Detecting and mitigating biases in multi-modal generative AI models

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

ISROI meeting, Freiburg, 24 May 2024

Background image generated by Stable Cascade

janna.hastings@uzh.ch
@jannahastings
@jannahastings@mastodon.online
https://hastingslab.org/

Overview of today's talk

Generative AI: A new paradigm for digital medicine

Multi-modal pre-trained models: Opportunities and challenges in medicine

Evaluation, bias and mitigation strategies

Generative AI – a new paradigm in digital medicine

Medical Knowledge and Decision Support

“black box”, data bottleneck

1990s

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

“good old-fashioned AI”

2000s

growing applications of data-driven supervised machine learning in imaging, pathology, discovery, diagnosis and prognosis

“traditional machine learning”

2011+

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

2017+

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

“black box” *and* unknown data

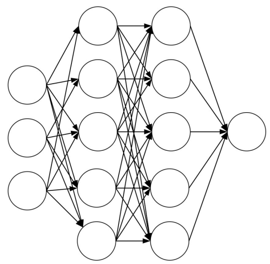
generative AI

Universität Zürich[UZH]

Universität St.Gallen School of Medicine

Generative pre-trained models are based on Transformer architecture

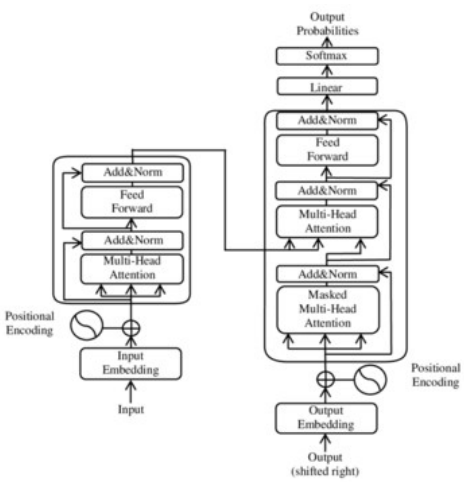
Pre-training



memorisation
meaning similarity
structural patterns

The cat

https://medium.com/@evertongomede/next-word-prediction-enhancing-language-understanding-and-communication-1322f3b57632

