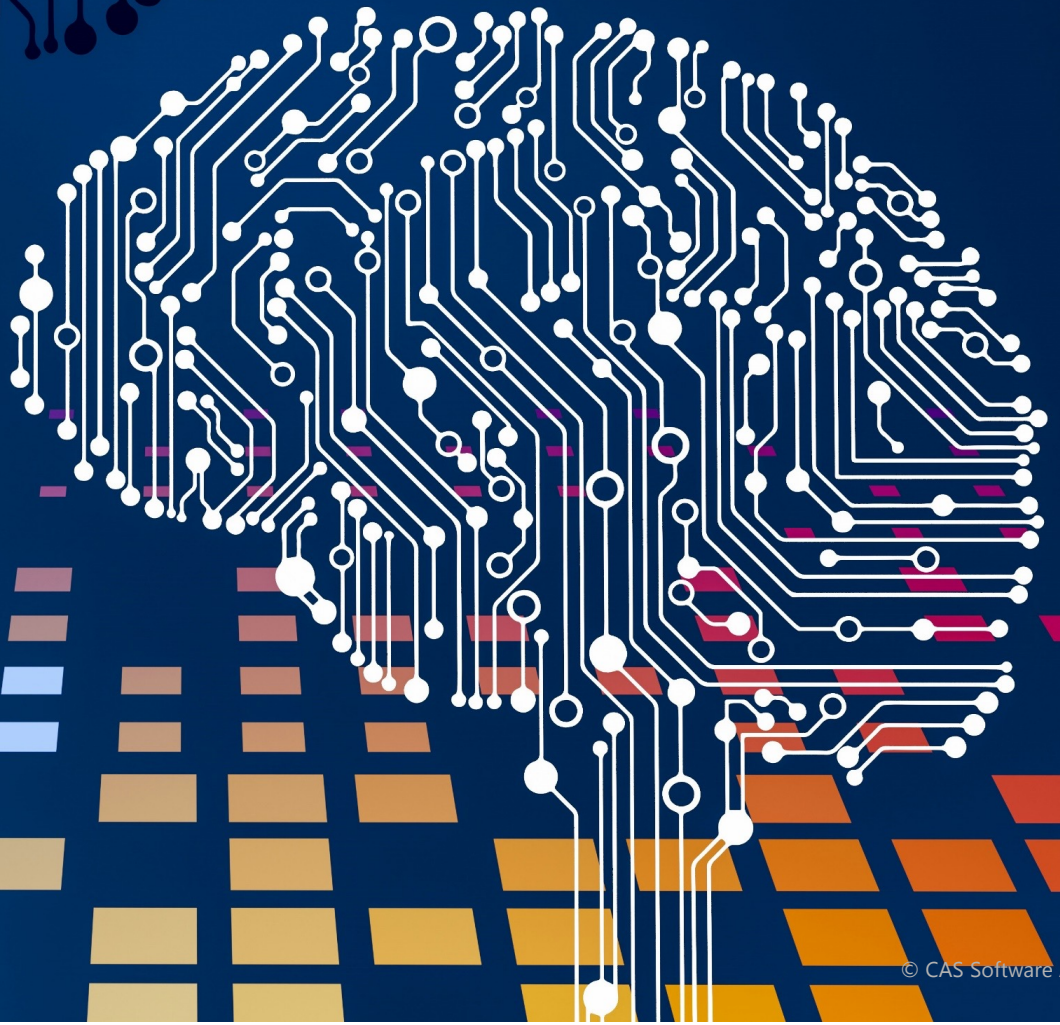
AI and LLMs

A short introduction

Dr. Markus Bauer, CAS Software AG

January 2024



Agenda

AI

- AI – a short introduction
- Machine Learning
- Neural networks
- LLMs
- Summary

What is AI?

