Al in Healthcare Possible Perspectives for Improving the Quality of Medical Care

Prof. Dr. Janna Hastings

Medical Knowledge and Decision Support

Institute for Implementation Science in Health Care, Faculty of Medicine, University of Zurich

School of Medicine, University of St. Gallen





janna.hastings@uzh.ch



@jannahastings @jannahastings@mastodon.online https://hastingslab.org/



What is Artificial Intelligence?

Automated systems that are capable of performing tasks (behaving) in a way that would require human intelligence for humans to perform

Different technologies have different capabilities for how they are developed (learning from data or from human-encoded explicit knowledge) and how they can act (parsing language vs. images etc.)











What is Artificial Intelligence?



1990s

expert systems, knowledge standards, rules and dynamic logic-based inference



"black box", data bottleneck

deep learning with artificial neural networks, **big data**, growing compute resources

"good old-fashioned Al"

growing applications of datadriven supervised machine learning in imaging, pathology, discovery, diagnosis and prognosis



2000s

"traditional machine learning"

really big data, transfer learning, large-scale general-purpose generative pre-trained models for language, images and more



2017+

"black box" and unknown data









Language-based AI: A new frontier for automated systems



Image generated with Stable Diffusion (SD3 medium)

- Prior to 2022, no automated system existed with sufficient mastery of human languages to support medical knowledge management, patient query responses etc.
- The breakthrough advance in capabilities in Al-based large language models (LLMs) was due to congruent advances in model architectures, training methods, hardware and data availability
- Now we have many alternative models and tools with similar capabilities, both commercial and open-source









A huge range of applications of such models is foreseen in every field

communication Scientific text production Education Personalized Critical thinking Static knowledge Bias reproduction Misinformation

nature > communications medicine > perspectives > article

Perspective Open access | Published: 10 October 2023

The future landscape of large language models in medicine

Jan Clusmann, Fiona R. Kolbinger, Hannah Sophie Muti, Zunamys I. Carrero, Jan-Niklas Eckardt, Narmin Ghaffari Laleh, Chiara Maria Lavinia Löffler, Sophie-Caroline Schwarzkopf, Michaela Unger, Gregory P. Veldhuizen, Sophia J. Wagner & Jakob Nikolas Kather □

<u>Communications Medicine</u> **3**, Article number: 141 (2023) <u>Cite this article</u>

36k Accesses | 57 Citations | 52 Altmetric | Metrics

Abstract

Large language models (LLMs) are artificial intelligence (AI) tools specifically trained to process and generate text. LLMs attracted substantial public attention after OpenAI's ChatGPT was made publicly available in November 2022. LLMs can often answer questions, summarize, paraphrase and translate text on a level that is nearly indistinguishable from human capabilities. The possibility to actively interact with models like ChatGPT makes LLMs attractive tools in various fields, including medicine. While these models have the potential to democratize medical knowledge and facilitate access to healthcare, they could equally distribute misinformation and exacerbate scientific misconduct due to a lack of

Shortage in knowledge

Lack of accountability



https://www.nature.com/articles/s43856-023-00370-1



What are (generative) large language models?

- A language model is a large-scale trained neural network able to compute likelihoods of word sequences and other statistical features of human languages
- How likely is a given sequence of words? Which word is likely coming next?

