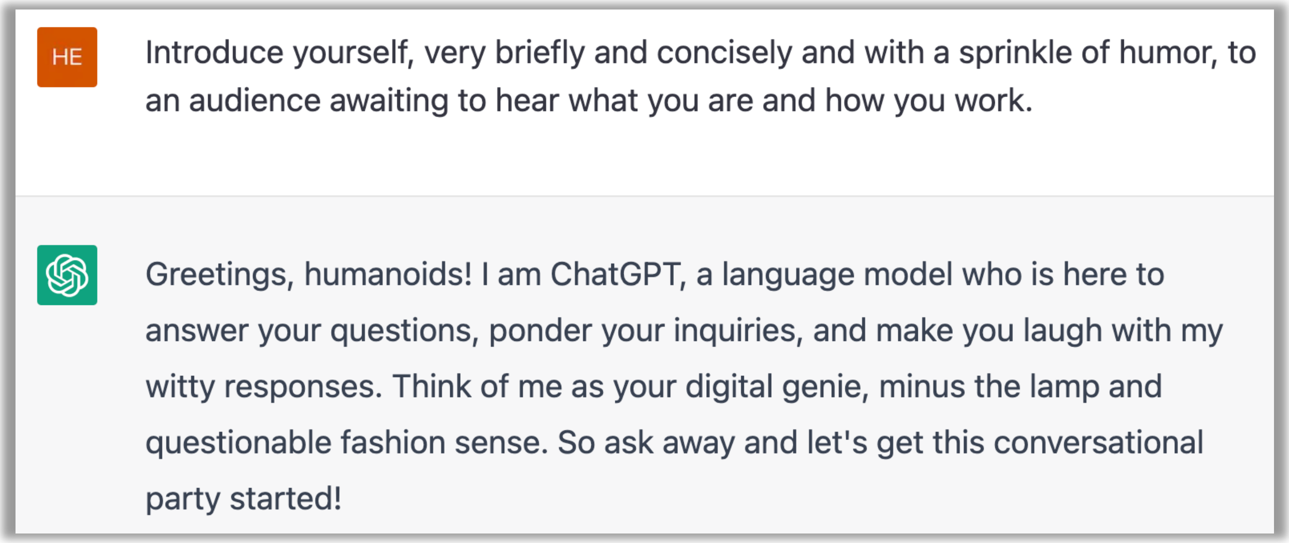
Explaining the Sensation: An Accessible Introduction to ChatGPT

Dr. Michael Heck

13th March 2023



Introduce yourself, very briefly and concisely and with a sprinkle of humor, to an audience awaiting to hear what you are and how you work.

Greetings, humanoids! I am ChatGPT, a language model who is here to answer your questions, ponder your inquiries, and make you laugh with my witty responses. Think of me as your digital genie, minus the lamp and questionable fashion sense. So ask away and let's get this conversational party started!

The Omnipresence of **Natural Language Processing** (NLP)

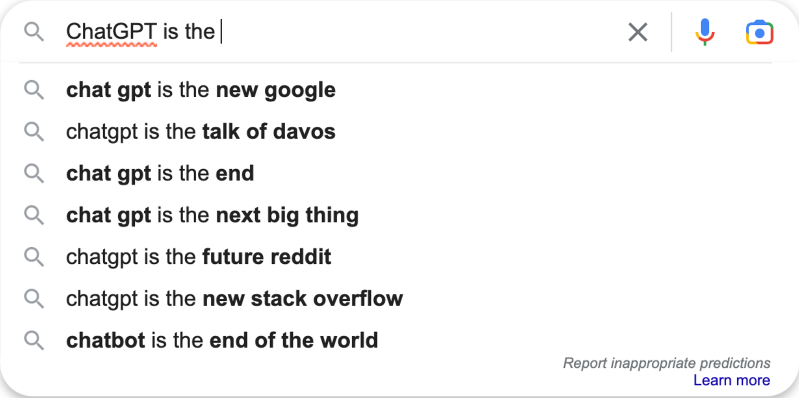
- Everyday tech relies on solutions for **NLP**, powered by **artificial Intelligence** (AI)
  - NLP is so pervasive that all of you are affected by it, often without realizing

Enabling computers to understand natural language

Relies on **machine learning**, e.g., **neural networks**, **deep learning**

The Omnipresence of **Natural Language Processing** (NLP)