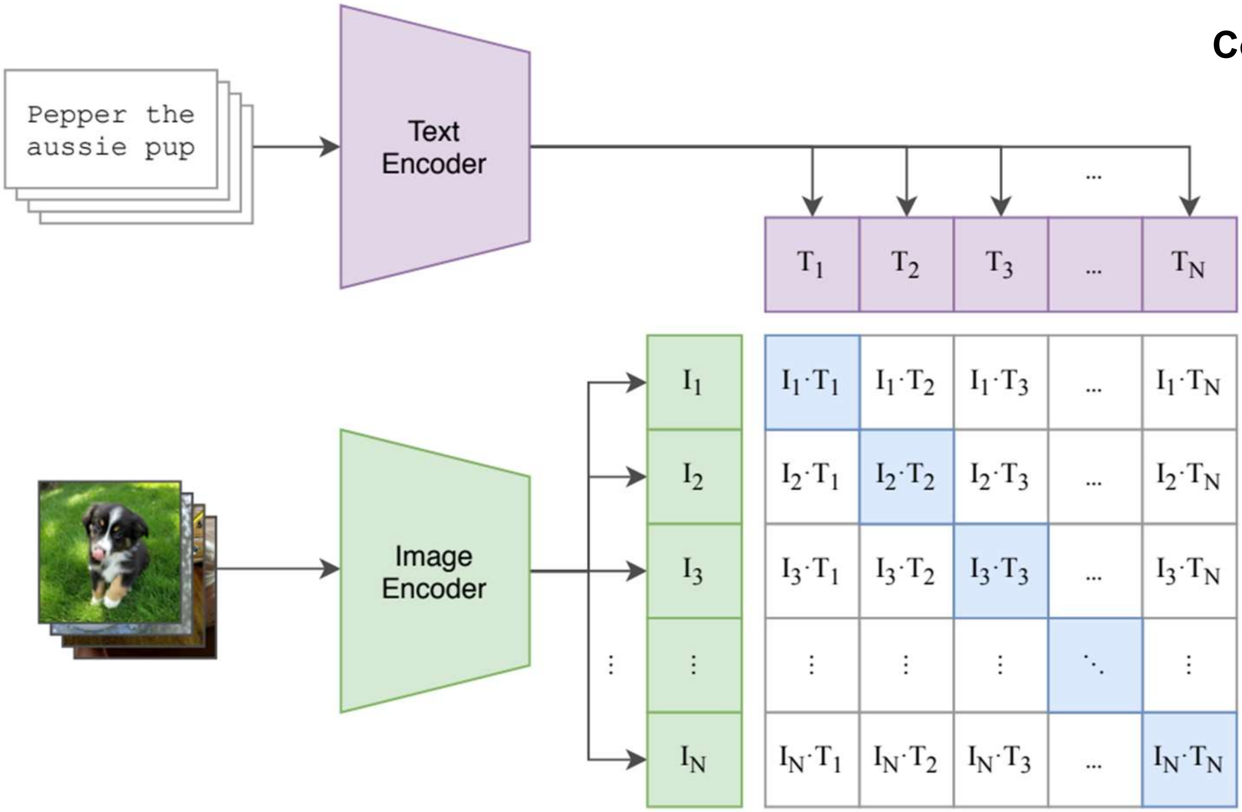


Attention is all you need

A Vaswani, N Shazeer, N Parmar… - Advances in neural …, 2017 - proceedings.neurips.cc

The dominant sequence transduction models are based on complex recurrent orconvolutional neural networks in an encoder and decoder configuration. The best performing such models also connect the encoder and decoder through an attentionm echanisms. We propose a novel, simple network architecture based solely onan attention mechanism, dispensing with recurrence and convolutions entirely. Experiments on two machine translation tasks show these models to be superiorin quality while being more …

Multi-modal language/image pre-training: CLIP



Contrastive Language-Image Pre-Training

Supports **visual understanding** => generating text descriptions of visual inputs

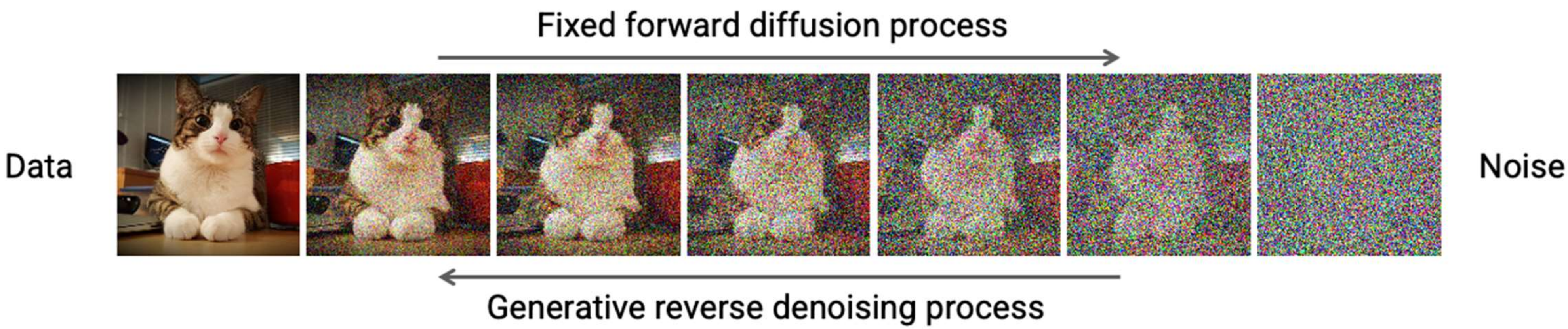
Learning Transferable Visual Models From Natural Language Supervision

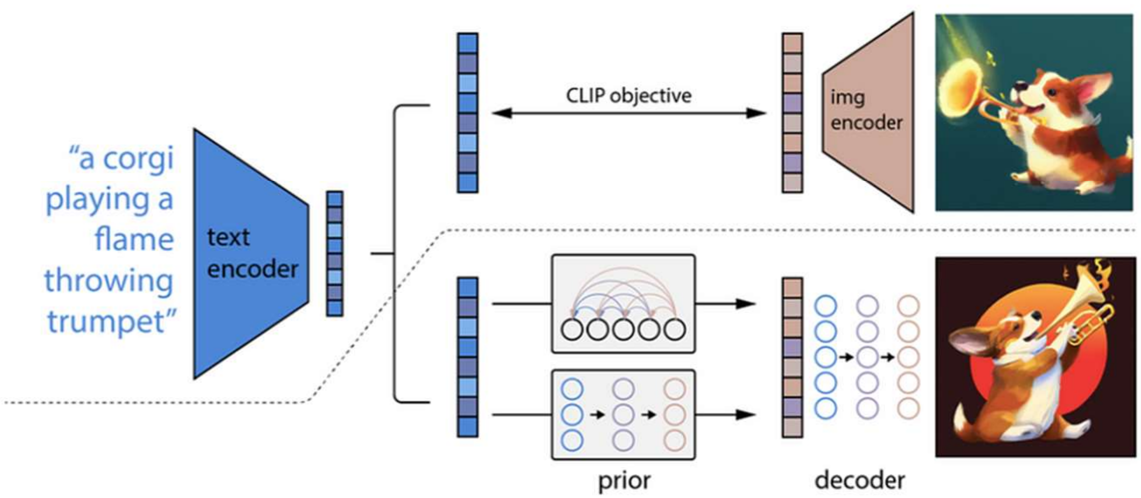
Alec Radford [*1] Jong Wook Kim [*1] Chris Hallacy [1] Aditya Ramesh [1] Gabriel Goh [1] Sandhini Agarwal [1] Girish Sastry [1] Amanda Askell [1] Pamela Mishkin [1] Jack Clark [1] Gretchen Krueger [1] Ilya Sutskever [1]

