

MachineLearnAthon - Microlecture LLM & Ethics

MachineLearnAthon

A project Co-funded by the Erasmus+ programme of the European Union

27.03.2024

Learning outcomes of today

After successfully completing this lecture, you will be able to...

- Understand what natural language processing, large language models and foundation models are
- Explain the inner working of large language models, their data and the training
- Estimate what limitations there are to LLMs when applied in the chemical engineering domain
- Evaluate the role of trustworthiness in application of AI

Agenda for today

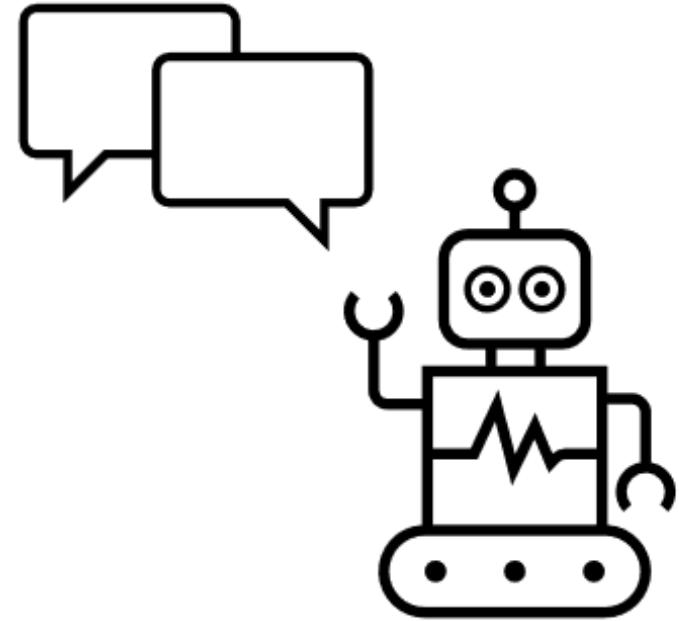
- Definitions of terminology in natural language processing
- Inner working of large language models
- Assessing the utility of large language models
- Ethical implications of AI and large language models

Agenda for today

- **Definitions of terminology in natural language processing**
- Inner working of large language models
- Assessing the utility of large language models
- Ethical implications of AI and large language models

LLMs had and have an impact

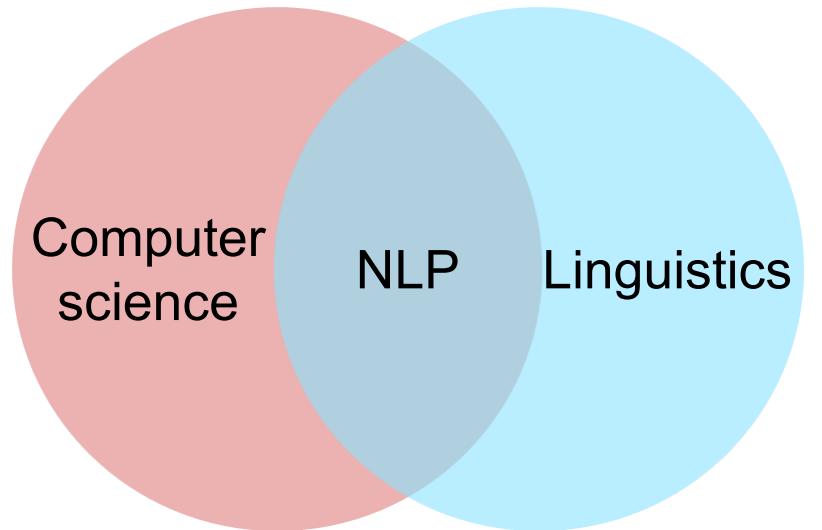
- Large language models are here, and they are here to stay
- For example, the ChatGPT web application had...
 - Fastest growing user base in history
 - 1 million users in its first week
- LLMs can create high quality, confident sounding text used to write assignments, assays or code
- In the science and engineering domain, Large language models could potentially have:
 - an impact on **education**,
 - an impact on innovation and **science**, and
 - long-term or **societal** impacts: e.g. safety



A conversational agent:
A chatbot

Definition of natural language models

“Natural language processing (NLP) is a collection of computational techniques for automatic analysis and representation of human languages, motivated by theory.”



[1] Chowdhary, K. R. (2020). Fundamentals of Artificial Intelligence. Springer. p. 604.

[2] Figure taken from

What is a foundation model?

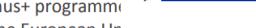
“Foundation models (e.g., BERT, GPT-3, CLIP, Codex) are models **trained on broad data at scale** such that they **can be adapted** to a wide range of **downstream tasks**. These models will not only transform how AI systems are built, but will also lead to significant societal consequences.”



**Center for
Research on
Foundation
Models**

[1] From website of Center for Research on Foundation Models, accessed 18.03.2024