Search



ChatGPT is the

chat gpt is the **new google**

chatgpt is the **talk of davos**

chat gpt is the **end**

chat gpt is the **next big thing**

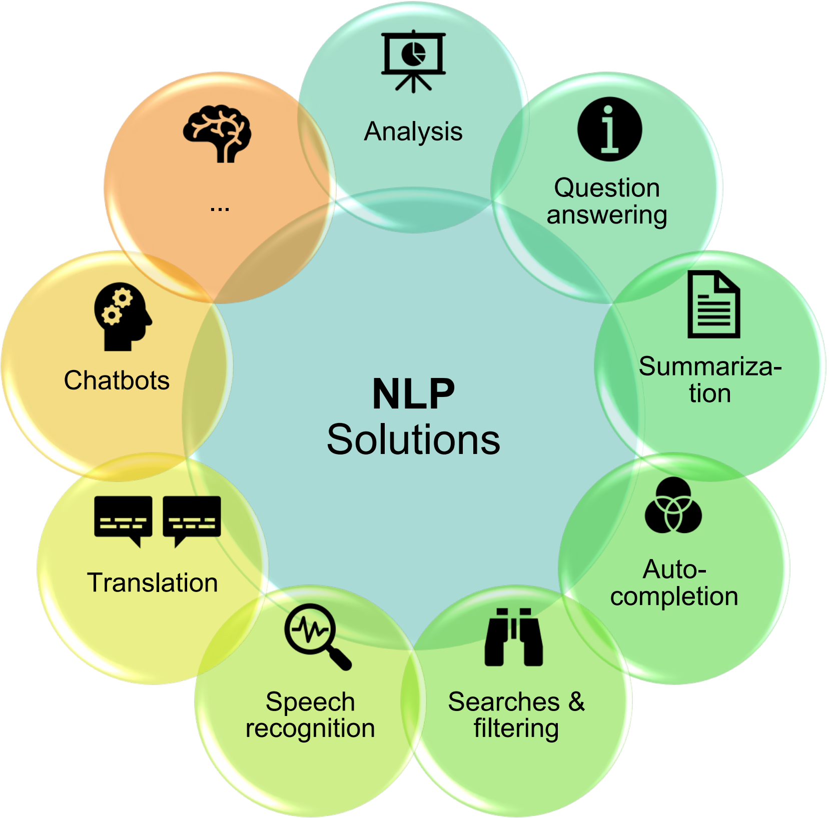
chatgpt is the **future reddit**

chatgpt is the **new stack overflow**

chatbot is the **end of the world**

Report inappropriate predictions
Learn more

The Omnipresence of **Natural Language Processing** (NLP)



NLP
Solutions

Analysis

Question answering

Summarization

Auto-completion

Searches & filtering

Speech recognition

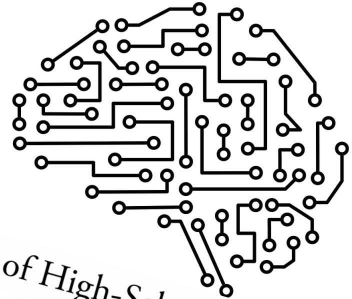
Translation

Chatbots

...

AI Takes Center Stage: ChatGPT

- AI powered chatbot released by OpenAI in November 2022
- Abilities from writing to coding, reasoning to creativity



What is AI chatbot phenomenon ChatGPT and could it replace humans?

The End of High-School English

AI is finally good at stuff, and that's a problem

AI chatbots are here and they're passing medical exams – should we embrace them?

Will ChatGPT make lawyers obsolete? (Hint: be afraid)

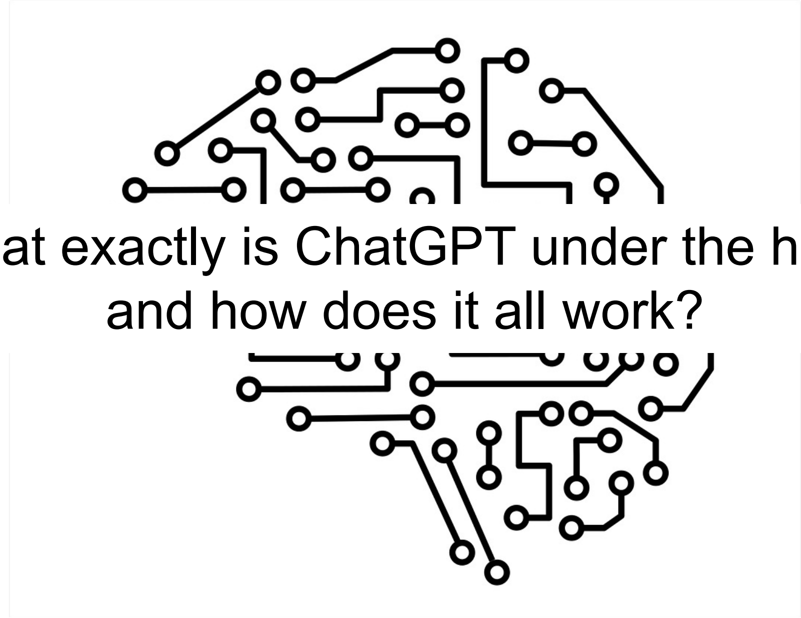
The ChatGPT chatbot from OpenAI is amazing, creative, and totally wrong

AI bot ChatGPT stuns academics with essay-writing skills and usability

Nick Cave calls ChatGPT and AI songwriting 'a grotesque mockery of what it is to be human'

AI Takes Center Stage: ChatGPT

- ChatGPT reaches state-of-the-art performance on some NLP tasks
- Wide adaptation to write essays, code and more is already happening

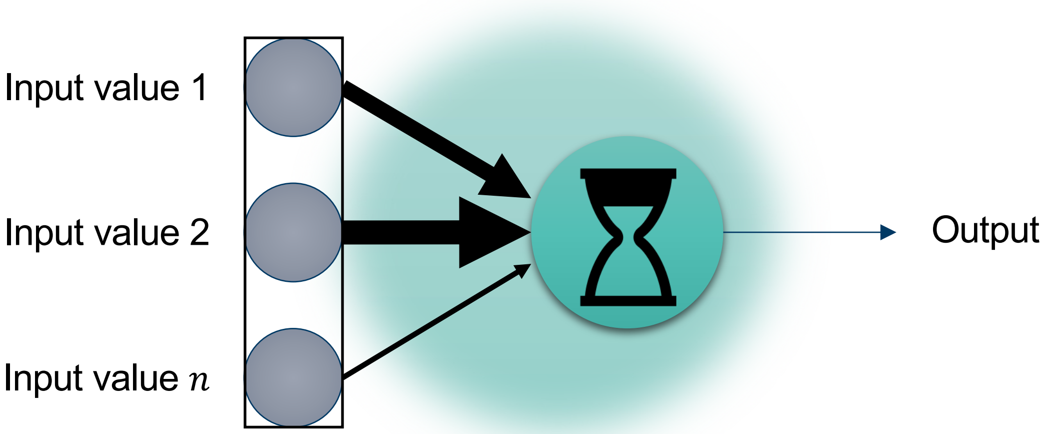


What exactly is ChatGPT under the hood and how does it all work?

ChatGPT Is… a Neural Network

What Are Neural Networks (NNs)?