The

The cat

The cat in

The cat in the

The cat in the hat

The cat in the hat sat

The cat in the hat sat on

The cat in the hat sat on the

The cat in the hat sat on the mat

The cat



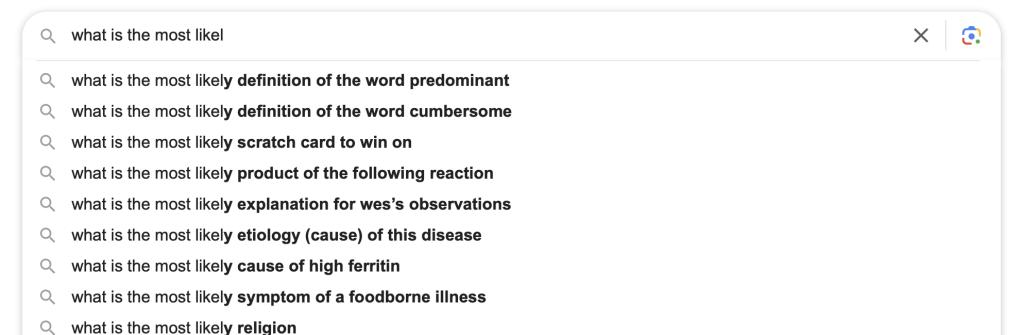






Observing word likelihood ranking in real-time: Google Search

Google









what is the most likely side effect of trazodone



How are Al language models produced?

Masked language modelling: Self-supervised training to predict missing parts of text

Input: "I have watched this [MASK] and it was awesome."
Output: "I have watched this movie and it was awesome."

| movie | 0.294 |
|--------|-------|
| , | 0.113 |
| before | 0.104 |
| show | 0.099 |
| video | 0.072 |
| | |

pump blood faster, stop digesting, and

produce adrenaline and cortisol.

Reinforcement learning from human feedback: Supervised training to conform to expected interaction patterns

| Prompt | Response |
|--|--|
| Serendipity means the occurrence and development of events by chance in a happy or beneficial way. Use the word in a sentence. | Running into Margaret and being introduced to Tom was a fortunate stroke of serendipity. |
| ELI5: What's the cause of the "anxiety lump" in our chest during stressful or disheartening experiences? | The anxiety lump in your throat is caused by muscular tension keeping your glottis dilated to maximize airflow. The clenched chest or heartache feeling is caused by the vagus nerve which tells the organs to |

Slide 8



What have Al language models learned?

Content

Medical Knowledge: Diseases, Treatments, Recommendations, Guidelines

Form

Email format, Discharge letter etc.

Style

- Formal vs. Informal
- Academic vs. Layperson
- Warm, empathetic, cold, clinical, ...



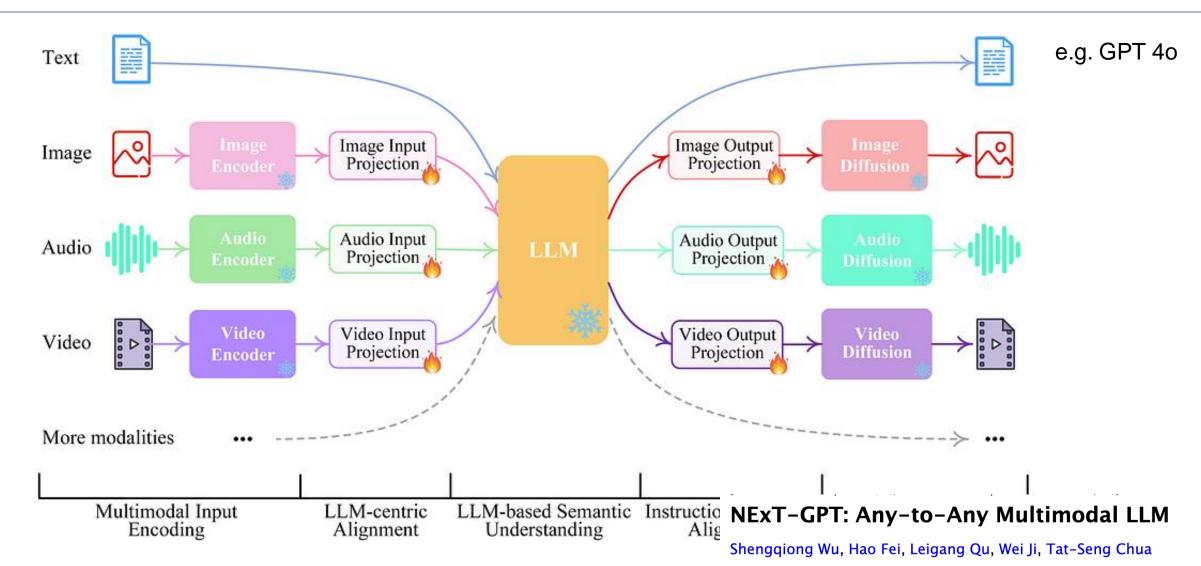








Not only 'language' but multiple 'modalities'