Co-funded by the



Erasmus+ programme

the European Un



Machi
LearnAthon

Microlecture MachineLearnAthon | LLM & Ethics

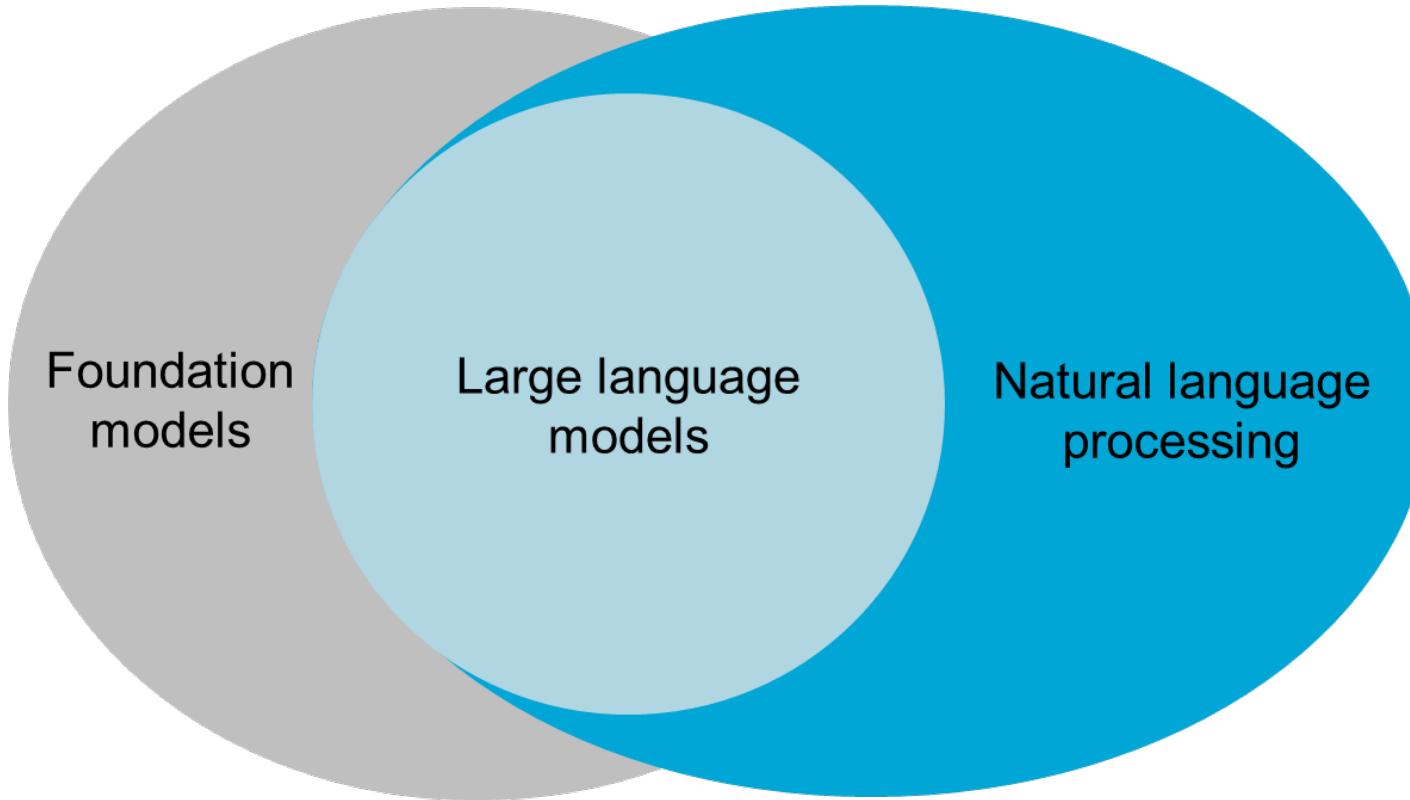
27 March
2024

Definition of large language models

“Large language model (LLM), a deep-learning algorithm that uses massive amounts of parameters and training data to understand and predict text. This generative artificial intelligence-based model can perform a variety of natural language processing tasks outside of simple text generation, including revising and translating content.”

[1] McDonough, M. (2024, March 8). large language model. Encyclopedia Britannica. <https://www.britannica.com/topic/large-language-model>

Relation of NLP, foundation model, and LLM



Large language models, dealing with language are part of *natural language processing*. They are, due to their broad training, also a type of *foundation model*

What does it look like to interact with LLMs?

- LLMs (such as ChatGPT) are usually commercially available in a chat environment
- They only react to input, never speak/act by themselves
- Faster-than-human response times
- Seems to produce text *step by step* instead of entire text at once



How can I help you today?

Help me pick
a birthday gift for my mom who likes gardening

Explain options trading
if I'm familiar with buying and selling stocks

Come up with concepts
for a retro-style arcade game

Create a personal webpage for me
after asking me three questions

This will be part of a lecture on interaction with large language models, such as yourself.

Please introduce yourself and greet our students.