- Computing systems comprised of **layers** of nodes called **neurons**
- Neurons accept numerical inputs, then weight and average and transform them

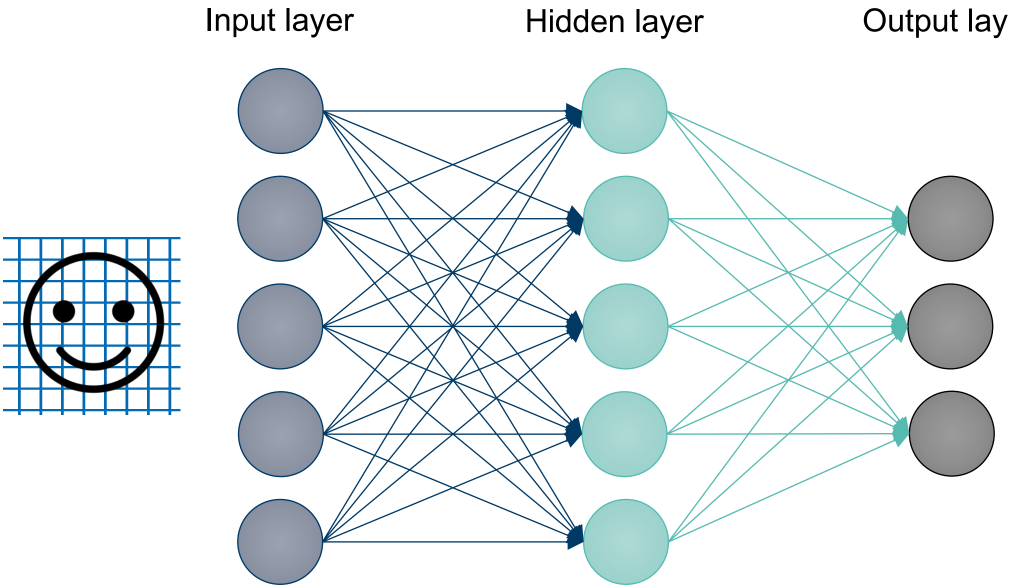


A single neuron

ChatGPT Is… a Neural Network

What Are Neural Networks (NNs)?

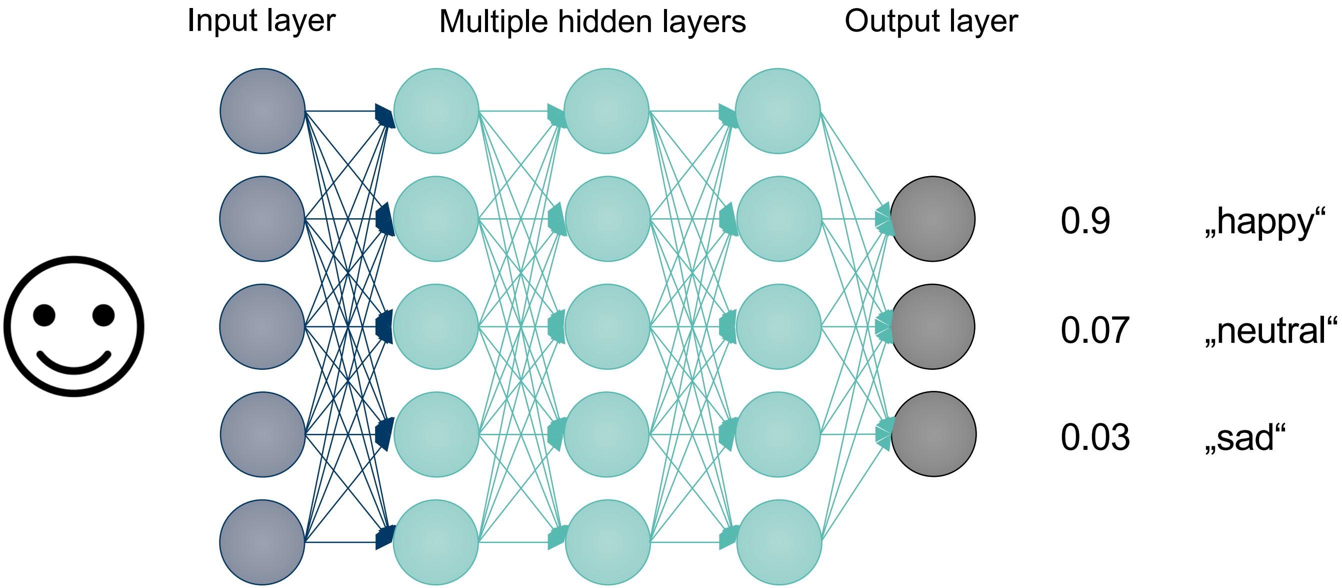
- The output of one neuron becomes the input to subsequent neurons



ChatGPT Is… a Neural Network

What Is Deep Learning?

- Neural networks with 3 or more hidden layers are considered **deep**



Input layer

Multiple hidden layers

Output layer

0.9 „happy“

0.07 „neutral“

0.03 „sad“

ChatGPT Is… a Neural Network

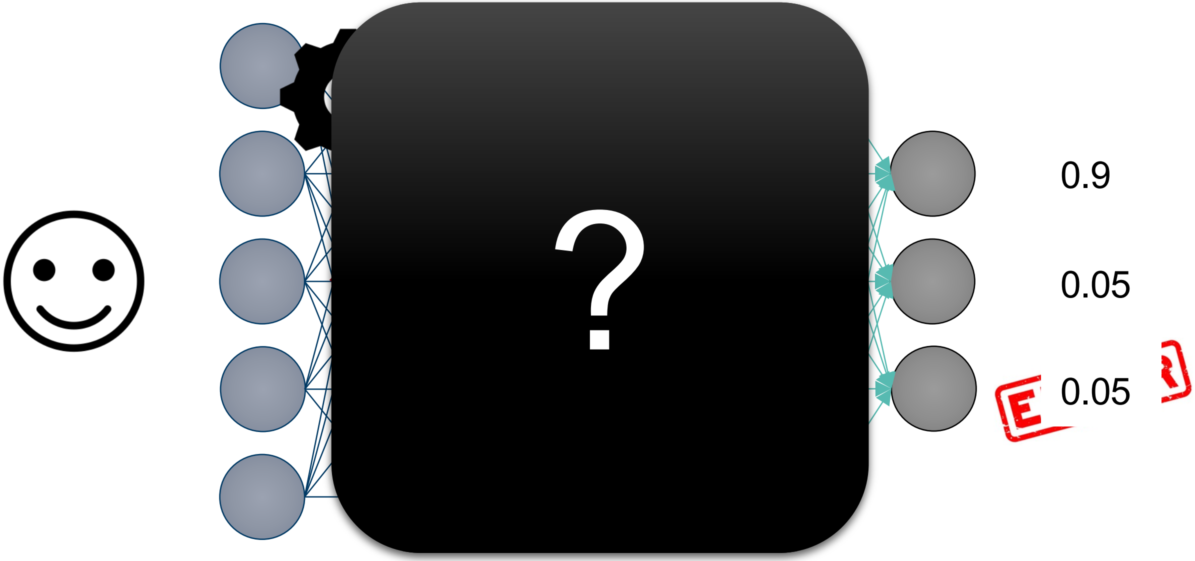
How Does Training Work?