https://openai.com/index/clip/

Universität Zürich[UZH] Universität St.Gallen School of Medicine

Conditional denoising diffusion – generating coherent outputs



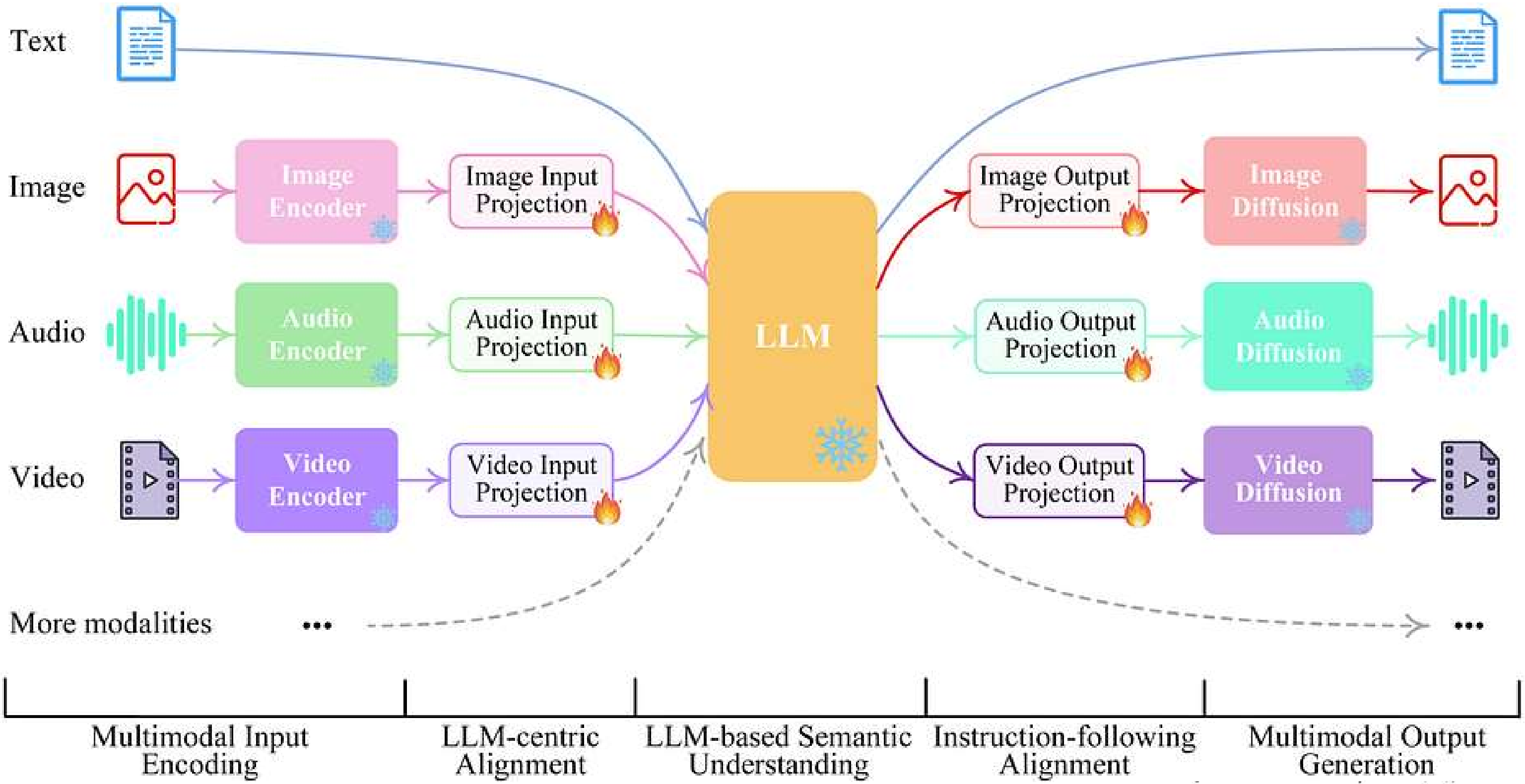


CLIP + De-noising diffusion

Figure from DALL-E2

Medical Knowledge and Decision Support

Additional modalities may be added – centred around language



here, Next-GPT

but also

ChatGPT-4o

(o=omni)

NExT-GPT: Any-to-Any Multimodal LLM

Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, Tat-Seng Chua

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 c

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

Examples of visual understanding in ChatGPT-4o

Sure, I'll create a puzzle for you. Let

Riddle:

I speak without a mouth and hear w
What am I?