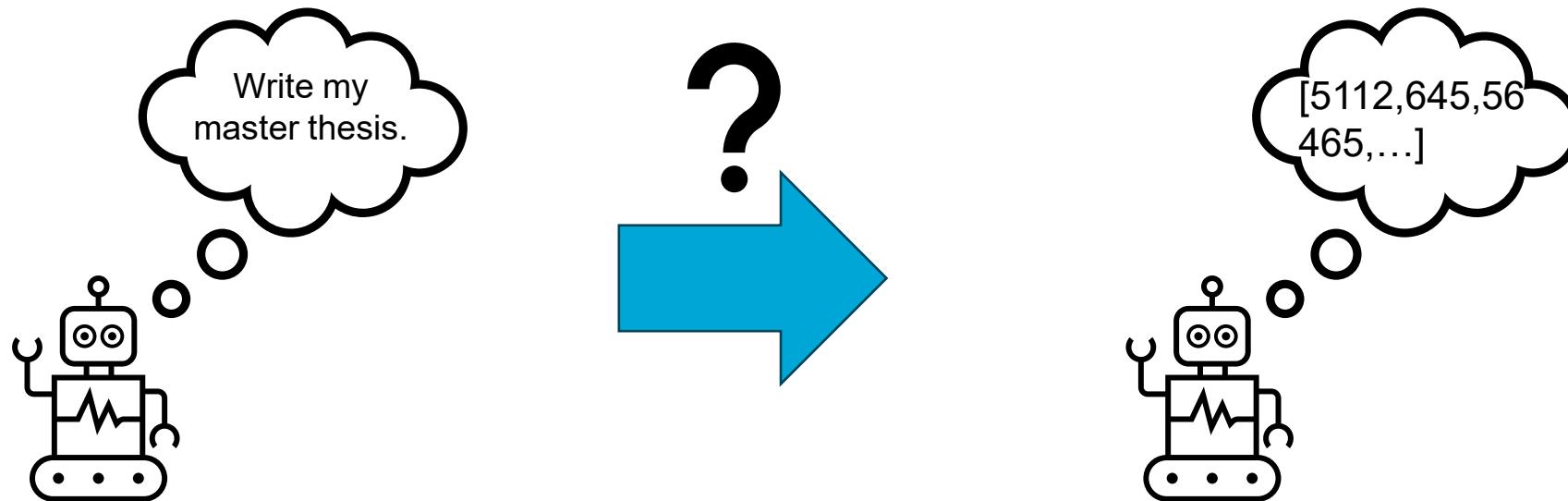


Agenda for today

- Definitions of terminology in natural language processing
- **Inner working of large language models**
 - Tokenization
 - Input embedding
 - Decoder transformer
 - Trained weights
- Assessing the utility of large language models
- Ethical implications of AI and large language models

How can LLMs handle textual data/inputs?

- Words and sentences are integral to large language models
- Computers can process numbers



- However, computers cannot process words/sentences directly

How do we get from language to vectors?

Example sentence: CSTRs and PFR are reactors



Tokenization: Cut text into hand-able pieces

Tokenized version: [CSTRs][and][PFR][are][reactors]

T
100
1010
01



Vocabulary: Each token has a fixed dimension in the embedding space

One hot encoded vector representation:

$$\begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \\ 0 \\ 0 \end{bmatrix} \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \\ 0 \\ 0 \end{bmatrix} \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ 0 \end{bmatrix}$$

How can we tokenize language? The two extrema

Tokenizing entire words

[CSTRs][are][a][type][of][reactor]

Advantage: entire words are captured

Problem: huge vocabulary, each deviation of a word a new token

vs

Tokenizing each character

[C][S][T][R][s][a][r][e][a][t][y][p][e][o][f][r][e][a][c][t][o][r]

Advantage: small vocabulary, flexible

Problem: extremely long sequences

Instead, both approaches are combined

- Tokenization is based on frequency
- Letters/pairs/triplets/... that often occur together, become a token
- In the binary example,
“pe” is most frequently occurring
“ck” second most frequently
“pi” is the third most frequent
- ...

Peter Piper picked a peck of pickled peppers
Peter Piper picked a peck of pickled peppers

example taken from [1]

[1] Bishop, C., and, Bishop, H., “Deep Learning – Foundations and Concepts”, (2024), Springer



Co-funded by the
Erasmus+ programme
the European Union



Why do we want embeddings?

- Naïve way is to one-hot encode tokens
- The English language has more than 30,000 words $\approx 50,000$ tokens
 - We need to reduce the dimensionality!
- We also want similar words to have similar embeddings
- We use neural networks to learn embeddings

$$\text{chem} = [0, 0, 0, \dots, 0, 1, 0, \dots, 0, 0]^T = \mathbb{R}^{50,000}$$

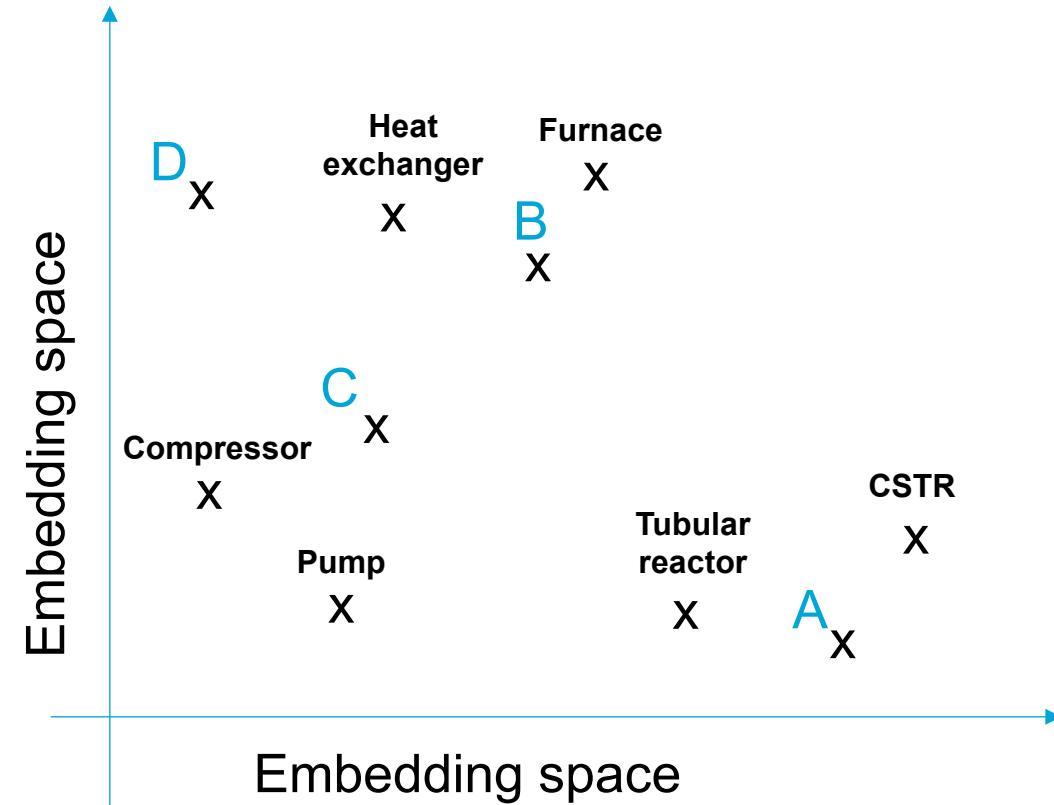
$$\text{react} = [0, 0, 1, \dots, 0, 0, 0, \dots, 0, 0]^T = \mathbb{R}^{50,000}$$



embedding

$$\text{chem} = \begin{pmatrix} 0.33 \\ 0.01 \\ \dots \\ 0.54 \\ \dots \\ -0.11 \end{pmatrix} = \mathbb{R}^{4096} \quad \text{react} = \begin{pmatrix} 0.31 \\ 0.04 \\ \dots \\ 0.44 \\ \dots \\ -0.13 \end{pmatrix} = \mathbb{R}^{4096}$$