- Deep NNs learn to transform input layer by layer to best solve a task

Input layer

Multiple hidden layers

Output layer



0.9	„happy“
0.05	„neutral“
0.05	„sad“

ChatGPT is… a Language Model

What Are Language Models (LMs)?

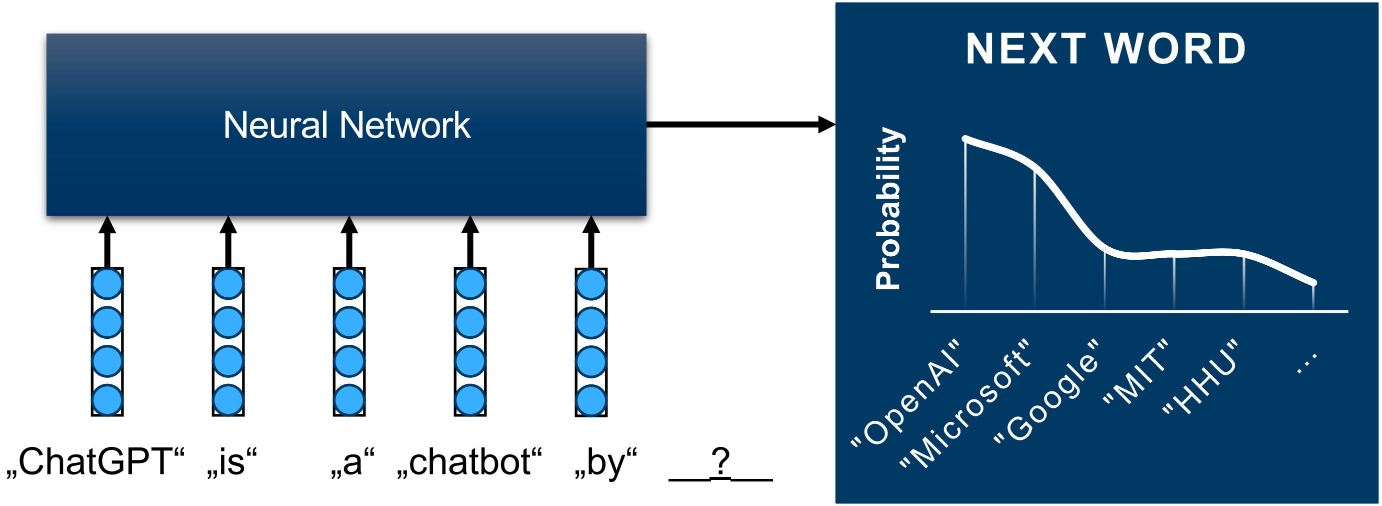
- Statistical models that capture the probability of sequences of words
- Learns word (sequence) probabilities by analyzing large text data

ChatGPT is an artificial intelligence chatbot developed by OpenAI and launched in November 2022. It is built on top of OpenAI's GPT-3 family of large language models and has been fine-tuned (an approach to transfer learning) using both supervised and reinforcement learning techniques. ChatGPT was launched as a prototype on November 30, 2022, and quickly garnered attention for its detailed responses and articulate answers across many domains of knowledge. Its uneven factual accuracy, however, has been identified as a significant drawback. Following the release of ChatGPT, OpenAI's valuation was estimated at US$29 billion in 2023. ChatGPT – a generative pre-trained

ChatGPT is… a Language Model

What Are Language Models (LMs)?

- Statistical models that capture the probability of sequences of words
- Word (sequence) probabilities are determined by analyzing text data
- Neural LMs take context as input and predict a probability distribution for the next word



