

ChatGPT: the parrot in the classroom?

Ildikó Pilán

Snakk om seminar, 22 March 2023



ChatGPT can pass the bar exam in 90th percentile

Will ChatGPT make coding tests for engineers obsolete?

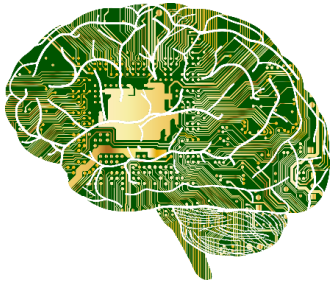
AI in higher education – A tool for better learning?



ChatGPT can pass the bar exam in 90th percentile

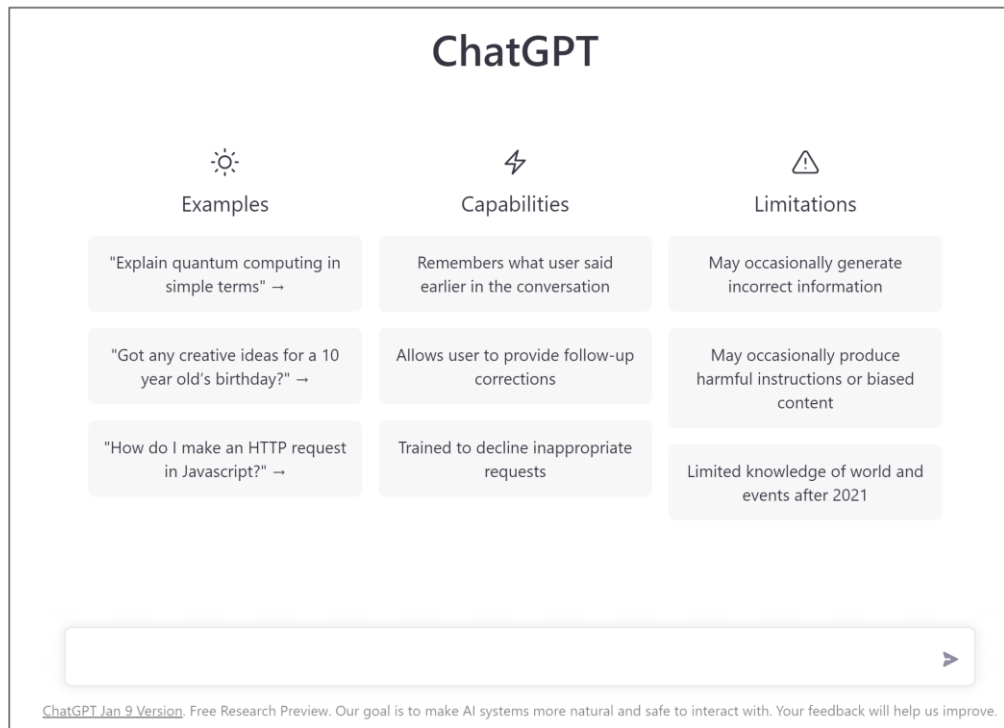
Will ChatGPT make coding tests for engineers obsolete?

AI in higher education – A tool for better learning?



What is ChatGPT?

- ▶ A **chatbot** interacting in a conversational way.
- ▶ Aims at **generating** human-like **text** as response.
- ▶ Relies on large AI-based **language models** (GPT).
- ▶ Released in Nov 2022 by OpenAI (but the technology is from 2018).



Example: text correction + feedback

PI

Can you correct and give feedback on this: Jeg sitter og tester en verktøy. Det er effektiv men jeg ikke er sikker hvordan jeg kan bruke det i skolen.

Example: text correction + feedback

PI

Can you correct and give feedback on this: Jeg sitter og tester en verktøy. Det er effektiv men jeg ikke er sikker hvordan jeg kan bruke det i skolen.



Sure, I can help you with that! Here's a corrected version of your text with some feedback:



"Jeg sitter og tester et verktøy. Det er effektivt, men jeg er ikke sikker på hvordan jeg kan bruke det på skolen."

Feedback:

- "Verktøy" is a neuter noun in Norwegian, so you should use "et" instead of "en" to indicate the correct gender.
- You used "effektiv" correctly, but you forgot to add the "-t" at the end to match the neuter gender of "verktøy".
- In the second sentence, it's better to say "på skolen" instead of "i skolen". This is because "skolen" refers to a place where you are present, so you use "på" to indicate that you are physically at the school.