ChatGPT reaches 100 million users two months after launch

Unprecedented take-up may make AI chatbot the fastest-growing consumer internet app ever, analysts say

The World Ahead | Business in 2024

Generative AI will go mainstream in 2024

TECH

World's first major law for artificial intelligence gets final EU green light

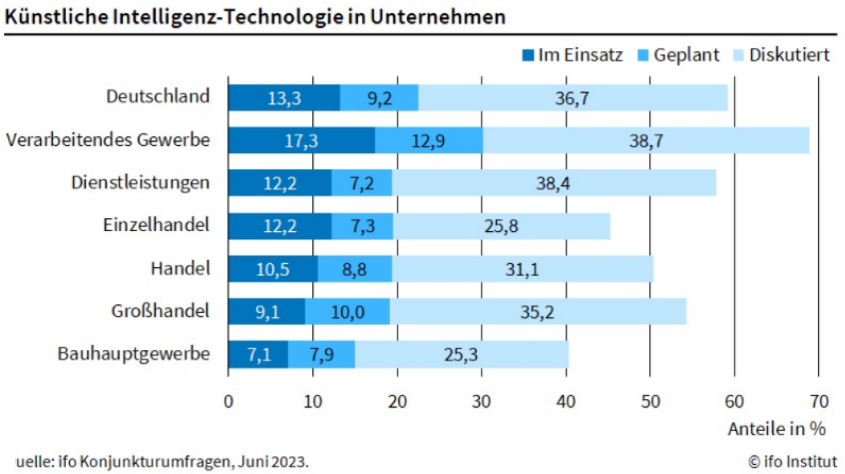
BBC

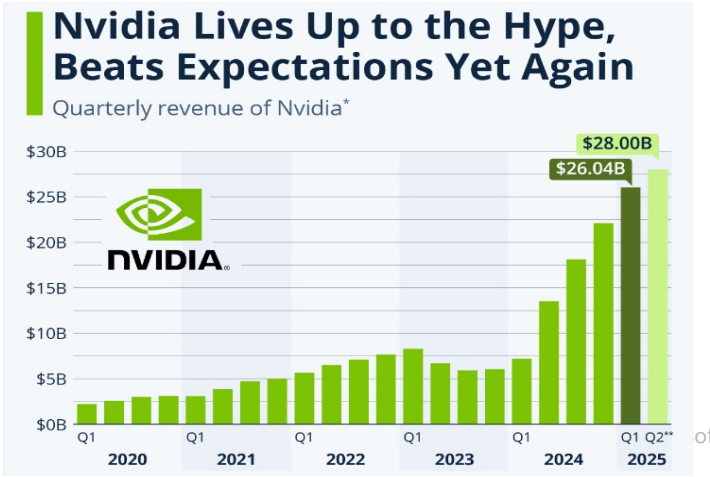
AI could replace equivalent of 300 million jobs - report

Pope Francis Raises Alarm About AI

Revolution durch künstliche Intelligenz

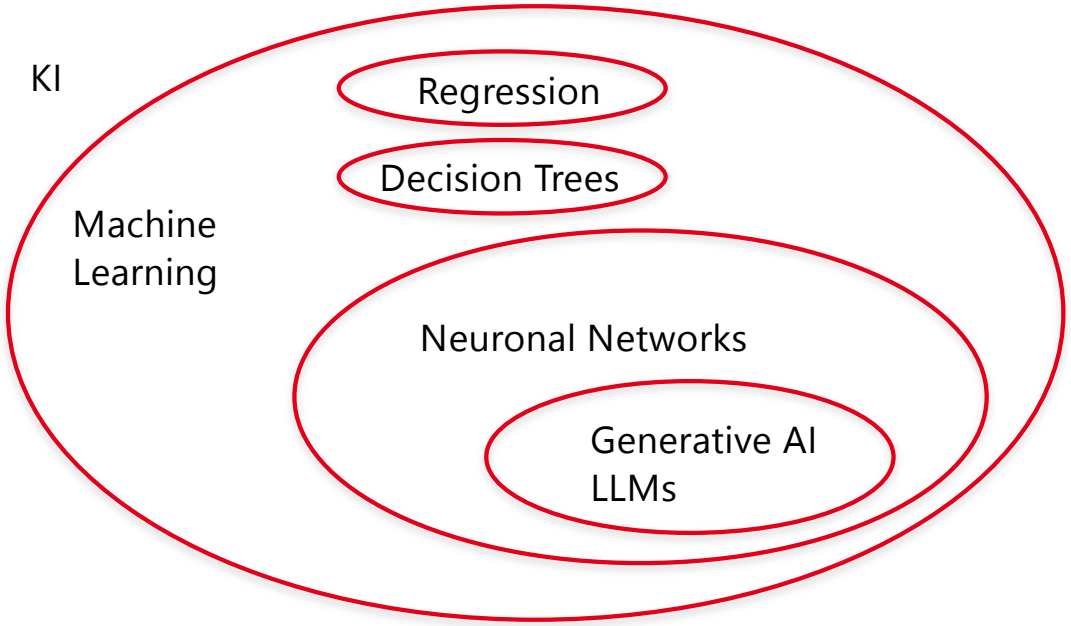
Wofür brauchen wir Schulen überhaupt noch?