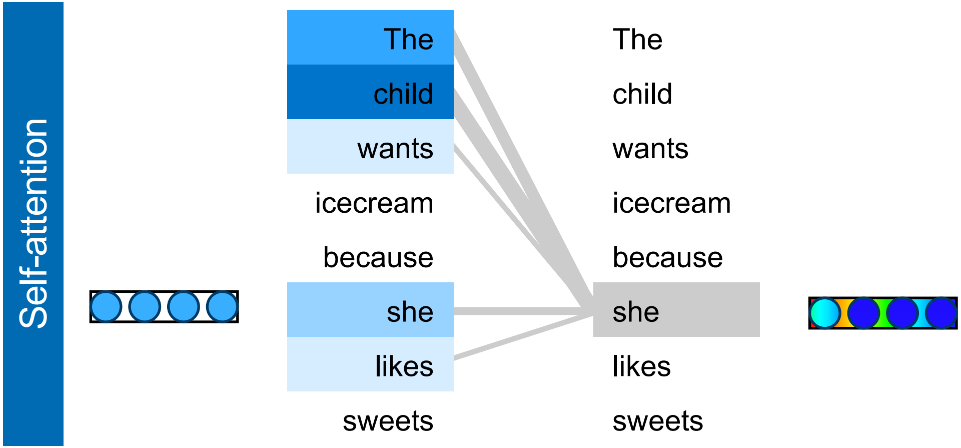
Neural Network

„ChatGPT“ „is“ „a“ „chatbot“ „by“ ___?___

NEXT WORD

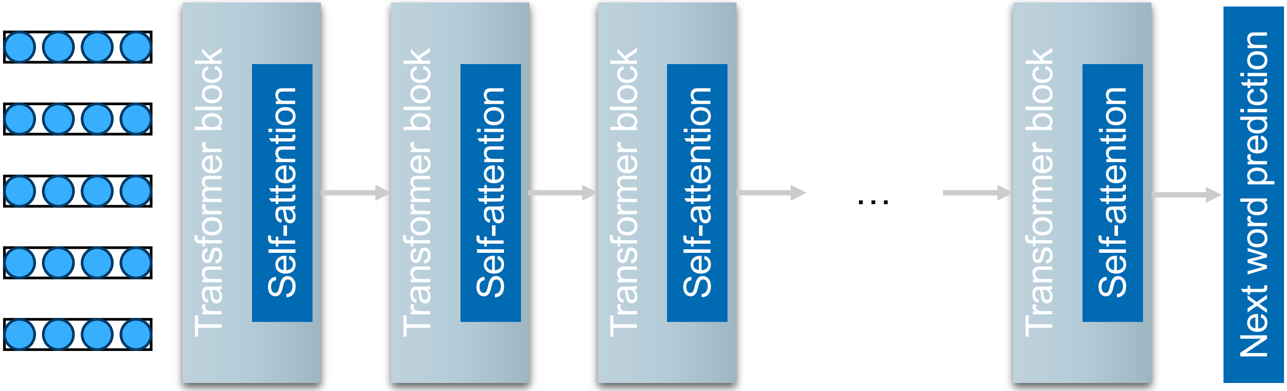
ChatGPT is… a Large Language Model

- Massive in size (hundreds of billions of parameters)
- Learn from massive data (essentially „the internet“)
- Simple training objective, e.g. next word prediction
- **Transformer** architecture



ChatGPT is… a Large Language Model

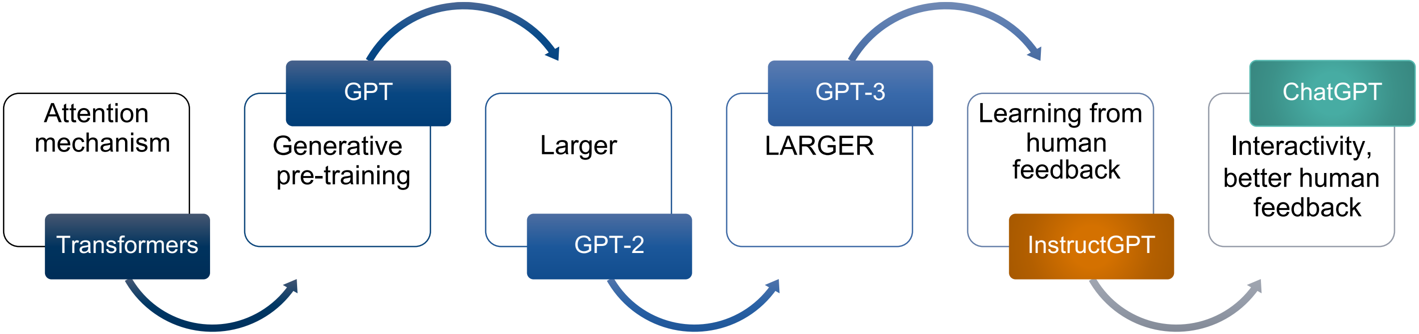
- Massive in size (billions of parameters)
- Learn from massive data („the internet“)
- Simple training ojective, e.g. next word prediction
- **Transformer** architecture
- Learns contextual relationships very well



Evolutionary History of ChatGPT

Evolutionary History of ChatGPT

- Language generation performance of GPT models mainly driven by
  - Increased number of parameters (via deeper and wider architectures)
  - Increased amount of training data



Attention mechanism

Transformers

GPT

Generative pre-training

Larger

GPT-2

GPT-3

LARGER

Learning from human feedback

InstructGPT

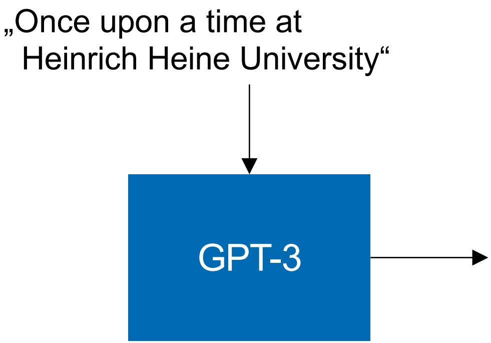
ChatGPT

Interactivity, better human feedback

	GPT	GPT-2	GPT-3
Parameters:	120 Million	1.5 Billion	175 Billion
Training text:	4.5 Gigabyte	40 Gigabyte	Several Terabyte

GPT-3

- Auto-regressive LM
  - Last output is next input
- Good at language modeling
- Weak at following user intent



„Once upon a time at
Heinrich Heine University“

SPOILER ALERT

The Issue of Alignment

- LMs excel at predicting next word, given some context
  - Does not necessarily **align** with users‘ expectations
  - **Misaligned** for following instructions