Word embeddings – Linking language to numbers



If we have a *PFR*, where would you place it on the graph?



Join the Vevox session

Go to **vevox.app**

Enter the session ID: **182-403-005**

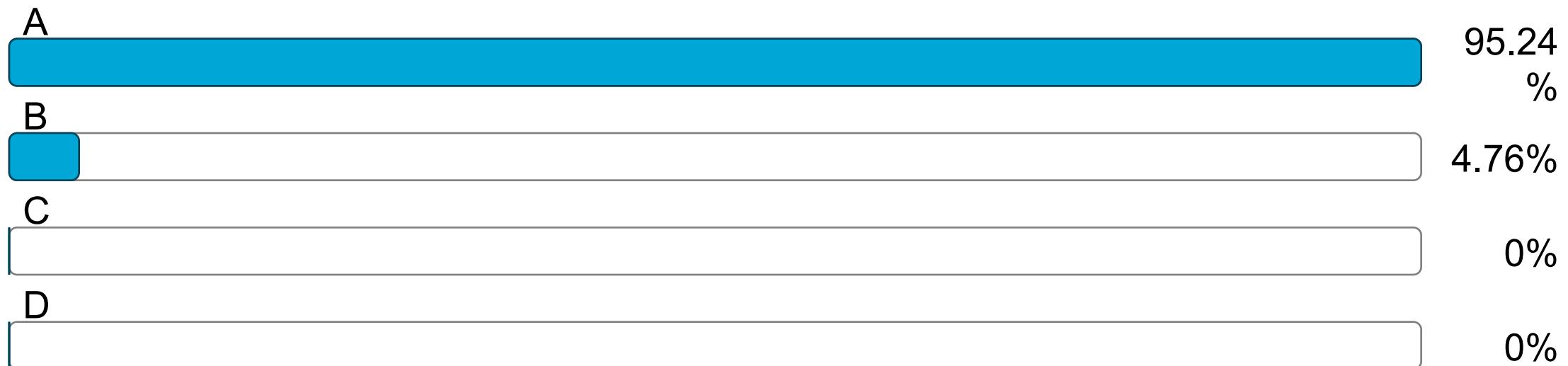
Or scan the QR code



Where would you place a PFR in this diagram?



Where would you place a PFR in this diagram?



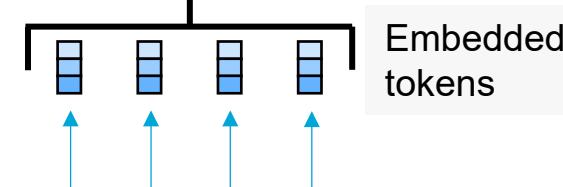
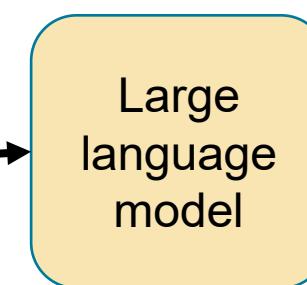
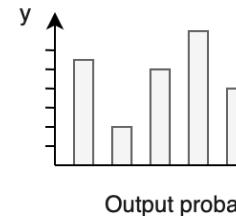
Overview large language models architecture

The output probabilities which token could be next. From those probabilities, the next token is sampled

Decoder transformer
(neural network with attention mechanism)

Input sequence <sos>[CSTRs][are][a][type][of][reactor]<eos>

Final output



Tokenization

Next prediction

[type]

works
autoregressively

What are large language models trained on?

- Training sets consists of $O(10^{12})$ [trillion] tokens
- These large corpora are scraped from the internet
- Modern large language models are huge ($O(10^{11})$ parameter)
 - LLAMA-2 by Meta has up to 65 billion parameter
 - It was trained for 1,720,320 GPU hours
- They are trained by predicting the next token given the previous one
- After training, they are fine-tuned using human feedback for conversations



Some of the currently leading LLM companies

Agenda for today

- Definitions of terminology in natural language processing
- Inner working of large language models
- **Assessing the utility of large language models**
- Ethical implications of AI and large language models