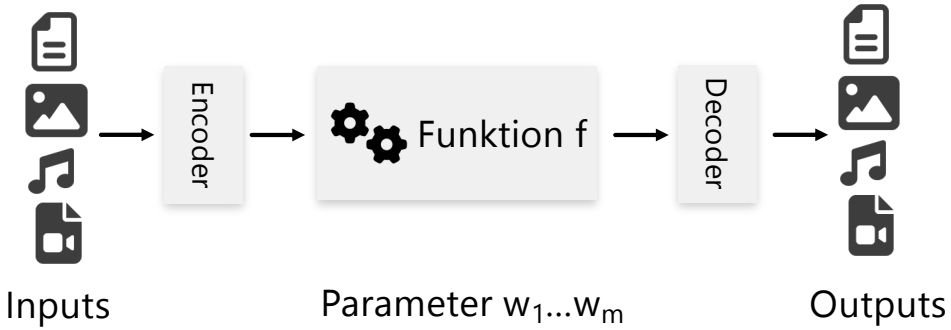


What is AI?

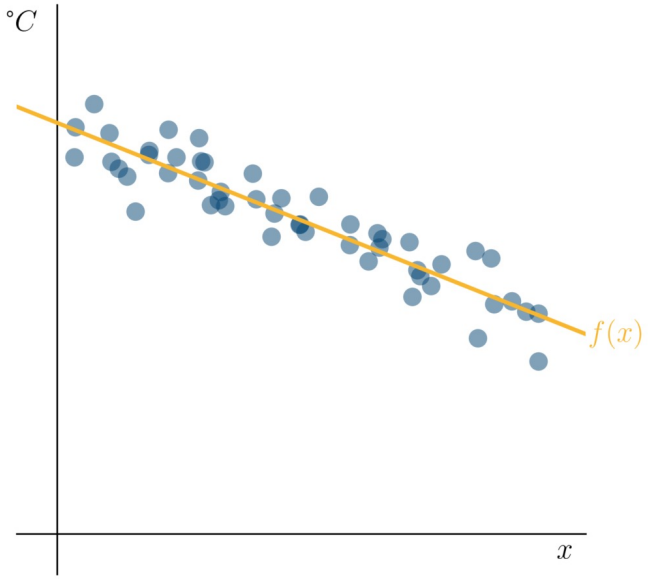
- Definition: *AI is the science of "creating intelligent machines that complement human reasoning to augment and enrich our experience and competencies."* *(Microsoft 2020)*