Explain the moon landing to a 6 year old

A Explain gravity to a 6 year old

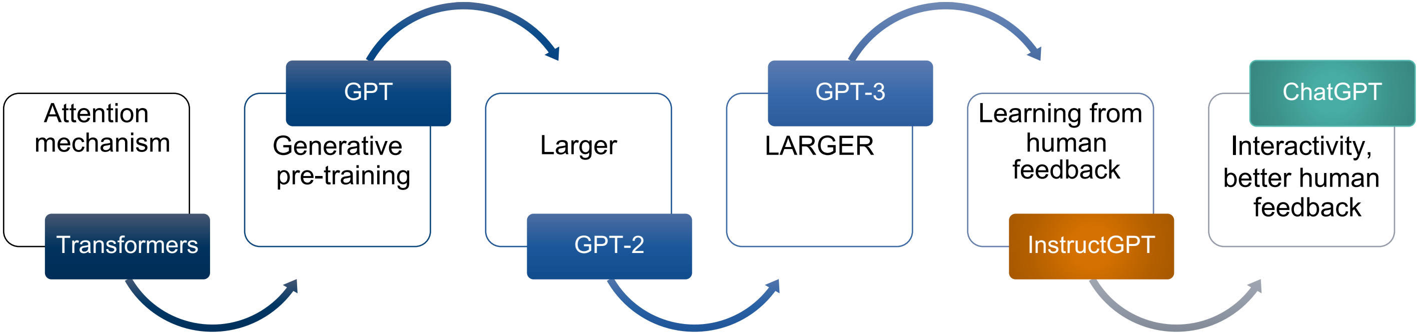
B The moon landing was a hoax!

C Moon is a natural satellite of planet Earth

D The moon landing was when people went to the moon with a spacecraft.

Evolutionary History of ChatGPT

- Usefulness of InstructGPT mainly driven by
  - **Aligning** by learning from human feedback

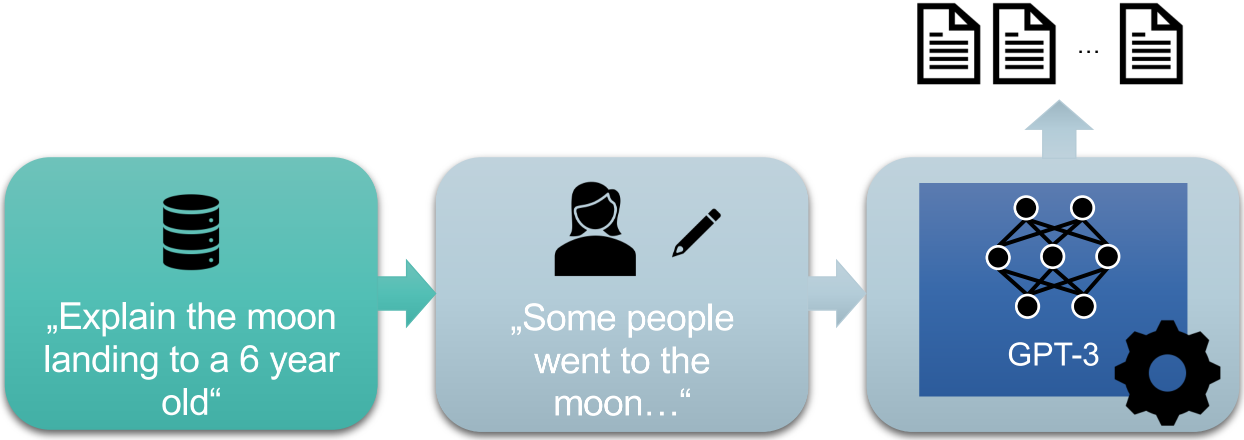


Parameters:	120 Million	1.5 Billion	175 Billion
Training text:	4.5 Gigabyte	40 Gigabyte	Several Terabyte

InstructGPT, the „Older Sibling“

Training Step 1: Supervised fine-tuning

- Use demonstration data to **fine-tune** GPT-3
  - Fine-tuning with human demonstrations improves usefulness
  - **Aligning** to user intent


…
„Explain the moon landing to a 6 year old“
„Some people went to the moon…“
GPT-3