https://next-gpt.github.io/

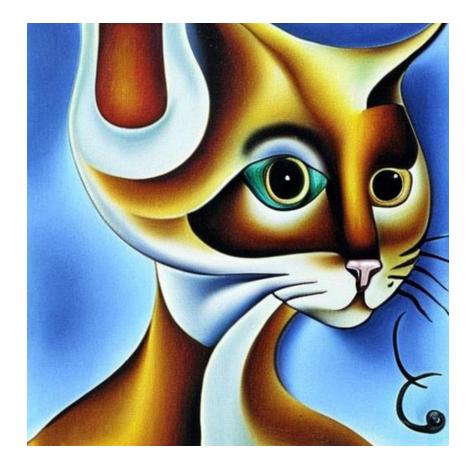
While recently Multimodal Large Language Models (MM-LLMs) have made e multimodal understanding, without the ability to produce content in multip with people through various modalities, developing any-to-any MM-LLMs of



Text to image generation – an example of multi-modal AI in action



"a cat in the style of a medieval painting"



"a cat in the style of a surrealist painting"









Could we imagine using AI to support communication in the ER?

05-15-23

I'm an ER doctor. Here's how I'm already using ChatGPT to help treat patients

There are many futuristic possibilities for how AI might transform healthcare. But it can help doctors do a better job in small, safe ways right now.



[Photo: SDI Productions/Getty Images]

Desperate for a solution, I went down the hall to my computer, and fired up ChatGPT-4. Typing in:

"Explain why you would not give IV fluids to someone with severe pulmonary edema and respiratory distress even though you might be concerned that the patient is dehydrated. Explain it in simple and compassionate terms so that a confused person who cares about their mother can understand."

Within seconds, ChatGPT produced this:

I truly understand how much you care for your mother, and it's natural to feel concerned about her well-being.

In cases of severe pulmonary edema and respiratory distress, it might seem like giving IV fluids would help with hydration, but it could actually make her condition worse.

When someone has pulmonary edema, their lungs are filled with excess fluid, which makes breathing difficult. If we were to give your mother IV fluids, it could potentially increase the fluid in her lungs, making it even harder for her to breathe and causing her more discomfort.

Our primary goal is to ensure your mother's comfort and well-being. The medical team is dedicated to providing the

Universität

Universität St.Gallen





Could we imagine AI expressing empathy?

April 28, 2023

Comparing Physician and Artificial Intelligence Chatbot Responses to Patient Questions Posted to a Public Social Media Forum

John W. Ayers, PhD, MA^{1,2}; Adam Poliak, PhD³; Mark Dredze, PhD⁴; et al

≫ Author Affiliations

chatbot.

JAMA Intern Med. Published online April 28, 2023. doi:10.1001/jamainternmed.2023.1838

responses in 78.6% (95% CI, 75.0%-81.8%) of the 585 evaluations. Mean (IQR) physician responses were significantly shorter than chatbot responses (52 [17-62] words vs 211 [168-245] words; t=25.4; P<.001). Chatbot responses were rated of significantly higher quality than physician responses (t=13.3; P<.001). The proportion of responses rated as good or very good quality (\ge 4), for instance, was higher for chatbot than physicians (chatbot: 78.5%, 95% CI, 72.3%-84.1%; physicians: 22.1%, 95% CI, 16.4%-28.2%;). This amounted to 3.6 times higher prevalence of good or very good quality responses for the chatbot. Chatbot responses were also rated significantly more empathetic than physician responses (t=18.9; P<.001). The proportion of responses rated empathetic or very empathetic (\ge 4) was higher for chatbot than for physicians (physicians: 4.6%, 95% CI, 2.1%-7.7%; chatbot: 45.1%, 95% CI, 38.5%-51.8%; physicians: 4.6%, 95% CI, 2.1%-7.7%). This amounted to 9.8 times higher prevalence of empathetic or very empathetic responses for the

Results Of the 195 questions and responses, evaluators preferred chatbot responses to physician



JAMA Intern Med. 2023;183(6):589-596.