Intuition: Machine Learning

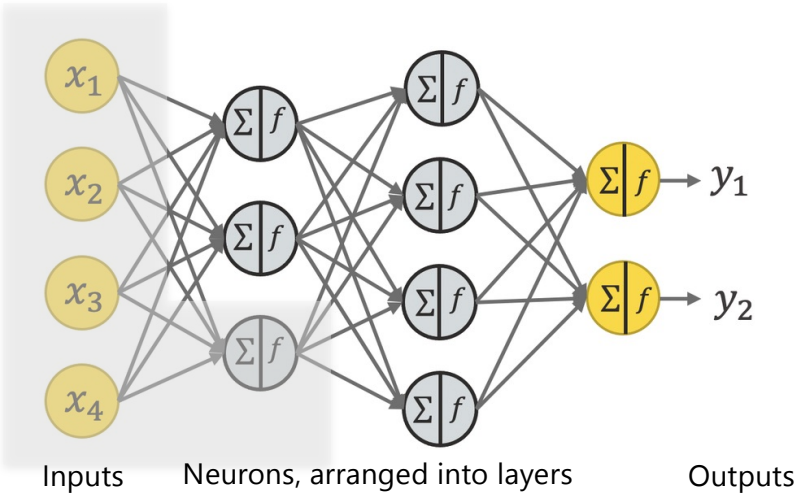


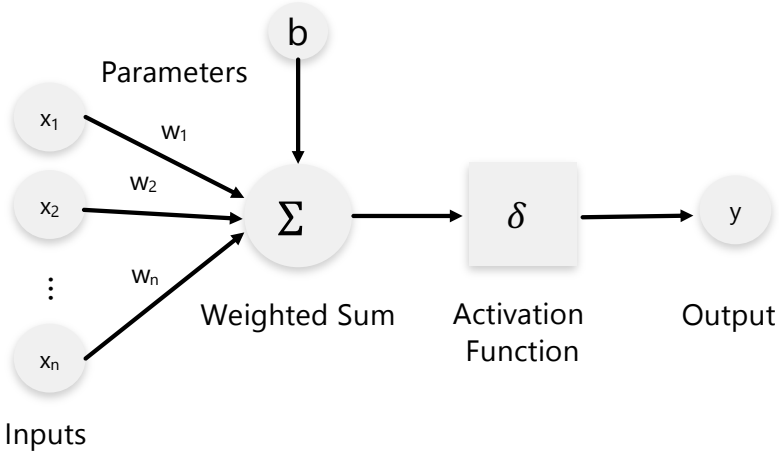
- Model: Parameterized Function that transforms Inputs into Outputs
- Training: Determine the parameters $w_1...w_m$ using sample data (= large amounts of known “correct” inputs, outputs



Example: linear regression
$f(x) = w_1 * x + w_2$

Neuronal Networks

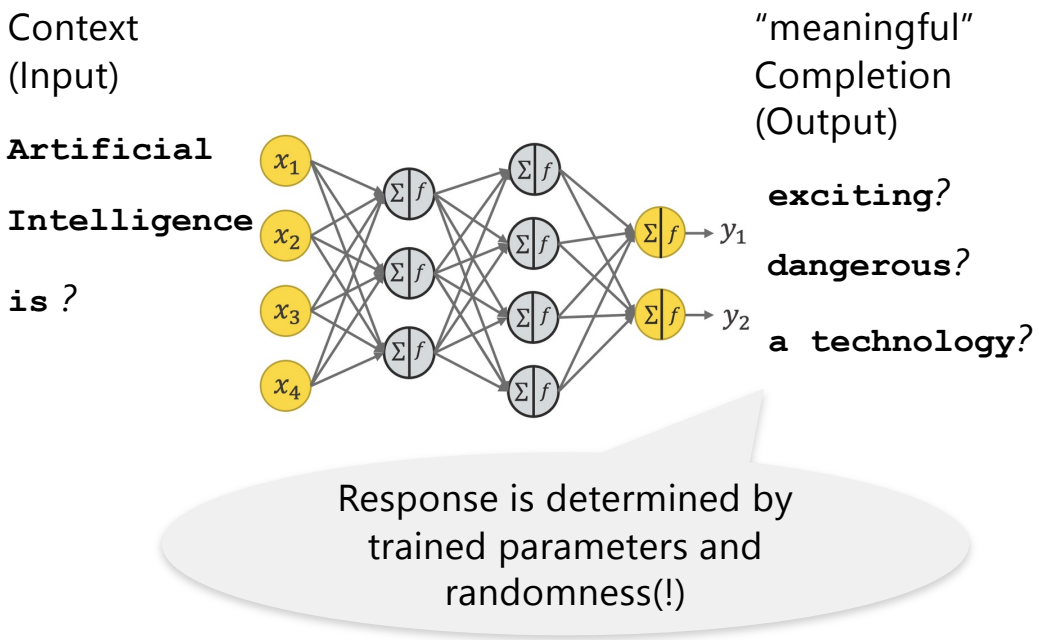




- Model: Network of neurons, each neuron is a function with parameters: weights, bias, activation functions
- Training: determine the parameters for each neuron, activation function leads to non-linear behaviour
- Training is costly: many "pairs" of inputs and outputs, many parameters – gradient descent algorithm speeds this up, but it's still costly
  Computations can be expressed as vector-/matrix-operations => GPUs
- Well explored application: character recognition (OCR)