It looks like you've uploaded an X-ray image. While I can provide some general information about interpreting X-rays, it is important to note that a detailed analysis and diagnosis should be performed by a qualified medical professional.

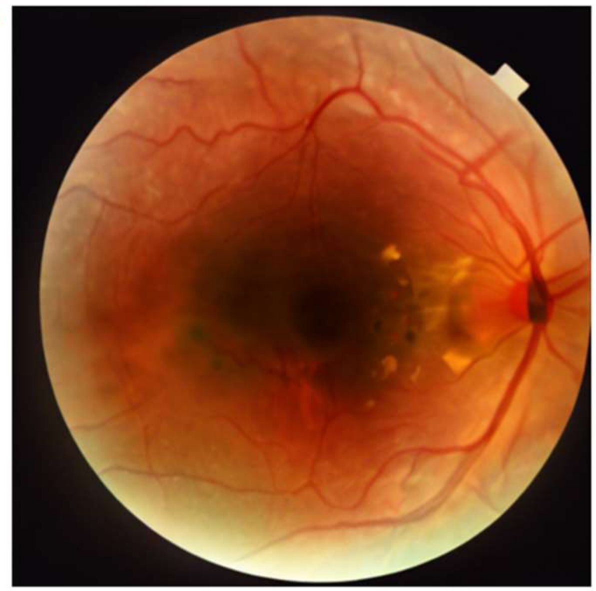
In a standard chest X-ray:

- The heart and great vessels appear in the center of the image.
- The lungs should appear as dark areas since they are filled with air.

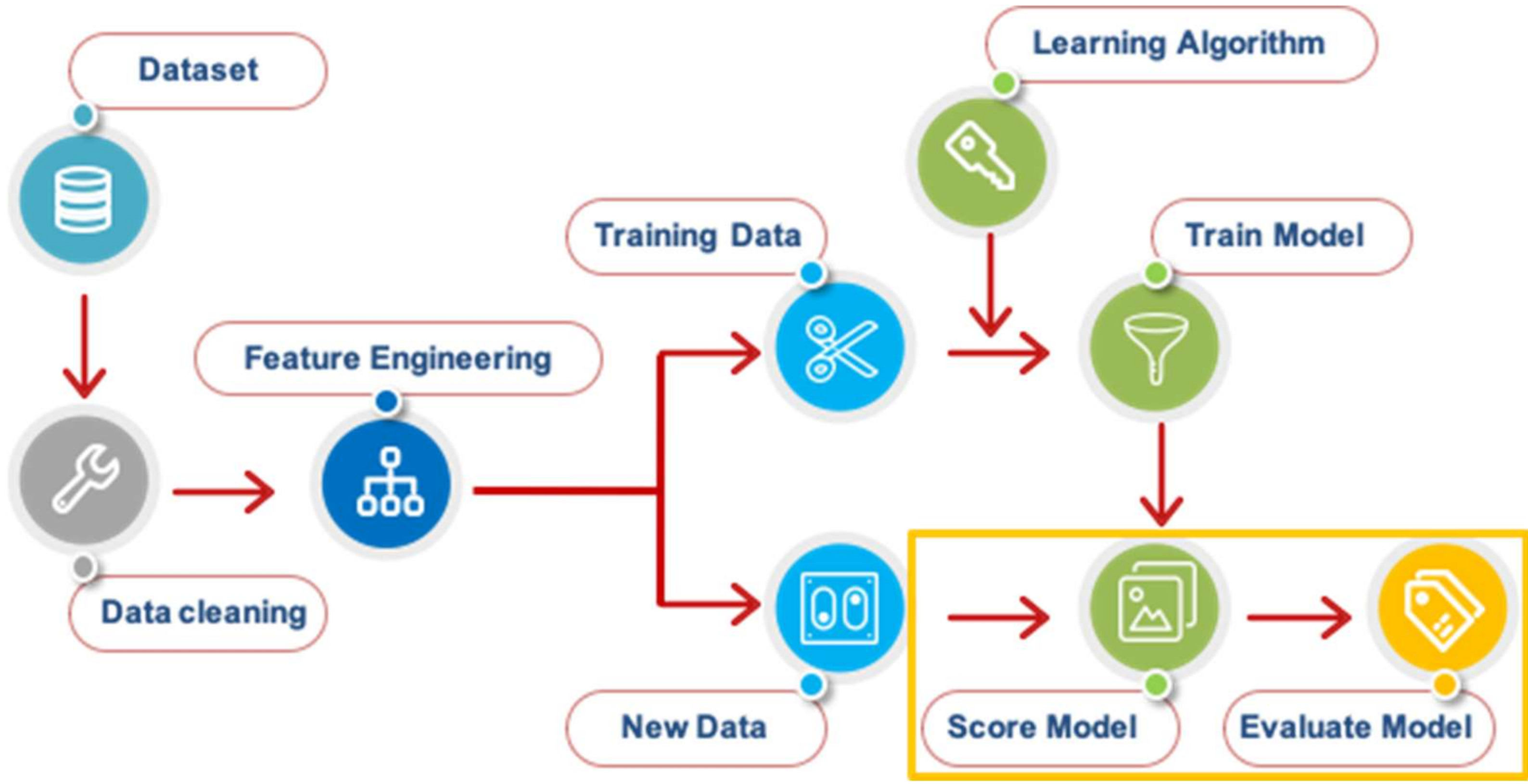
Medical applications for multi-modal generative models