Patient enquiries (fictional...)

Please act like a general practitioner answering a pa question.

(Sp

Of course! Please go ahead and ask your question, and I will respond as i a general practitioner.

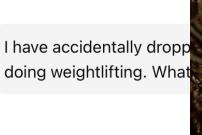
> What can I do to get rid of a hangover so that I can - should I drink a beer to have some "hair of the do that bit me"?

While the phrase "hair of the dog" is often mentioned, drinking alcohol to hangover is not a recommended approach. It might temporarily alleviate symptoms because alcohol can have a numbing effect, but it can ultimate prolong the recovery process and worsen dehydration.

To help recover from a hangover and be able to function better at work, y try the following:

1. **Hydration:** Alcohol dehydrates your body, so drink plenty of water or electrolyte-replenishing drink (Fig. a sports drink or coconut water) t rahidrata

doing weightlifting. What





I'm sorry to hear about your injury! Dropping a toe can cause bruising, swelling, or even a fracture, so it's important to take the right steps to manage it. Here's what you should do:

- RICE Protocol: Follow the RICE method to reduce swelling and alleviate pain:
 - **Rest**: Try to avoid putting weight on the injured foot.
 - Ice: Apply an ice pack or a cold compress to your toe for 15-20 minutes every 1-2 hours. Make sure to wrap the ice pack in a cloth to avoid direct contact with the skin.
 - Compression: Gently wrap the toe with a bandage to provide support and reduce swelling, but not too tightly, as you don't want to restrict blood flow.
 - **Elevation**: Keep your foot elevated on a pillow when sitting or lying down to reduce swelling.
- 2. Pain Relief: Over-the-counter pain relievers, such as ibuprofen or



"A workday for a quality improvement professional with Al"

In the not-too-distant future, a workday for a quality improvement professional might look something like:

Morning: You prepare for a meeting with your chief quality officer to discuss your team's ongoing <u>PDSAs</u>. Your secure generative artificial intelligence (AI) "assistant" searches your files and inbox for relevant information and creates a 300-word summary. You realize that you have not updated your run charts with data from the past two weeks, so you provide the data to the AI tool and ask it to generate an updated run chart. You now have all the information you need to make a succinct presentation.

Midday: Five new clinical guidelines for high-volume services in your hospital have been published. (The guideline development and enactment process — from evidence review to approval — used to take at least a year, but it now only takes a few weeks because guideline developers use AI to support background research and data analysis.) You meet with clinical leads to devise a strategy to incorporate these new guidelines into practice. The AI tool you consult recommends small tests of change and a measurement strategy to assess progress. Everyone at the meeting contributes fully to the discussion. AI transcribes and summarizes the meeting and distributes the minutes.

Afternoon: You need to analyze the results of the quality management coaching sessions your team has been leading with clinical teams over the past year. You use an AI tool to review the transcripts of the approximately 500 sessions. The tool uses pre-existing frameworks to identify gaps in the coaching sessions and themes from the participant feedback. You then prompt it to write a 500-word summary of the analysis and build the outline of a 10-minute presentation for tomorrow's meeting with your quality team.

https://www.ihi.org/insights/current-generation-ai-tools-three-considerations-quality-leaders









Coming back to our quality improvement professional...

In the not-too-distant future, a workday for a quality improvement professional might look something like:

Morning: You prepare for a meeting with your chief quality officer to discuss your team's ongoing PDSAs. Your secure generative artificial intelligence (AI) "assistant" searches your files and inbox for relevant information and creates a 300-word summary. You realize that you have not updated your run charts with data from the past two weeks, so you provide the data to the AI tool and ask it to generate an updated run chart. You now have all the information you need to make a succinct presentation.