Example: hallucinations



When can you see the northern lights in Hungary?



The Northern Lights, also known as the Aurora Borealis, are typically visible in Hungary during the winter months when the sky is clear and dark. The best time to see them is typically between September and April, with the peak activity occurring between November and February. However, it should be noted that the northern lights are not visible every night, and sightings are not guaranteed.





VS.



- ▶ Input: **keywords** of simple questions
 - ▶ Result: **web pages**
 - ▶ Information from a **matching document** collection (internet)
 - ▶ Reliability of information depends on web **page creator**
- ▶ Input: dialogue-like **prompts**
 - ▶ Result: **generated** human-like **text** (Story, explanation, translation etc.)
 - ▶ Information from **patterns** previously **learned** from training data
 - ▶ Reliability of information depends on **algorithm & training data**



VS.



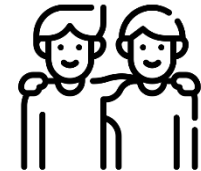
- ▶ Input: **keywords** of simple questions
- ▶ Result: **web pages**
- ▶ Information from a **matching document** collection (internet)
- ▶ Reliability of information depends on web **page creator**

- ▶ Input: dialogue-like **prompts**
- ▶ Result: **generated** human-like **text** (Story, explanation, translation etc.)
- ▶ Information from **patterns** previously **learned** from training data
- ▶ Reliability of information depends on **algorithm** & training **data**

“asking a librarian”



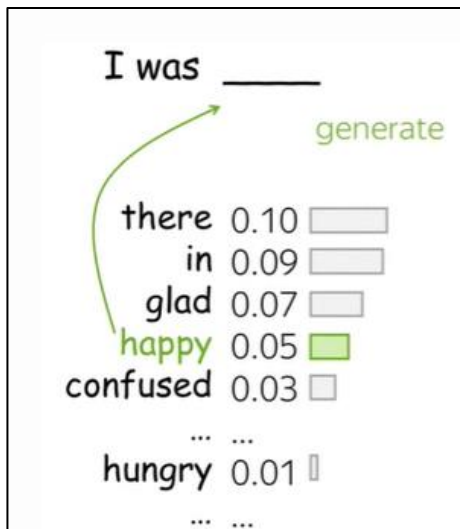
“asking a friend”