- Extracting structured data from unstructured clinical notes for real-world clinical research
- Automating note generation, reducing clinical documentation time burden
- Generation of synthetic datasets, unlocking additional research possibilities while preserving patient privacy

```
In [9]: image = pipe("A retinal fundus image", num_inference_steps=50).images[0]
        image
          0%|          | 0/50 [00:00<?, ?it/s]
Out[9]:
```



Note: Evaluation of generative AI performance is challenging



"Traditional" machine learning is evaluated on **unseen data** with known **ground truth**

In generative pre-trained models – you often don't know what the model has seen, and you often don't know the ground truth

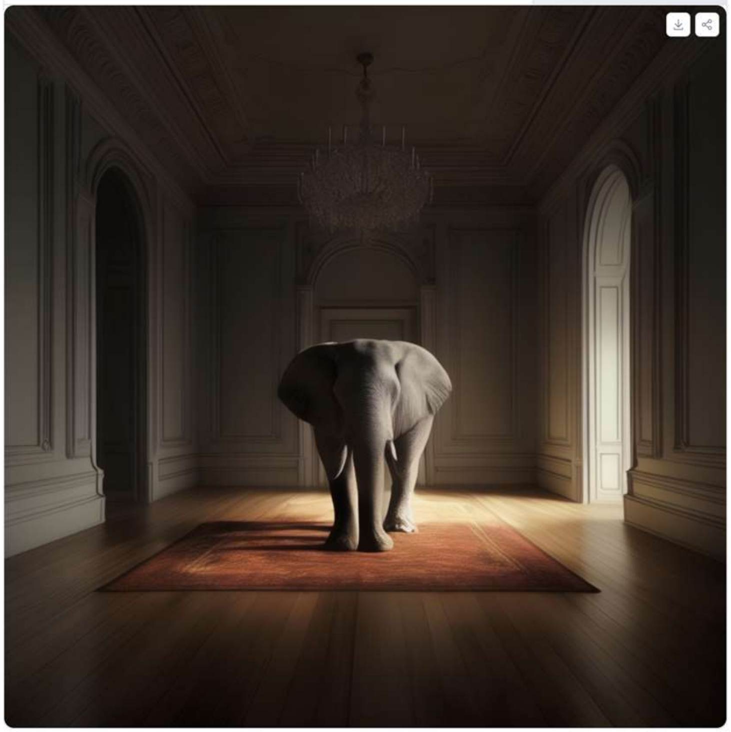
https://www.chrisgiler.com/blog/model-evaluation-metrics-for-classification-algorithms

Medical Knowledge and Decision Support

Models are powerful similarity and conditional generation engines. But they do not ‘understand’ in the sense we usually use for humans

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

Run



Stable Diffusion (stable cascade)

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

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.

