep Learning for image recognition
- 2011 IBM Watson wins Jeopardy
- 1997 Long Short-Term Memory (LSTM) networks
- 1966 ELIZA chatbot
- …

»To appreciate the huge technological advance, we must consider the progress across several scientific disciplines during these decades, including theoretical foundations established, technological outputs, and the powerful interaction between these disciplines.« (Floyd 2023, p. 5)

Infrastructural context of LLMs



Infrastructural context of LLMs

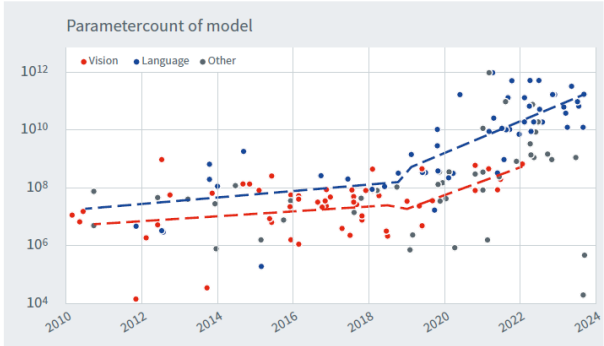
- AI-Systems: Algorithm, Infrastructure, Data, Use

- Infrastructure needs raise ecological and economic concerns

    - Energy and water consumption in training and inference phases
    - Raw materials for technology components
    - Growth of LLMs and in numbers of LLMs (Rohde et al. 2024; Patel/Ahmad 2023; Patel et al. 2024)

    - Questions arise whether costs of provision are in balance with rewards from use (Proschofsky/Der Standard 2024)
    - When are LLMs »hitting a wall« (Marcus 2022)?

Microsoft data center in West Des Moines, Iowa (Charlie Neibergall/AP Photo/The Des Moines Register)



Exponential growth of models (SustainAI/Rohde et al. 2024)

2. Information and its context(s)

Information and its context(s)

- AI-Systems: Algorithm, Infrastructure, Data, Use

- How is information processed by humans vs. LLMs? (Chemero 2023)

Humans	**LLMs**
embodied	multimodal, but isolated
motivated by needs	general purpose
socially oriented	no social relations
→ world context	→ abstract model

- Attributing information to a source establishes honesty and trust, opens up the possibility of questioning, commenting and correcting information (Ford 2023)

  »The knowledge embedded in ChatGPT responses is like ground meat: You cannot tell where it came from, and the process of obtaining it is not transparent.« (Floyd 2023, p. 18; »dead« vs. »alive« texts)

Information and its context(s)

- LLMs depend on the choice of the text corpus used for training (Floyd 2023)
- Problems: Black box, bias, and model collapse

- **Bias:** misrepresentation of categories such as race, gender, religious orientation, age etc. in the data
  - potentially leading to biased results
  - eventually resulting in discrimination

Machine translation used by immigration/criminial police (Guardian 2017, JT 2023)

Covert dialect prejudice in LLMs‘ judgements (Hofmann et al. 2024)

(Source: Gemini/narayan.somendra/medium.com)

- GPT-3 and ChatGPT have been found to be biased with regard to racializing stereotypes, gender and religious orientations, but the problem seems to be declining (Abid et al. 2021; Tamkin et al. 2021; Alba 2022; Biddle 2022; Borji 2023).

Information and its context(s)

- LLMs depend on the choice of the text corpus used for training (Floyd 2023)
- Problems: Black box, bias, and model collapse

- **Limits of Data:**

  »metrics and data are incomplete by their basic nature«
  (Nguyen 2024, p. 94)

  - Availability
  - Quantitative nature
  - Standardization
  - Categorization

  - Re-contextualize data
  - Awareness
    (countering automation bias)

  »Data is powerful but incomplete; don't let it entirely drown out other modes of understanding.« (Nguyen 2024, p. 101)