GPT: Generative Pre-trained Transformer



next word prediction

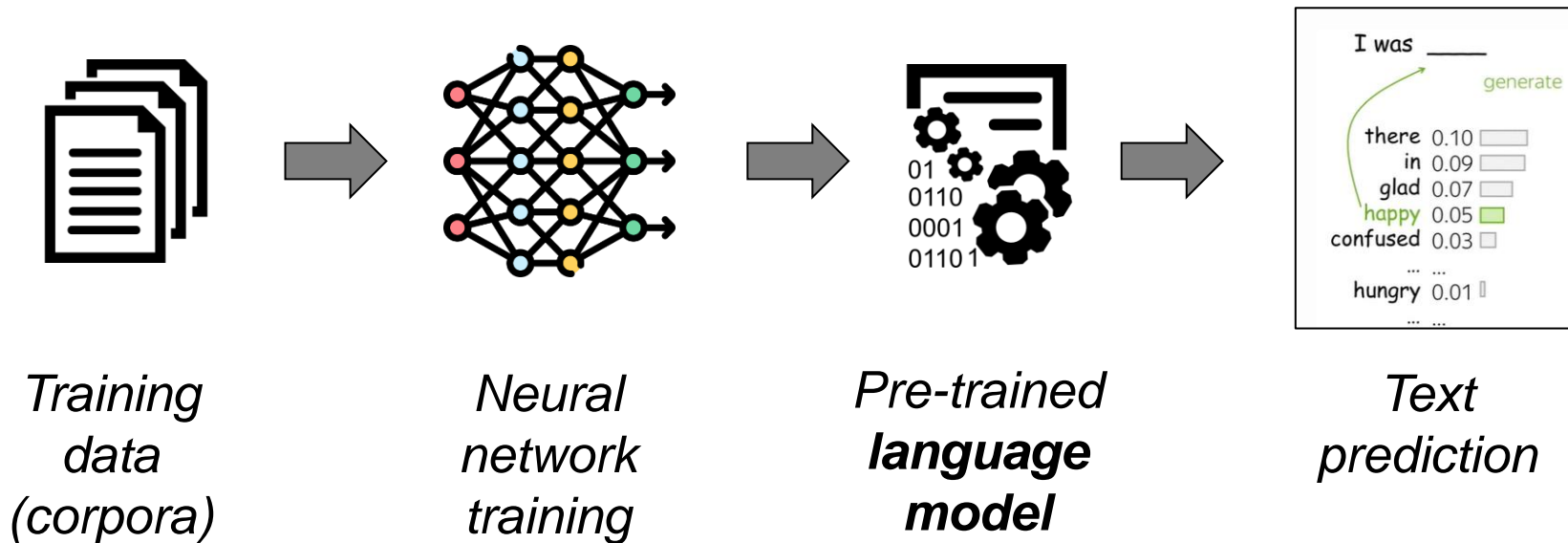


Probabilities learned from data:
Wikipedia, news, online forums,
web texts, books

I was there in 1990.
I was there, were you?
I was in a good shape then.
We were in Oslo last year.
I was glad it happened.
I was happy you arrived.
I was confused about everything.
They were hungry before lunch.

... ..

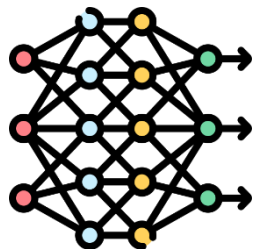
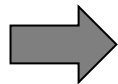
GPT: Generative **Pre-trained** Transformer



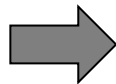
GPT: Generative **Pre-trained** Transformer



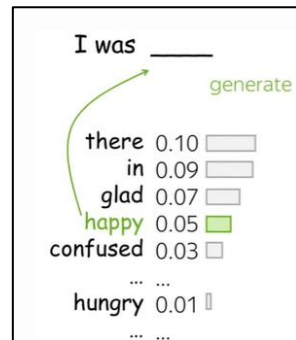
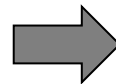
*Training
data
(corpora)*