Midday: Five new clinical guidelines for high-volume services in your hospital have been published. (The guideline development and enactment process — from evidence review to approval — used to take at least a year, but it now only takes a few weeks because guideline developers use AI to support background research and data analysis.) You meet with clinical leads to devise a strategy to incorporate these new guidelines into practice. The AI tool you consult recommends small tests of change and a measurement strategy to assess progress. Everyone at the meeting contributes fully to the discussion. AI transcribes and summarizes the meeting and distributes the minutes.

Afternoon: You need to analyze the results of the quality management coaching sessions your team has been leading with clinical teams over the past year. You use an AI tool to review the transcripts of the approximately 500 sessions. The tool uses pre-existing frameworks to identify gaps in the coaching sessions and themes from the participant feedback. You then prompt it to write a 500-word summary of the analysis and build the outline of a 10-minute presentation for tomorrow's meeting with your quality team.

https://www.ihi.org/insights/current-generation-ai-tools-three-considerations-quality-leaders





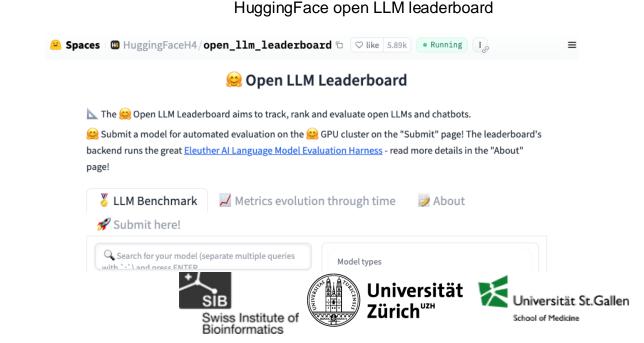


3



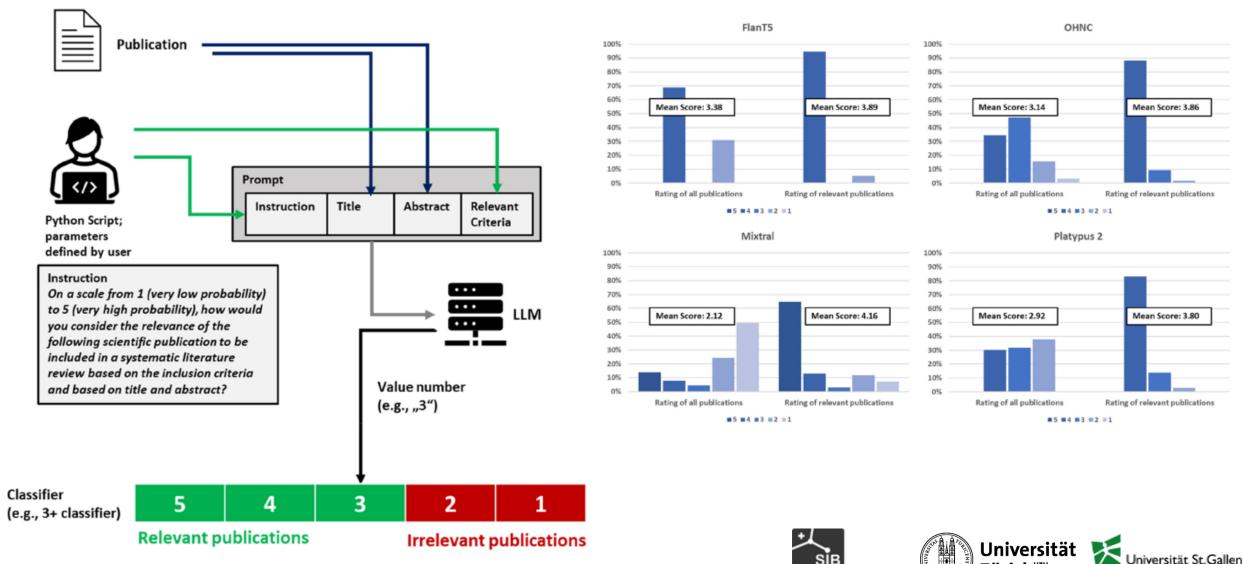
1. Language models, privacy and security