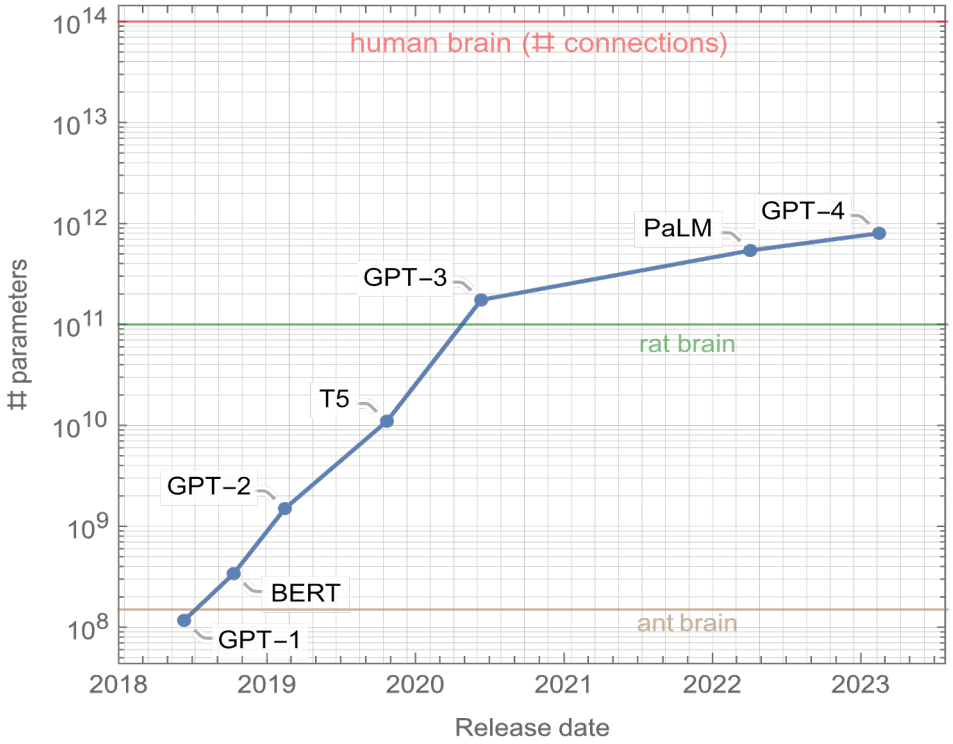
How do LLMs work?

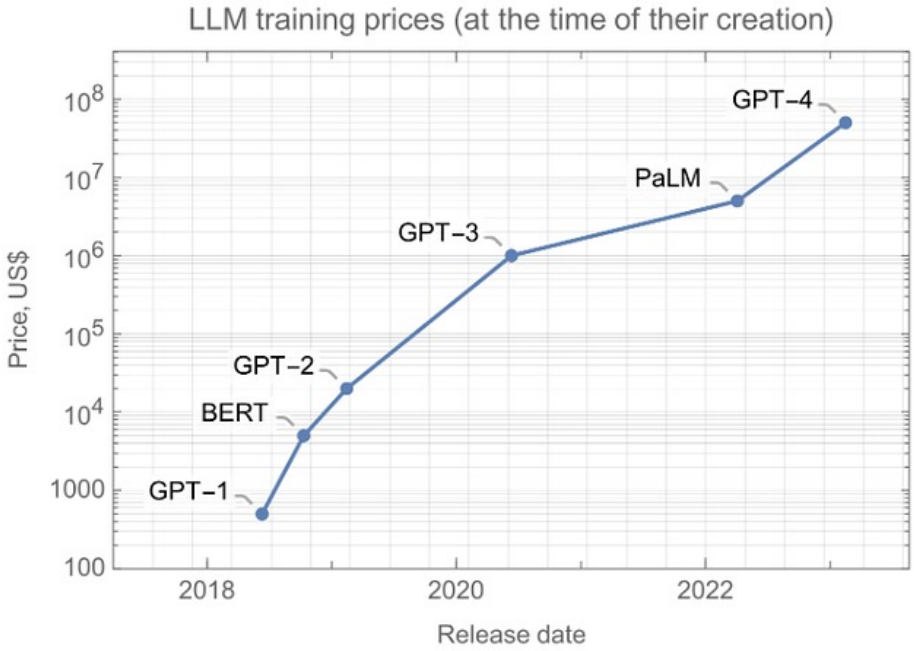
- Neural networks are used to generate/complete content (token by token)
- Why is suddenly possible
  - Network architecture, network size, training methodology
  - Computation power



- Two training phases:
  - Pretraining: Fully automatic, using a large base of knowledge (internet, libraries,…)
  - Fine-Tuning: Improvement of the parameters with high quality question/answer pairs/scenarios and expert feedback