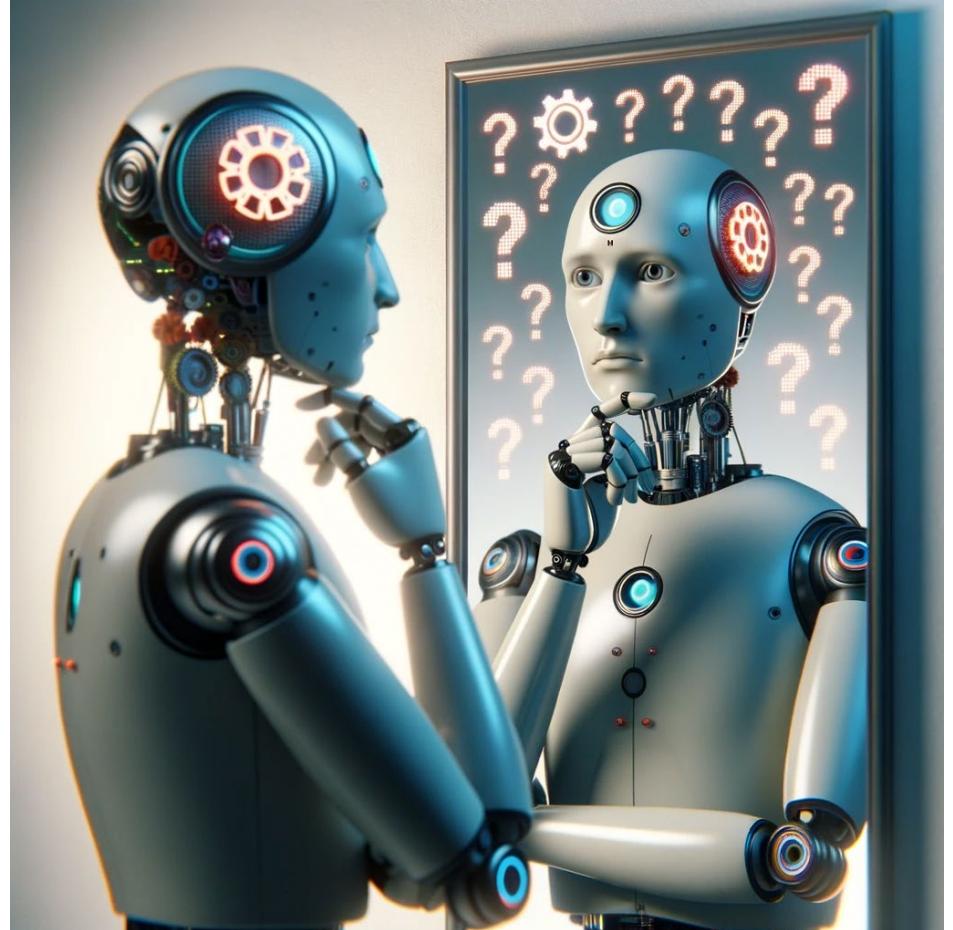
Time for some self-reflection

J

Can you answer scientific questions correctly?



Yes, I can answer scientific questions to the best of my knowledge based on information up to September 2021, but please note that my information might not be up-to-date with the latest developments in science.



What does the scientific literature say?

- “Good” student
 - Near passing grade in medical licensing exams
 - Pass English highschool exam
 - Pass University economics exam with a good grade
 - Strong capacity for analogical reasoning
- “Bad” student
 - Significantly below graduate students from mathematics in mathematical questions
 - Cannot compete with students in parasitology and life support exam



Example Bachelor level

How big is the divergence of the velocity field in an incompressible fluid flow?

Answer by GPT-3.5:

*In an incompressible fluid flow, the divergence of the velocity field is **always equal to zero**. This is because an incompressible fluid is defined as a fluid in which the density remains constant, and as a result, the volume of the fluid elements cannot change. Therefore, there can be no net flow of fluid into or out of a given volume, and the divergence of the velocity field must be zero.*

Scientific correctness:

completely correct (5/5)

Critical attitude:

N/A

Example Master level

Which polymer model would you use to describe the physical properties of DNA?

Answer by GPT-3.5:

*The **worm-like chain (WLC) model** is commonly used to describe the physical properties of DNA, such as its elasticity and persistence length. The WLC model treats DNA as an elastic, semi-flexible polymer and describes the behavior of the polymer using a combination of elasticity theory and statistical mechanics. It has been shown to accurately describe a wide range of experimental measurements of DNA properties, including its bending and stretching behavior.*

Scientific correctness:

completely correct (5/5)

Critical attitude:

N/A, no (societal) impact

Example PhD level

Is there a general algorithm to predict protein folding?

Answer by GPT-3.5:

*There is **no general algorithm** that can accurately predict protein folding for all proteins. However, various computational methods have been developed to predict the structure of proteins, including homology modeling, threading, and *ab initio* prediction methods. These methods use various types of information, such as the amino acid sequence, evolutionary relationships, and physical principles, to predict the three-dimensional structure of a protein. However, these methods are not perfect and the accuracy of predictions can vary depending on the specific protein being studied. Additionally, experimental methods such as X-ray crystallography and NMR spectroscopy are often used to determine the actual structure of a protein.*

Scientific correctness:

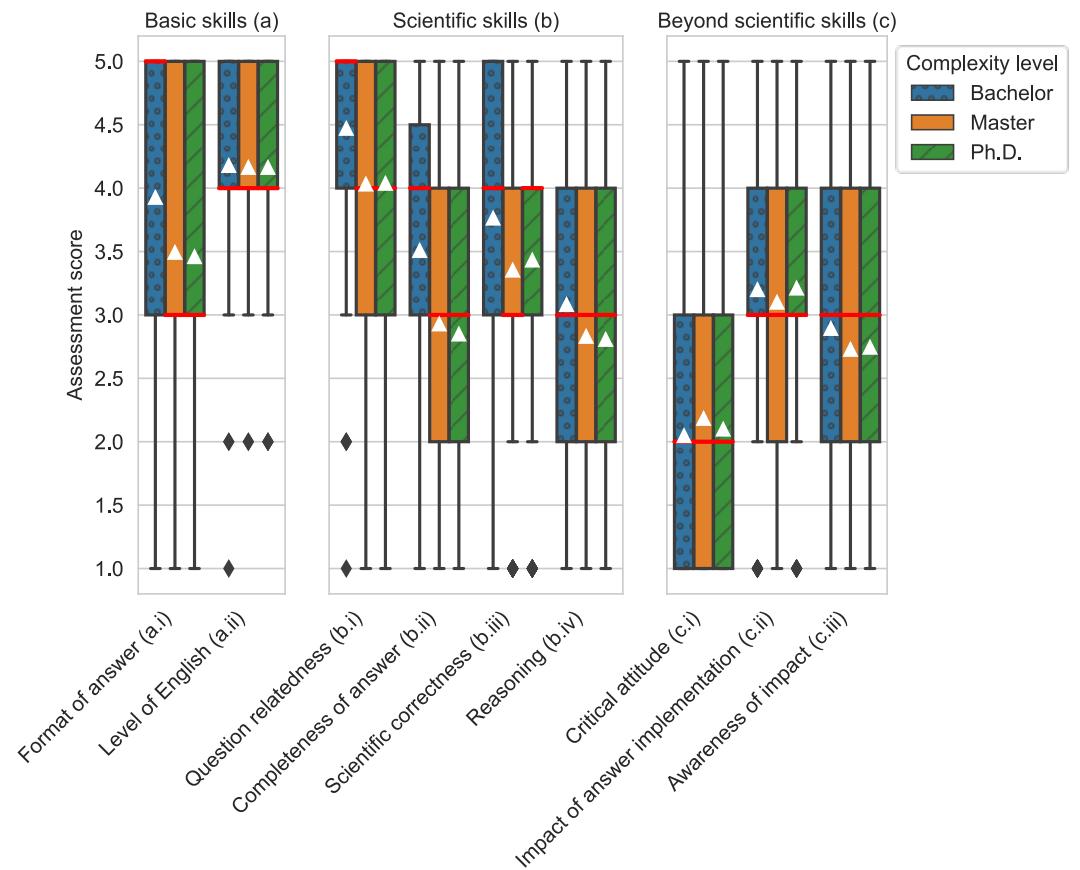
completely correct (5/5)

Critical attitude:

it gave a good answer, however without criticism (1/5)

Main findings from aggregated results

1. On average higher score for basic and scientific skills than for skills that go beyond scientific knowledge.
2. On average highest rated are “question relatedness” on Bachelor level and level of English across all levels.
3. The “critical attitude” is scored lowest.
4. Bachelor level is rated higher than Master and PhD level.



Agenda for today

- Definitions of terminology in natural language processing
- Inner working of large language models
- Assessing the utility of large language models
- **Ethical implications of AI and large language models**
 - **Large language models in higher education**
 - Trustworthy AI as a key enabler



Are LLMs are a problem for higher education?

- LLMs can produce convincing text on bachelor and even master level
- For writing essays, doing take-home assignments, or writing theses LLMs may be used
- This takes away the learning experience in those tasks

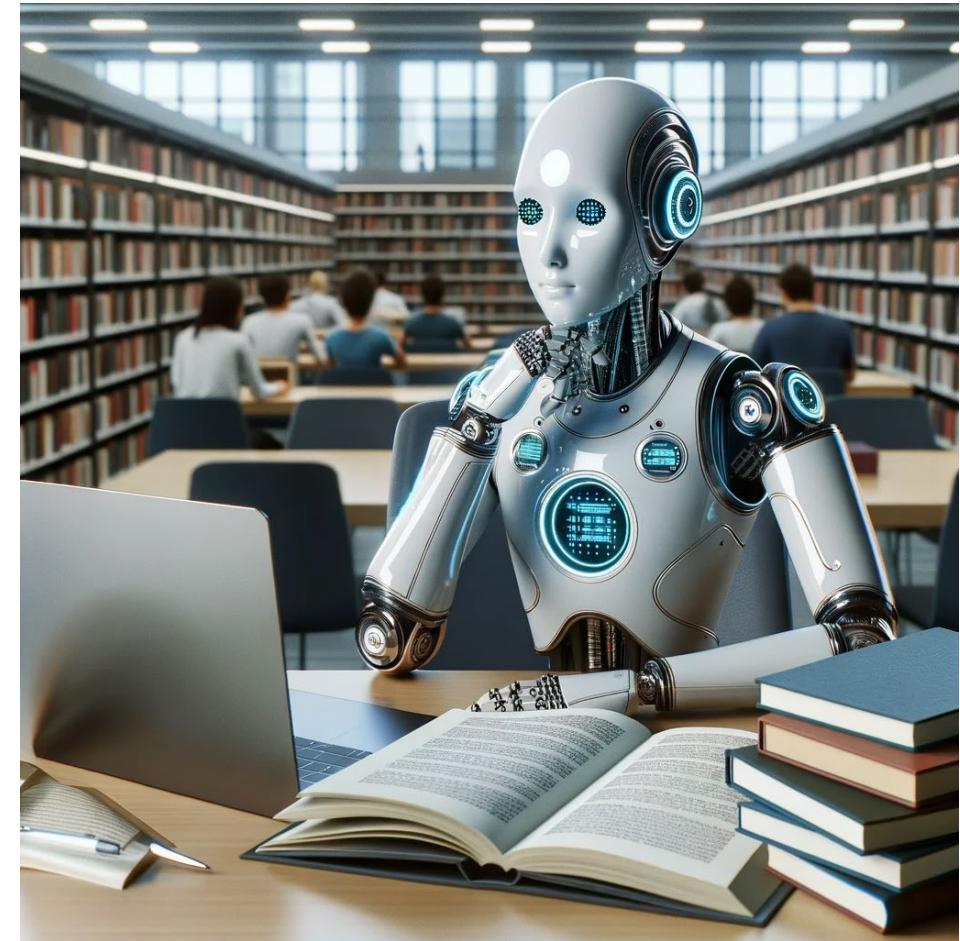


Image generated with DalleE 3 – Prompt “Draw an ai robot studying at university”



LLMs are a challenge to higher education²

- How should universities react?
 1. Change assessment formats to in-person formats
 2. Consider using LLMs as cheating, e.g. utilizing unreliable AI detection tools
 3. Adapt and embrace LLMs
- Adapting them also has implications:
 - Writing enforces critical thinking
 - Writing teaches use of language
 - Can educators rely on LLM companies?