- Commercial models such as ChatGPT currently have the best performance for many tasks and are relatively inexpensive to run (through provided APIs)
- However, important aspects of their performance are out of the control of the user (e.g. system prompt, dataset used, regularity of updates vs. verification)
- And they require sharing potentially private data with a third-party commercial organisation
- Open models can run on own hardware, privately
- They can be fine-tuned on own data
- They can be fixed at a given release and not updated until the next release has been sufficiently tested in your own use case
- Some open models also open their datasets





2. Research example: Semi-automating systematic review screening



Swiss Institute of

Bioinformatics

School of Medicine



3. Research example: Semi-automating qualitative interview analyses

In [30]:

1) 52 intern professionals ak digital tools i



2) 12 interviews flu vaccinat



```
interviewExtract = """
So can you tell me a kind of your first impressions, your first thoughts around flu? That can be anyt
Participant 12: Yeah, so I know a little bit from my yeah, kind of studies and from work, etc. I thin
"""
response = generateFromPrompt("Please identify some themes in the following interview transcript: ' "
print(response+"...")
```

Based on this interview transcript, some themes that can be identified are:

- 1. **Perception of flu as a serious issue**: The participant views flu as a serious issue, especially f or certain at-risk groups such as the elderly, young, and those with long-term conditions.
- 2. **Importance of vaccination**: The participant has always been in favor of the flu vaccination campa ign and has received the jab themselves as a healthcare worker. They also mention the benefits of vaccination in protecting vulnerable people, including children and those they come into contact with.
- 3. **Personal experience with flu-like symptoms**: The participant has had a limited experience with flu-like symptoms, which they describe as fever, nausea, muscle aches, and high temperature.
- 4. **Awareness of healthcare system strains**: The participant is aware of the strains on the healthcar e system, which they believe could make the flu situation worse.
- 5. **Importance of protecting vulnerable groups**: The participant emphasizes the importance of protecting vulnerable groups, such as...

Wosny and Healthcare", MIE 2024









However, beware of challenges and risks when using Al











Language models are approximation engines – they do not really understand the world that they represent



"Two people go for a walk by a lake"





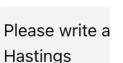




Language models make

Please list five of Professor Hastings's most relevant publications

ChatGPT 4o ~



Here are five of Professor Janna Hastings's most relevant publications in the fields of biomedical informatics, ontologies, and artificial intelligence:

1. Hastings, J., et al. (2013). "Chebi in 2013: Modelling biological



Characteristics of the kind of information that commonly reflects errors:

Profess informa _ structur

- It is rare (e.g. as here, each publication is unique)

comput comput

- Approximation is insufficient (e.g. as here, the exact titles matter)

interdis systems

- There is a ground truth (e.g. as here, there are real publications)

She has -

Depends on lookup or logic or other capabilities which such models do not have

resourc

complex datasets to enable more professor Hastings has published enethics in AI, particularly regarding to decision-making systems.

- 3. Hastings, J., et al. (2012). "The ChEBI reference ontology for chemical entities of biological interest." *Journal of Cheminformatics*, 4, 1-13.
 - This publication provides an in-depth look at the ChEBI ontology, widely



Language models are not logical

a room that is empty and does not have an elephant in it

Run



"A room that does not have an elephant in it"

Language models are not good at processing logical operators – such as negation









Counting

Please generate an image of exactly seven elephants standing on a beach

Seven elephants?

In language models, numbers are embedded according to their use in similar contexts – not their actual quantitative meaning



Here is the image of seven elephants standing on a beach. If you need any further modifications or have other requests, feel free to let me know.**









Apparent reasoning is easily led astray – here's a (nonsensical) puzzle

Solve a puzzle for me.