*Neural
network
training*



*Pre-trained
language
model*



*Text
prediction*

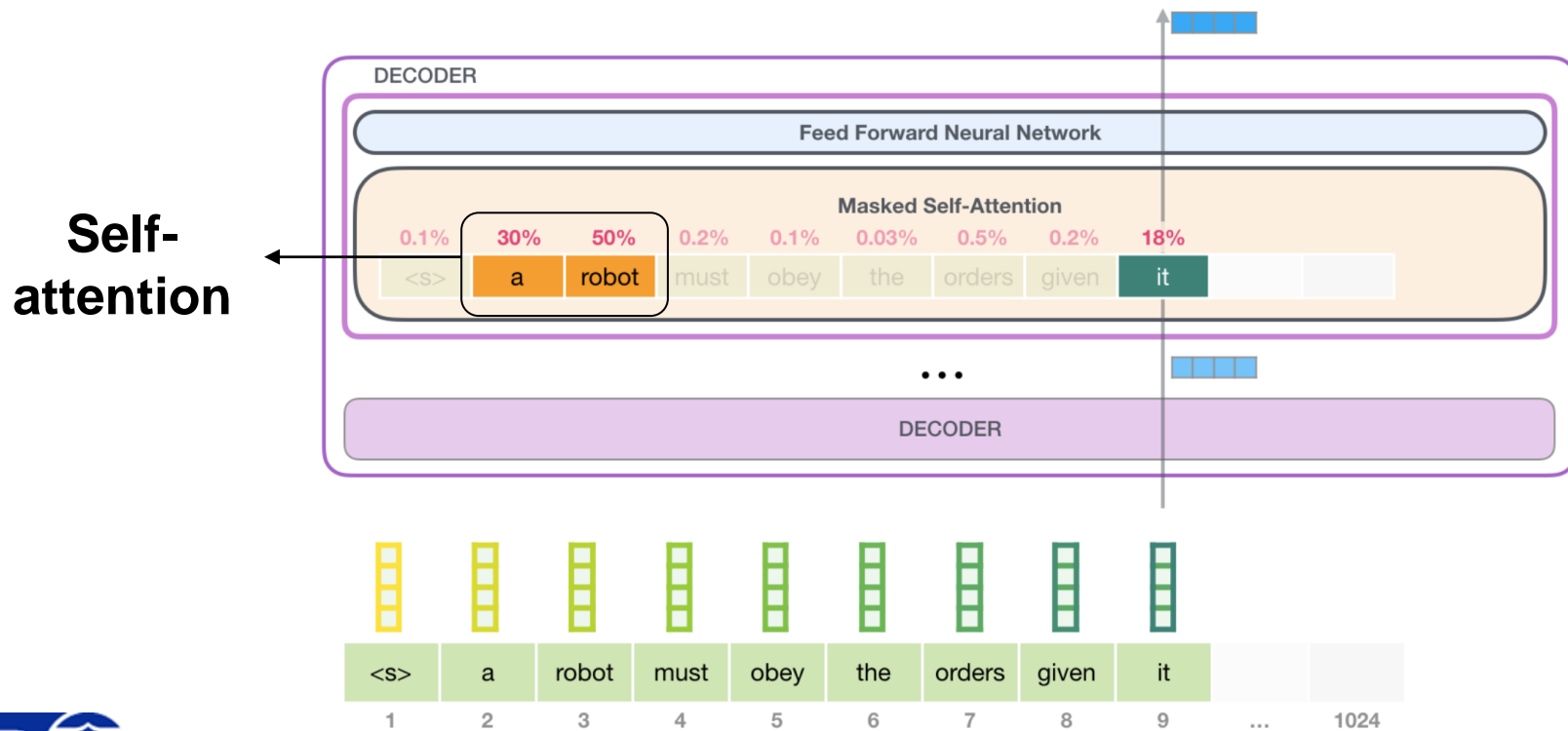
Fine-tuning



*Text
classification*




GPT: Generative Pre-trained Transformer



Inside language models: vectors

Context (left)	Target	Context (right)
<i>The happy</i>	<i>cat</i>	<i>meowed.</i>
<i>A small</i>	<i>kitten</i>	<i>meows desperately.</i>
<i>My neighbor's</i>	<i>dog</i>	<i>barks all night.</i>



<i>cat</i> →	0.6	0.9	0.1	0.4	-0.7	-0.3	-0.2
<i>kitten</i> →	0.5	0.8	-0.1	0.2	-0.6	-0.5	-0.1
<i>dog</i> →	0.7	-0.1	0.4	0.3	-0.4	-0.1	-0.3

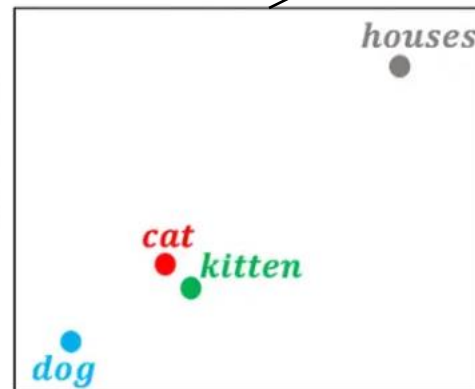
Inside language models: vectors

Context (left)	Target	Context (right)
<i>The happy</i>	cat	<i>meowed.</i>
<i>A small</i>	kitten	<i>meows desperately.</i>
<i>My neighbor's</i>	dog	<i>barks all night.</i>
<i>The agency sold three</i>	houses	<i>with a pool.</i>



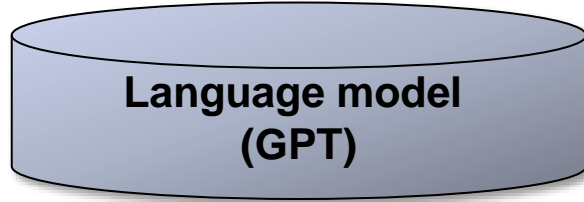
cat →	0.6	0.9	0.1	0.4	-0.7	-0.3	-0.2
kitten →	0.5	0.8	-0.1	0.2	-0.6	-0.5	-0.1
dog →	0.7	-0.1	0.4	0.3	-0.4	-0.1	-0.3
houses →	-0.8	-0.4	-0.5	0.1	-0.9	0.3	0.8

Dimensionality
reduction of
word
embeddings
from 7D to 2D



**Words with similar
contexts are closer
together.**

Language models: ChatGPT's base



Reinforcement learning in ChatGPT

- Clues about human expectations
- Increases helpfulness



Reinforcement Learning with
Human Feedback (RLHF)



Language model
(GPT)



ChatGPT



Explain reinforcement
learning to a 6 year old.