3. Context of invention

Context of invention

- Who is developing LLMs/AI systems?
  - »Teams of developers need to be more diverse« – male dominance (Eisner/Wiener Zeitung 2024)
  - Big tech firms dominate the development
  - Exploitation/reinforcement of global inequalities

- What happens with the data?
  - Data protection rules are often neglected
  - Data collection enhances oligarchy/market concentration

- Who is responsible?
  - Responsibility for sustainable business
  - Responsibility for results
  - Responsibility for impacts of use
  - Responsibility for regulation

TechCrunch/flickr (mod.), CC BY 2.0; TechCrunch/flickr, CC BY 2.0; Rory Arnold/No10 Downing Street - OGL 3; TIME 2006

Conclusions

Conclusions

- LLMs are not hermetic monoliths,
  but depend on a technological, informational and social context.

- Recognizing their de-contextualized appearance helps to more realistically assess their potentials and impacts.

- They are built by humans and can be controlled by humans.

- Humans should
  - be aware of the importance of context
  - build/extend competencies
  - work interdisciplinarily
  - develop/deploy/use AI systems responsibly.

To recognize the potentials and limits of current AI systems
and to identify ways to improve them,
we need to look beyond their technological functions
and consider the contexts of their development and use.

References

- Abid, A.; Farooqi, M.; Zou, J. 2021: Persistent Anti-Muslim Bias in LLMs. Proc. AIES '21. ACM, New York, NY, 298–306
- Alba, D. 2022: OpenAI Chatbot Spits Out Biased Musings, Despite Guardrails. 8.12.2022, Bloomberg (7.5.2024)
- Biddle, S. 2022: The Internet's New Favorite AI Proposes Torturing Iranians and Surveilling Mosques. 8.12.2022, The Intercept (7.5.2024)
- Borji, A. 2023: A Categorical Archive of ChatGPT Failures. arXiv
- Chemero. A. 2023: LLMs differ from human cognition because they are not embodied. Nature Human Behavior 7, 1828–1829
- Floyd, C. 2023: From Joseph Weizenbaum to ChatGPT. Weizenbaum Journal of the Digital Society 3(3)
- Ford, H. 2023: Is The Web Eating Itself? Seminar, UMass, 10.10.2023 (cited via Ethan Zuckerman, 7.5.2024)
- Guardian 2017: Facebook translates »good morning« into »attack them«, leading to arrest (A. Hern). 24.10.2017, The Guardian (7.5.2024)
- Hofmann, V.; Kalluri, P.R.; Jurafsky, D.; King, S. 2024: Dialect prejudice predicts AI decisions about people's character, employability, and criminality. arXiv

References (cont.)

- Japan Times 2023: AI's 'insane' translation mistakes endanger U.S. asylum cases. 24.9.2023, The Japan Times (7.5.2024)
- Marcus, G. 2022: Deep Learning Is Hitting a Wall. 10.3.2022, Nautilus (7.5.2024)
- Nguyen, C. Thi 2024: The Limits of Data. Issues in Science and Technology 40(2), 94–101
- Patel, D.; Ahmad, A. 2023: The Inference Cost Of Search Disruption – Large Language Model Cost Analysis. 9.2.2023, SemiAnalysis (7.5.2024)
- Patel, D.; Nishball, D.; Ontiveros, J.E. 2024: AI Datacenter Energy Dilemma - Race for AI Datacenter Space. 13.3.2024, SemiAnalysis (7.5.2024)
- Proschofsky, A. 2024: Der KI-Hype trifft auf die ersten Spuren von Realität. 20.4.2024, Der Standard (7.5.2024)
- Rohde, F. et al. 2024: Taking (policy) action to enhance the sustainability of AI systems. Institut für ökologische Wirtschaftsforschung (IÖW), Berlin
- Tamkin, A.; Brundage, M.; Clark, J.; Ganguli, D. 2021: Understanding the Capabilities, Limitations, and Societal Impact of Large Language Models. arXiv
- Weber, I.; Eisner, J. 2024: Eine Frauenquote für die KI-Forschung. 30.4.2024, Wiener Zeitung (7.5.2024)