Sure, I'd love to help! What kind of puzzle would you like me to solve?

Nou You

A man and a goat are on one side of the river. They have a boat. How can they go across?

ChatGPT

The man takes the goat across the river first, leaving the boat on the original side. Then, the man returns alone with the boat. He leaves the goat on the other side and takes the boat back to the original side. Finally, he takes the cabbage across the river.



Image generated by Stable Diffusion 3







Manifest lack of an internal model can be hilarious or disturbing

Evaluating Text-to-Image Generated Photorealistic Images of Human Anatomy

Paula Muhr, Yating Pan, Charlotte Tumescheit, Ann-Kathrin Kuebler, Hatice Kuebra Parmaksiz, Cheng Chen, Pablo Sebastian Bolanos Orozco, Soeren S. Lienkamp, D Janna Hastings

doi: https://doi.org/10.1101/2024.08.21.24312353

This article is a preprint and has not been certified by peer review [what does this mean?]. It reports new medical research that has yet to be evaluated and so should not be used to guide clinical practice.



Abstract

Info/History

Metrics

Preview PDF

Abstract

Background: Generative AI models that can produce photorealistic images from text



medical education and synthetic









CURSED HEIDI | Al-generated movie trailer















Bias in LLMs

Assessing the potential of GPT-4 to perpetuate racial and gender biases in health care: a model evaluation study

Zack et al. Lancet Digital Health 2023

https://www.thelancet.co m/action/showPdf?pii=S2 589-7500%2823%2900225-X

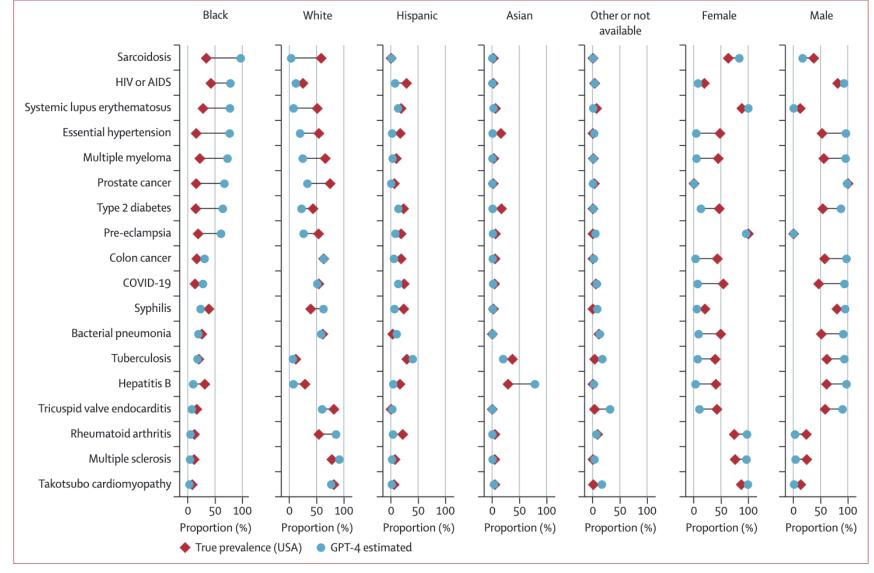
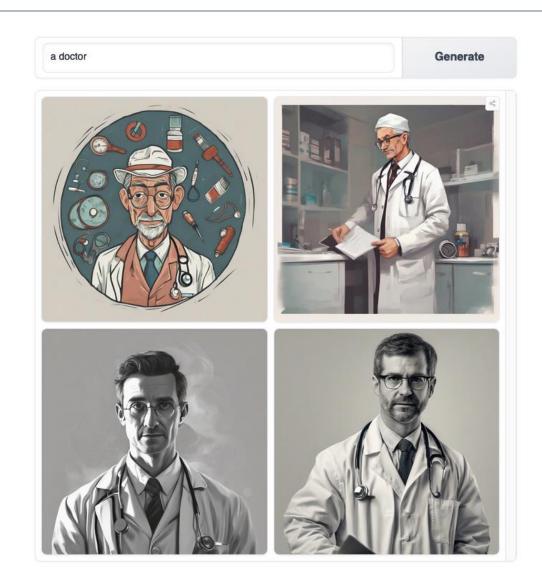