A

In reinforcement
learning, the
agent is...

B

Explain rewards...

C

In machine
learning...

D

We give treats and
punishments to
teach...



D > C > A > B

RM



D > C > A > B

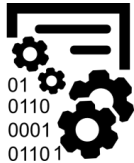
Safety layer in ChatGPT



Safety



Reinforcement Learning with
Human Feedback (RLHF)



Language model
(GPT)



Aim:

- Refuse harmful requests
- Reduce bias

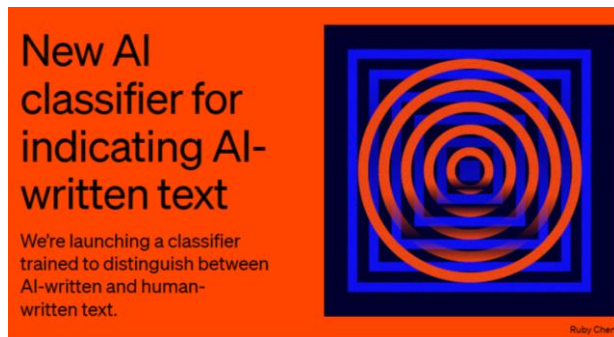


Based on:

- Safety RLHF prompt
- Rules

Latest updates

- ▶ Detection of AI-written text
 - not very reliable though



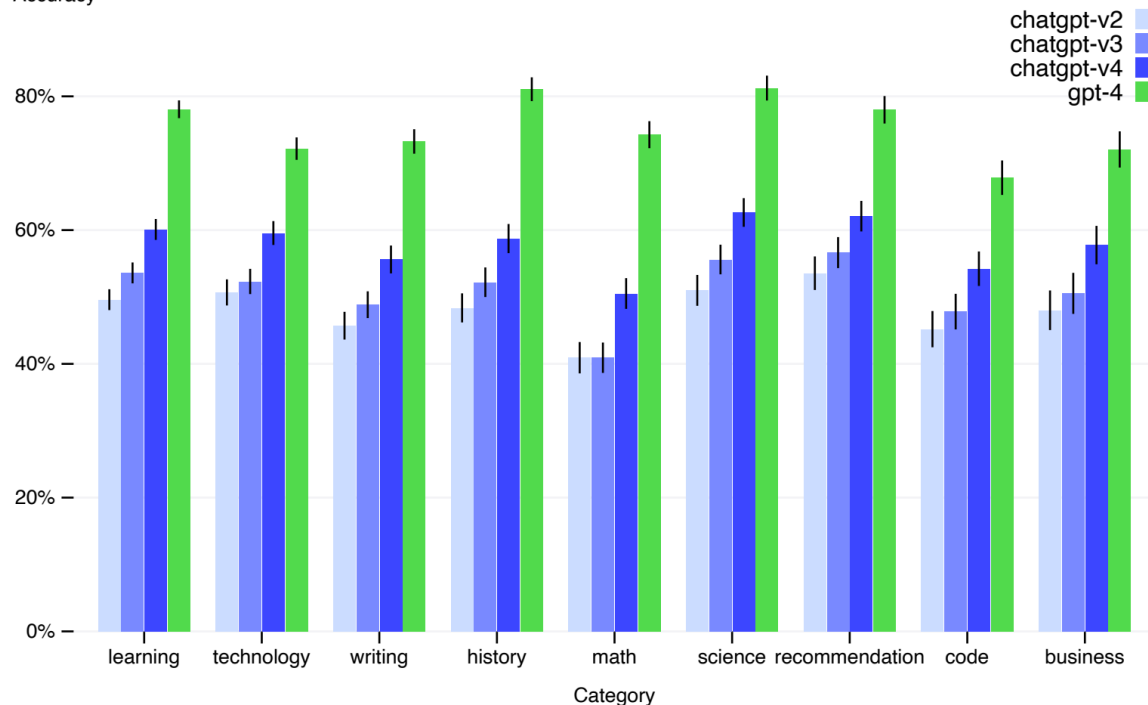
- ▶ New language model: **GPT-4**
 - 14 March 2023 (paid version)