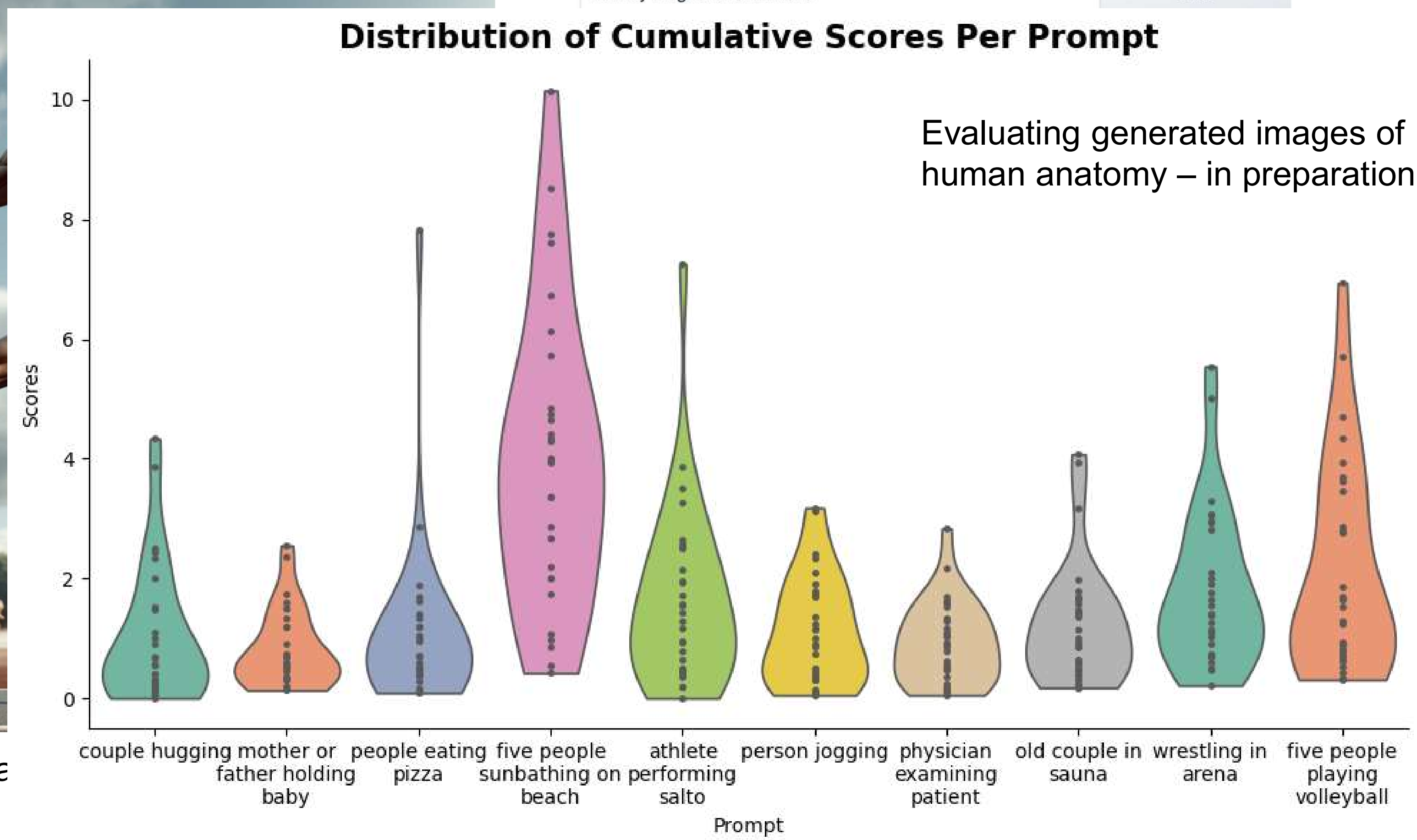
GPT-4o / DALL-E

Universität Zürich UZH

Universität St.Gallen School of Medicine

Models make non-human-like errors

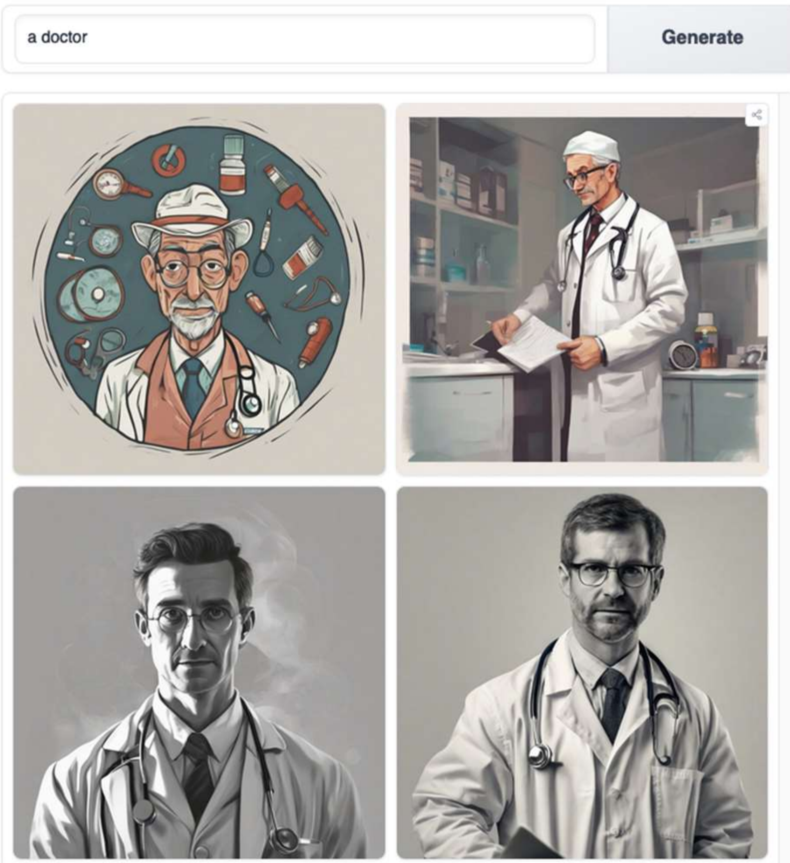
An x-ray image of a human hand
Run

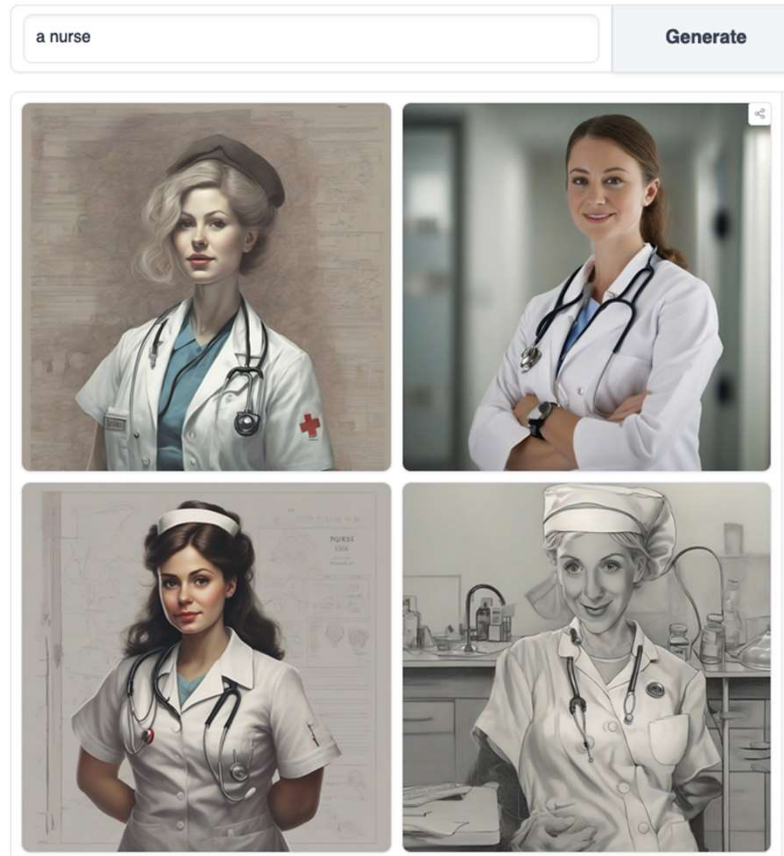


Evaluating generated images of human anatomy – in preparation

Over-idealiza

Models reflect many biases and stereotypes





Hastings, ”Preventing Harm from Non-Conscious Bias in Medical Generative AI”, *Lancet Digital Health*

Model biases may worsen inequalities for vulnerable populations

AI dermatology image-based diagnostic algorithm performs 50% worse on Black skin than advertised performance

Using artificial intelligence on dermatology conditions in Uganda: A case for diversity in training data sets for machine learning

Louis Henry Kamulegeya, Mark Okello, John Mark Bwanika, Davis Musinguzi, William Lubega, Davis Rusoke, Faith Nassiwa, Alexander Börve

doi: https://doi.org/10.1101/826057

of image body parts uploaded. Overall diagnostic accuracy of the AI app was low at 17% (21 out of 123 predictable images) with varying predictability levels correctness i.e. 1-8.9%, 2-2.4%, 3-2.4%, 4-1.6%, 5-1.6% with performance along individual diagnosis highest with dermatitis (80%).

Medical Knowledge and Decision Support

nature > nature medicine > articles > article

Article | Published: 19 April 2024

Demographic bias in mis[...] pathology models

Anurag Vaidya, Richard J. Chen, Drew F. K. Wil[...] Yang, Thomas Hartvigsen, Emma C. Dyer, Ming[...] Chen & Faisal Mahmood

Nature Medicine **30**, 1174–1190 (2024) | Cite

3876 Accesses | **1** Citations | **116** Altmetric

Abstract

Despite increasing numbers of regulatory [...] pathology systems often overlook the imp[...] potentially leading to biases. This concern [...] pathology has leveraged large public datas[...] groups. Using publicly available data from [...] tumor atlas, as well as internal patient data [...] models display marked performance disp[...] used to subtype breast and lung carcinom[...] example, when using common modeling a[...] under the receiver operating characteristi[...] for breast cancer subtyping, 10.9% for lung [...] prediction in gliomas. We found that riche[...]