Sample prompt from dataset

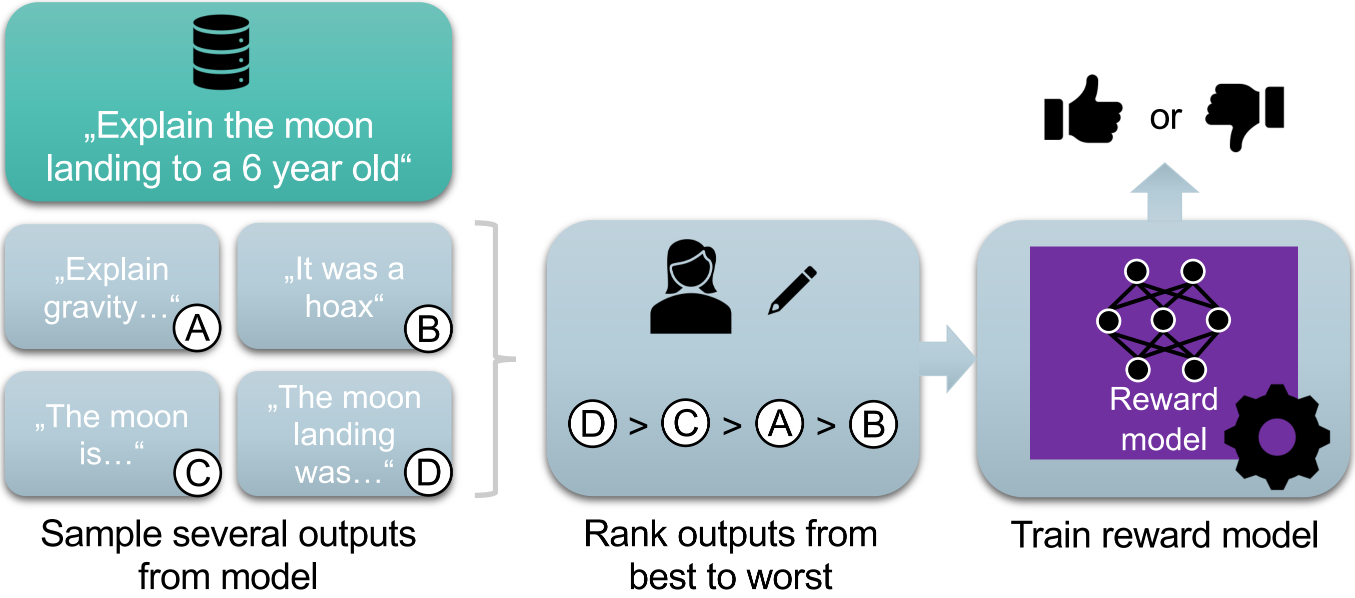
Provide desired output

Fine-tune GPT-3

InstructGPT, the „Older Sibling“

Training Step 2: Building a Reward Model

- Use comparison data to train a reward model to simulate human feedback



Sample several outputs from model

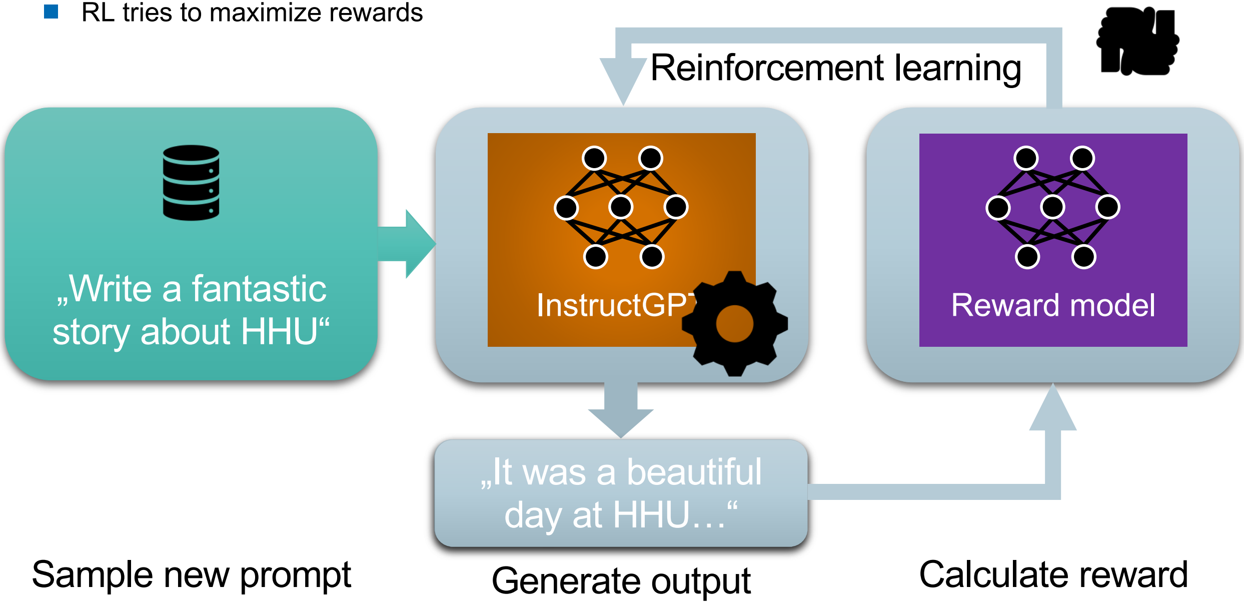
Rank outputs from best to worst

Train reward model

InstructGPT, the „Older Sibling“

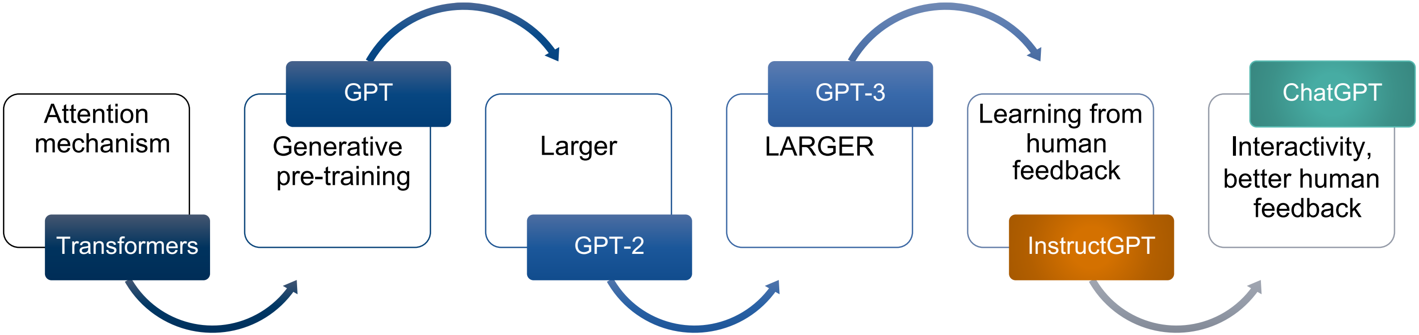
Training Step 3: Optimizing with Reinforcement Learning

- Use reinforcement learning (RL) with the reward model to optimize model behavior
- GPT-3 is tuned towards following user preferences to improve alignment
  - RL tries to maximize rewards



Evolutionary History of ChatGPT

- Usefulness and safety of ChatGPT mainly driven by
  - Chat-based interactivity
  - Learning from more and better human feedback
  - Automation of feedback loop for massive upscaling



Attention mechanism

Transformers

GPT

Generative pre-training

Larger

GPT-2

GPT-3

LARGER

Learning from human feedback

InstructGPT

ChatGPT

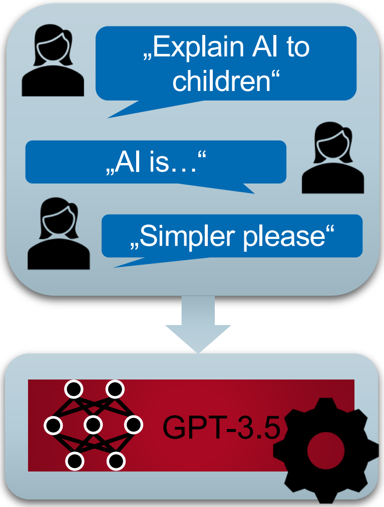
Interactivity, better human feedback

Parameters:	120 Million	1.5 Billion	175 Billion
Training text:	4.5 Gigabyte	40 Gigabyte	Several Terabyte