Complexity

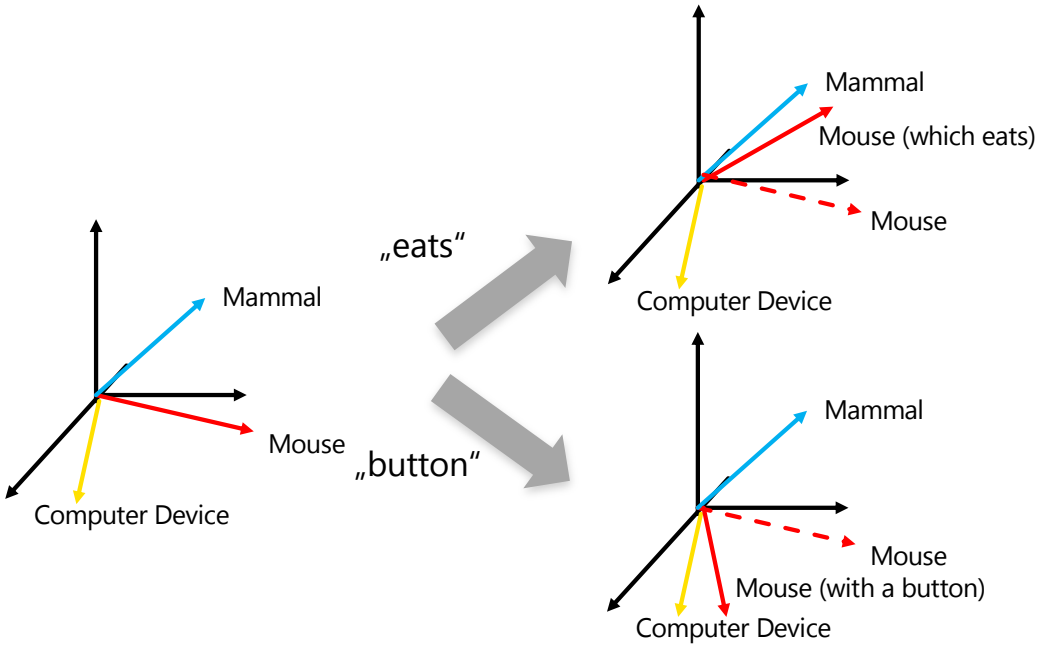




Source: https://www.numind.ai/blog/what-are-large-language-models

How does the LLM „know“ what we are talking about?

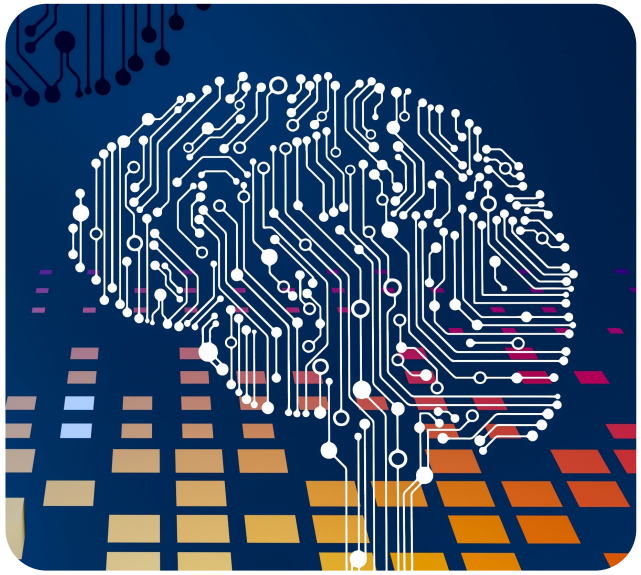
- Tokens (words) are encoded as very large vectors („embedding“)
- Vectors are manipulated by the previously trained LLM – while they „flow“ through the network, they are step-by-step refined towards a meaning (via matrix computations) („attention“) – note, that LLMs are not “simple forward-flowing networks”
- Similar vectors represent similar concepts
- At the end, vectors are converted back to tokens



My take on LLMs

- Large time savings for routine activities: well-trained LLMs such as ChatGPT are strong at generating content (texts, including source code)
- Efficiency: LLMs compress knowledge, but have high hardware requirements
- Abstraction: LLMs "use" patterns without human intervention, but cannot draw any conclusions (abductive reasoning)
- Emergence: contexts (prompts), chance and patterns create "new things"
- Correctness? Creativity? Hallucination? - a question of training?

Source: https://www.youtube.com/watch?v=2IK3DFHRFfw



Reminder:
LLMs work on correlation – not causality
– they do not reason, they are not really creative!

Tip: Watch https://www.youtube.com/watch?v=LPZh9BOjkQs by 3Blue1Brown