nature

Explore content | About the journal | Publish with us

nature > articles > article

Article | Open access | Published: 22 May 2024

A whole-slide foundation model for digital pathology from real-world data

Hanwen Xu, Naoto Usuyama, Jaspreet Bagga, Sheng Zhang, Rajesh Rao, Tristan Naumann, Cliff Wong, Zelalem Gero, Javier González, Yu Gu, Yanbo Xu, Mu Wei, Wenhui Wang, Shuming Ma, Furu Wei, Jianwei Yang, Chunyuan Li, Jianfeng Gao, Jaylen Rosemon, Tucker Bower, Soohee Lee, Roshanthi Weerasinghe, Bill J. Wright, Ari Robicsek, ... Hoifung Poon

+ Show authors

Nature (2024) | Cite this article

109 Altmetric | Metrics

Abstract

Digital pathology poses unique computational challenges, as a standard gigapixel slide may comprise tens of thousands of image tiles[1,2,3]. Prior models have often resorted to subsampling a small portion of tiles for each slide, thus missing the important slide-level context[4]. Here we present Prov-GigaPath, a whole-slide pathology foundation model

Predictive models may give accurate results, but for the wrong reasons

> Lancet Digit Health. 2022 Jun;4(6):e406-e414. doi: 10.1016/S2589-7500(22)00063-2.
Epub 2022 May 11.

AI recognition of patient race in medical imaging: a modelling study

Judy Wawira Gichoya [1], Imon Banerjee [2], Ananth Reddy Bhimireddy [3], John L Burns [4], Leo Anthony Celi [5], Li-Ching Chen [6], Ramon Correa [2], Natalie Dullerud [7], Marzyeh Ghassemi [8], Shih-Cheng Huang [9], Po-Chih Kuo [6], Matthew P Lungren [9], Lyle J Palmer [10], Brandon J Price [11], Saptarshi Purkayastha [4], Ayis T Pyrros [12], Lauren Oakden-Rayner [13], Chima Okechukwu [14], Laleh Seyyed-Kalantari [15], Hari Trivedi [3], Ryan Wang [6], Zachary Zaiman [16], Haoran Zhang [7]

Affiliations + expand

PMID: 35568690 PMCID: PMC9650160 DOI: 10.1016/S2589-7500(22)00063-2

Free PMC article

Abstract

Background: Previous studies in medical imaging have shown disparate abilities of artificial intelligence (AI) to detect a person's race, yet there is no known correlation for race on medical imaging that would be obvious to human experts when interpreting the images. We aimed to

Race is a confounder

AI models can predict race from medical images with high performance:

- X-ray – AUC 0.91-0.99
- chest CT – AUC 0.87-0.96
- mammography – AUC 0.81

Mitigating biases and stereotypes in models is challenging

- Model generation of outputs is stochastic
- Some biases affect the distribution of possible outputs
- Will not necessarily be evident in a single generated output

- Strategies to mitigate biases include:
  - Explicitly prompt for desired distribution
  - Use “retrieval-augmentation” strategies to supplement generation
  - Post-filter generated output checking for problems
  - Fine-tuning the model with more representative data
  - Longer term -- improve training data

Hastings, ”Preventing Harm from Non-Conscious Bias in Medical Generative AI”, *Lancet Digital Health*

Privacy concerns – and open source models

Open Source Models, Own Installation, Own Hardware

- 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 source models can be 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

LLaVA: Large Language and Vision Assistant

Visual Instruction Tuning

NeurIPS 2023 (Oral)

Haotian Liu*, Chunyuan Li*, Qingyang Wu, Yong Jae Lee

University of Wisconsin-Madison ▶ Microsoft Research ▶ Columbia University

NExT-GPT: Any-to-Any Multimodal LLM

Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, Tat-Seng Chua

While recently Multimodal Large Language Models (MM-LLMs) have made e:
multimodal understanding, without the ability to produce content in multip

St.Gallen

Cross-Care: Cataloguing what is in the training data, to study bias

Cross-Care: Assessing the Healthcare Implications of Pre-training Data on Language Model Bias

Shan Chen, Jack Gallifa
Janna Hastings, Hugo A

Large language models
and inaccuracies origina
biases and real world kr
systematically evaluate l
quantify discrepancies b
substantial misalignmer
indicating a pronounced
various alignment meth
further exploration and

Thank you!

Prof. Dr. Janna Hastings

Medical Knowledge and Decision Support

janna.hastings@uzh.ch
@jannahastings
@jannahastings@mastodon.online
https://hastingslab.org/

Acknowledgements

Universität Zürich[UZH]
Universität St.Gallen School of Medicine
SNSF

Marie Wosny
Livia Strasser
Charlotte Tumescheit
Joshua Sammet
Björn Gehrke
Paula Muhr

Otto von Guericke Universität Magdeburg
DFG

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

UCL Centre for Behaviour Change
Wellcome Trust

Susan Michie
Robert West
James Thomas
Alison Wright
Marta Marques

+ Many more colleagues and collaborators around the world

Universität Zürich[UZH] Universität St.Gallen School of Medicine