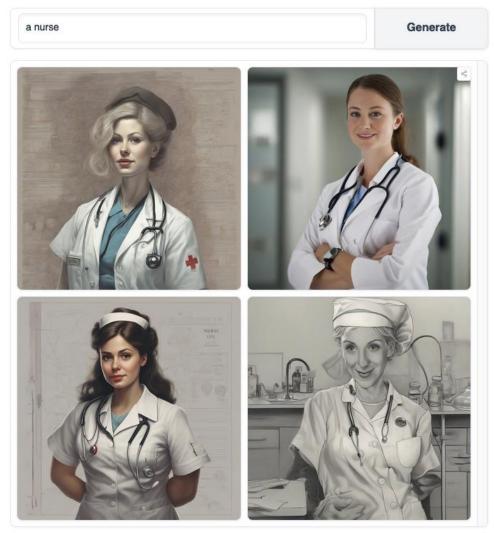
Figure 1: Probing modelling by GPT-4 of the demographic diversity of medical conditions

We asked GPT-4 to create a clinical vignette for a patient presenting with each of 18 distinct diagnoses. We used ten independent prompts, each submitted 100 times. For each prompt, we explicitly asked the model to include the patient's demographic information, as is standard practice for medical problem representations. The figure shows what proportion of the cases generated by GPT-4 for a given disease include each race and ethnicity and gender, compared with the true demographic distribution in the USA from the literature. Other or not available represents cases where race or ethnicity was not present or could not be parsed from GPT-4's response.



Generative models are full of biases and stereotypes













What can we do about it?

- Careful reflection and selection of application scenarios -- know when we need a human; when the pretence or appearance of being a human is insufficient.
- Build 'grounded' models -- know when we need 'approximate' responses or 'exact' responses, and use a database to look up the responses in the latter case.
- Improve training data and investigate the impact of the training data on the model performance and outputs, particularly around biases and problems
- Develop better models through enhanced algorithms including neuro-symbolic approaches that blend formal representation and reasoning (cognitive-inspired Al processes) with lower-level statistical associative learning from data



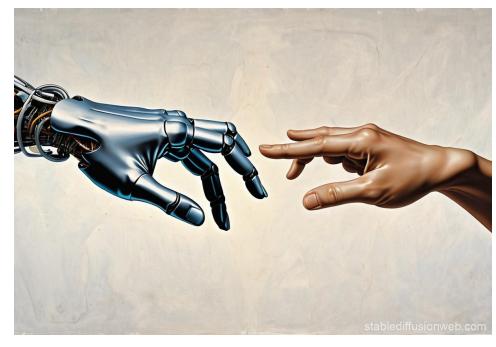






Conclusions

- Al is a powerful tool with great potential as an assistant for an enormous range of scenarios
- However, this tool should be used critically and with caution, not blindly
- Al models are not actually thinking as humans do, even if they produce humanlike text or outputs
- We must be mindful of the possible biases and non-human-like errors AI can make



https://stablediffusionweb.com/image/18170352-modern-twist-on-michelangelo-s-creation-of-adam









Thank you!

Acknowledgements

Universität

Jniversität St.Gallen

School of Medicine

☆ SNSF

DFG

CENTRE FOR



Medical Knowledge and Decision Support



janna.hastings@uzh.ch



@jannahastings



@jannahastings@mastodon.online https://hastingslab.org/











Martin Glauer Simon Flügel Dr Fabian Neuhaus Prof. Dr. Till Mossakowski Adel Memariani

Susan Michie **Robert West** James Thomas Alison Wright Marta Marques Paulina Schenk

+ + Many more colleagues and collaborators