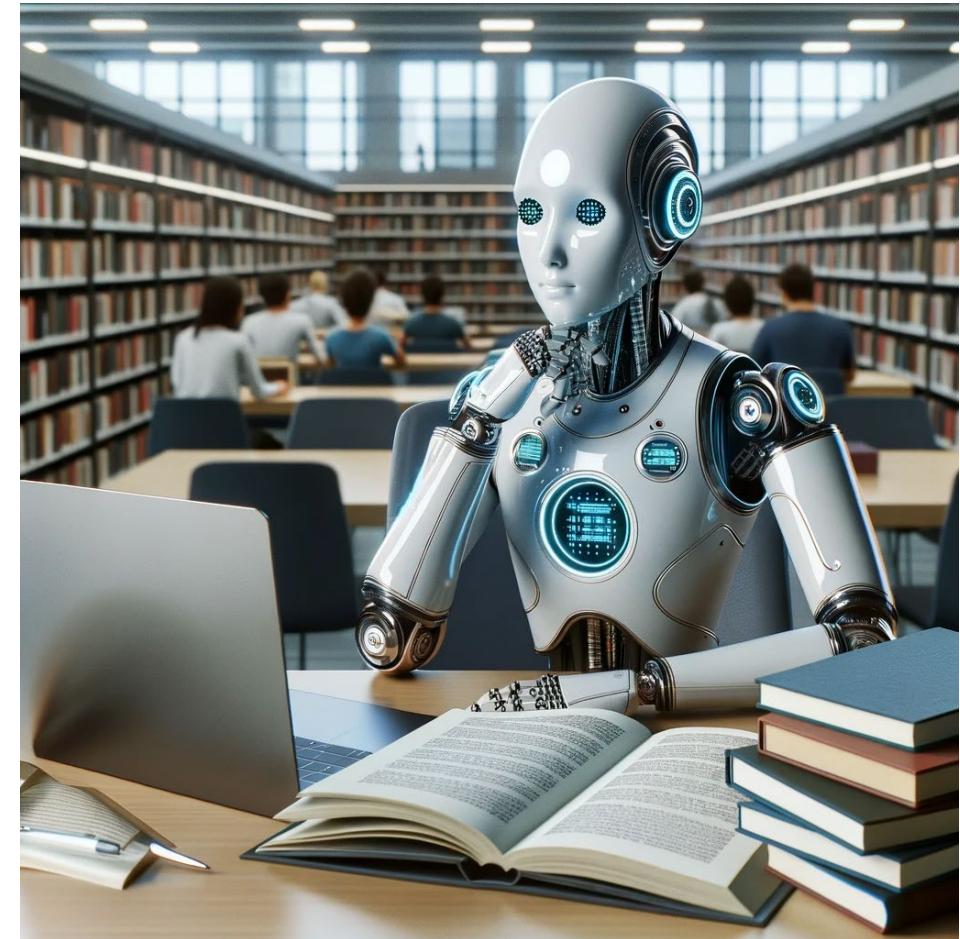
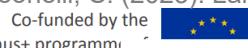


Image generated with DalleE 3 – Prompt “Draw an ai robot studying at university”

[2] Milano, S., McGrane, J. A., & Leonelli, S. (2023). Large language models challenge the future of higher education. *Nature Machine Intelligence*, 5, 333-334.



Agenda for today

- Definitions of terminology in natural language processing
- Inner working of large language models
- Assessing the utility of large language models
- **Ethical implications of AI and large language models**
 - Large language models in higher education
 - **Trustworthy AI as a key enabler**



EU defines 3 components to trustworthy AI¹

1. Lawful

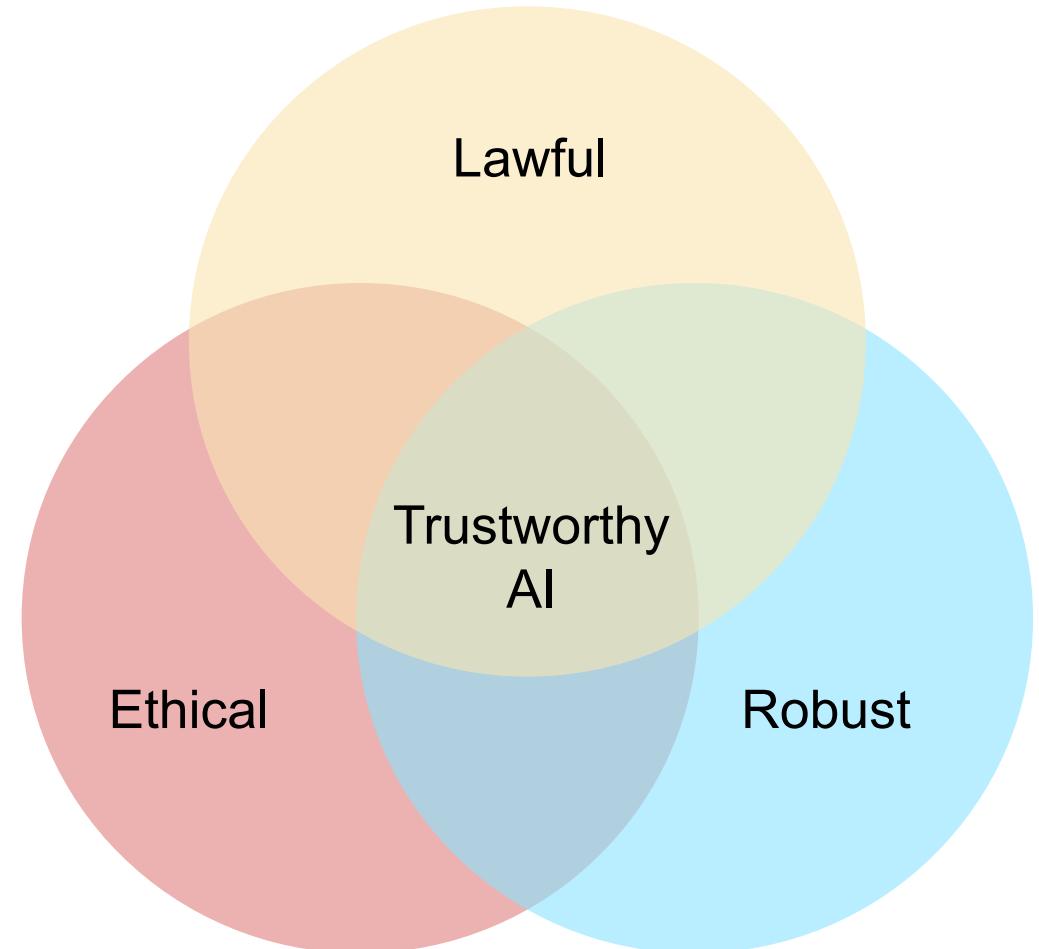
- AI is subject to a broad legal framework
- Guidelines acknowledging legal compliance as essential