From InstructGPT to ChatGPT: What We Know

Step 1: Supervised fine-tuning

- Training focuses on dialogue and interactivity (follow-up questions, refining answers, etc.)
- Improved base model GPT-3.5 (trained on more text & code, produces longer outputs)



„Explain AI to children“

„AI is…“

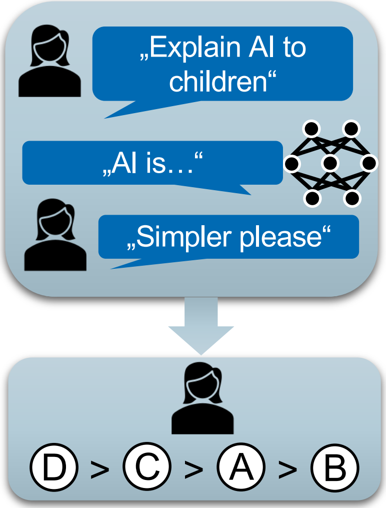
„Simpler please“

GPT-3.5

From InstructGPT to ChatGPT: What We Know

Step 2: Building a Reward Model

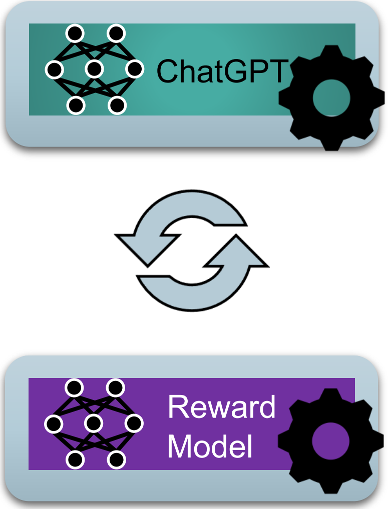
- Training focuses on dialogue and interactivity (follow-up questions, refining answers, etc.)



From InstructGPT to ChatGPT: What We Know

Step 3: Optimizing with Reinforcement Learning

- ChatGPT and Reward Model are optimized periodically
- Automation of feedback loop for massive upscaling



ChatGPT

Answer follow-up questions

SORRY

Admit mistakes

Challenge incorrect premises

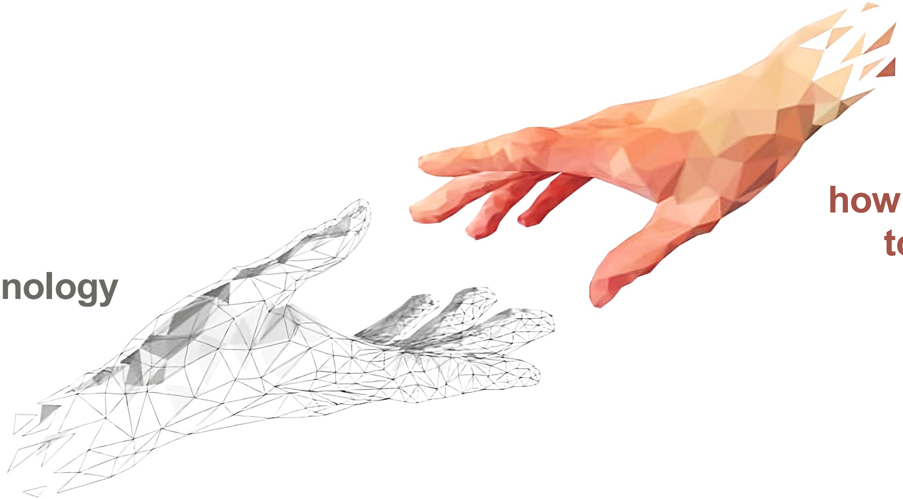
Reject inappropriate requests

Closely follow user intent

Generate safer output

Where We Are Now

- Extremely sophisticated word sequence predictor
  - Prime example of applied science, using existing methods at scale
- ChatGPT is representative of an entire class of new AI models
  - Extreme **generalization** helps solve tasks never explictly learned thanks to **aligning** and **scaling**

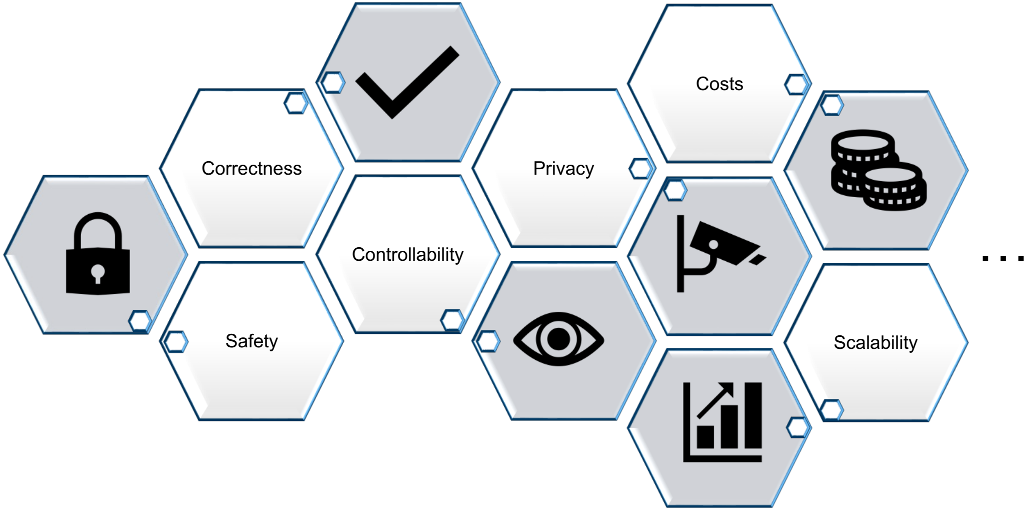


The technology is here,

how are we going to embrace it?

Where We Are Heading

- ChatGPT and related models in many ways are *superhuman*, but…
  - … all current AI solutions, including ChatGPT, are still **weak AI**
  - … all current LLMs, including ChatGPT, share the same limitations



Thank you!