GPT-4: improved factuality

Internal Factual Eval by Category

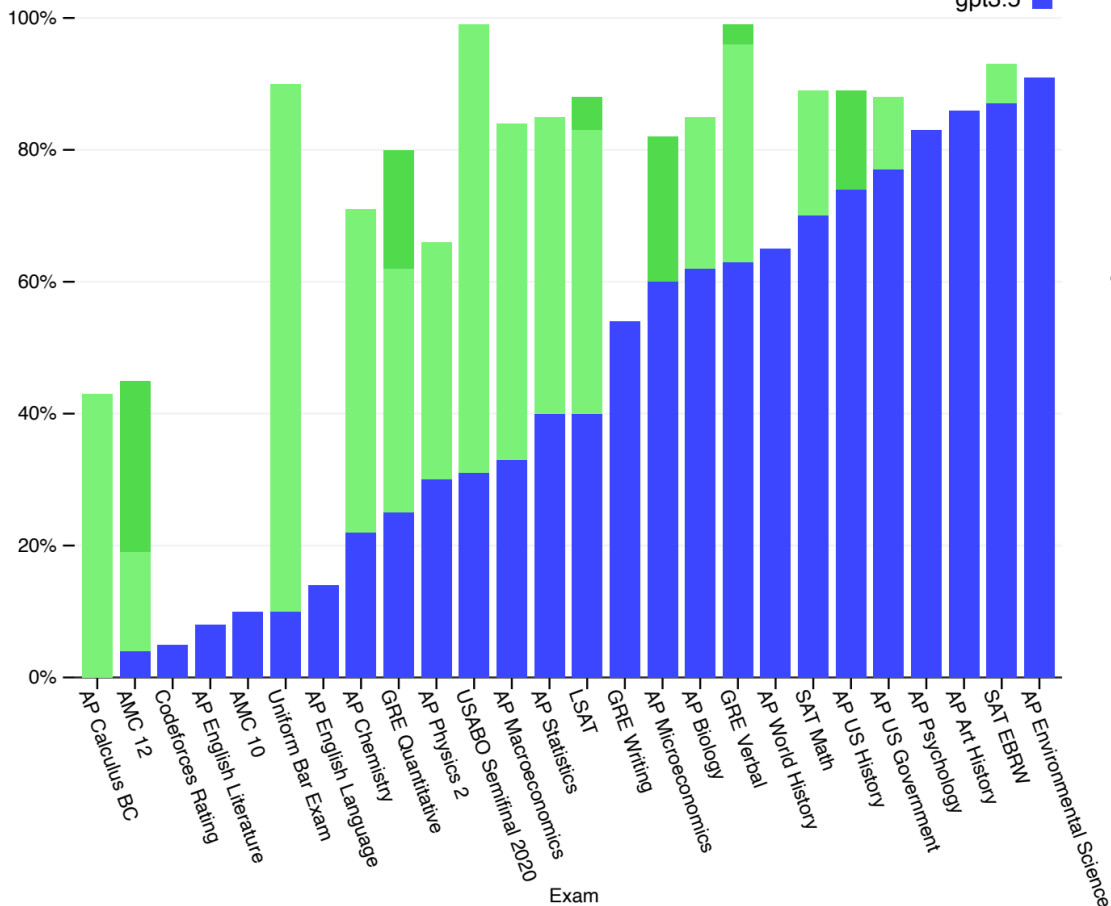
Accuracy



GPT-4 & exams

Exam results (ordered by GPT-3.5 performance)

Estimated percentile lower bound (among test takers)



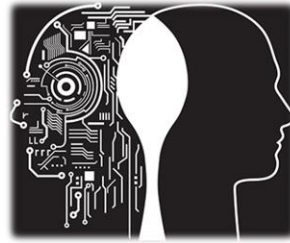
text +
images
as input!

Source:

<https://cdn.openai.com/papers/gpt-4.pdf>

What is ChatGPT good at?

- ▶ Generating and modifying **very** human-like texts
- ▶ Tailoring output to human expectations based on prompts
- ▶ Reproducing some well-known facts available in its underlying data



Educational use cases

- ✓ Edit and improve own writing
- ✓ Personalize assignments
- ✓ Provide feedback

ChatGPT's current limitations

- ▶ Data only **up to September 2021**.
- ▶ Sometimes still **harmful** or **biased**.
- ▶ Plausible-sounding, but **incorrect facts** together with correct ones.
- ▶ Does **not** produce **visual output**.
- ▶ **No reasoning**, common sense or any kind of awareness.



Generated by:  DALL·E