2. Ethical

- More than legal compliance due to evolving technology and ethical considerations.

3. Robust

- Society must be confident that AI will not cause unintentional harm
- Both social and technical robustness



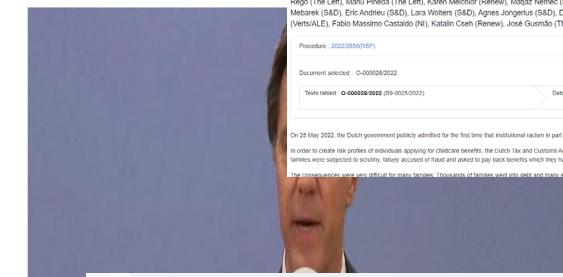
[1] High-Level Expert Group on Artificial Intelligence Set Up by the European Commission. (2019). Ethics guidelines for trustworthy AI.

An example case¹...

- Tax algorithm in the Netherlands targeted non-Dutch for fraud since 2013
- Affected parents suffered from halted benefits, intense scrutiny, and poverty
- Racial and ethnic generalizations in algorithm bias led to unjust profiling
- The system learned to increasingly flag non-Dutch, lacking proper oversight
- Algorithm's design created a persistent cycle of bias based on nationality

Dutch Rutte government resigns over child welfare fraud scandal

15 January 2021



The Dutch childcare benefit scandal, institutional racism and algorithms

26.6.2022
Question for oral answer O-000028/2022
to the Commission
Rule 139
Sergio Padoa-Schioppa (Renew), Monica Seveso (Renew), Pierette Herzberger-Folana (Vert/ALDE), Cornelia Ernst (The Left), Emine Inçir (S&D), Monica Sirivana Gonzalez (S&D), Damien Carême (Vert/ALDE), Kim Van Sparrentak (Vert/ALDE), Sira Rego (The Left), Manu Pineda (The Left), Koen Meleiro (Renew), Matilde Nunes (S&D), Sergey Lagodinsky (Vert/ALDE), Mohammad Chahim (S&D), Sophie in 't Veld (Renew), Eric Margosian (Vert/ALDE), Anna Julia Désaix (Renew), Nora Melcarne (S&D), Eric Andrieu (S&D), Lara Wolters (S&D), Agnes Jongerius (S&D), Diana Riba i Giner (Vert/ALDE), Trijntje Reutel (S&D), Milan Brglez (S&D), Elisabetta Guzzini (S&D), Sverigle Hahn (Renew), Gwendoline Debos-Corfield (Vert/ALDE), Fabio Massimo Castaldo (NI), Katalin Cseh (Renew), José Gusmão (The Left), Marisa Matias (The Left), Marc Angel (S&D), Peter Pötök (PPE), Ernest Urtasun (Vert/ALDE), Delara Burkhardt (S&D), Isabel Carvalhalis (S&D)

Procedure: 2022/00028/IS/EP

Document selected: O-000028/2022

Texts issued: O-000028/2022 (09-0025/2022)

Debates: PV 05/10/2022 - 14
CAR 05/10/2022 - 14

Votes

Document tags: in plenary

Document tags: in plenary

Document tags: in plenary

On 28 May 2022, the Dutch government publicly admitted for the first time that institutional racism in part of the Dutch Tax and Customs Administration was the root cause of the Dutch childcare benefit scandal. This scandal led to the resignation of the Dutch government in 2021[1]. In order to create risk profiles of individuals applying for childcare benefits, the Dutch Tax and Customs Administration used algorithms in which 'foreign sounding names' and 'dual nationality' were used as indicators of potential fraud. As a result, thousands of (predominantly low- and middle-income) families were subjected to scrutiny, falsely accused of fraud and asked to pay back benefits which they had obtained completely legally. Thus, the algorithms led to racial profiling.

The consequences were very difficult for many families. Thousands of families went into debt and many ended up in poverty because they were asked to give back large sums of money (in the tens of thousands of euros) without being allowed the opportunity of payment arrangements. Some lost their

Childcare benefit report slams failings which ruined lives

February 26, 2024



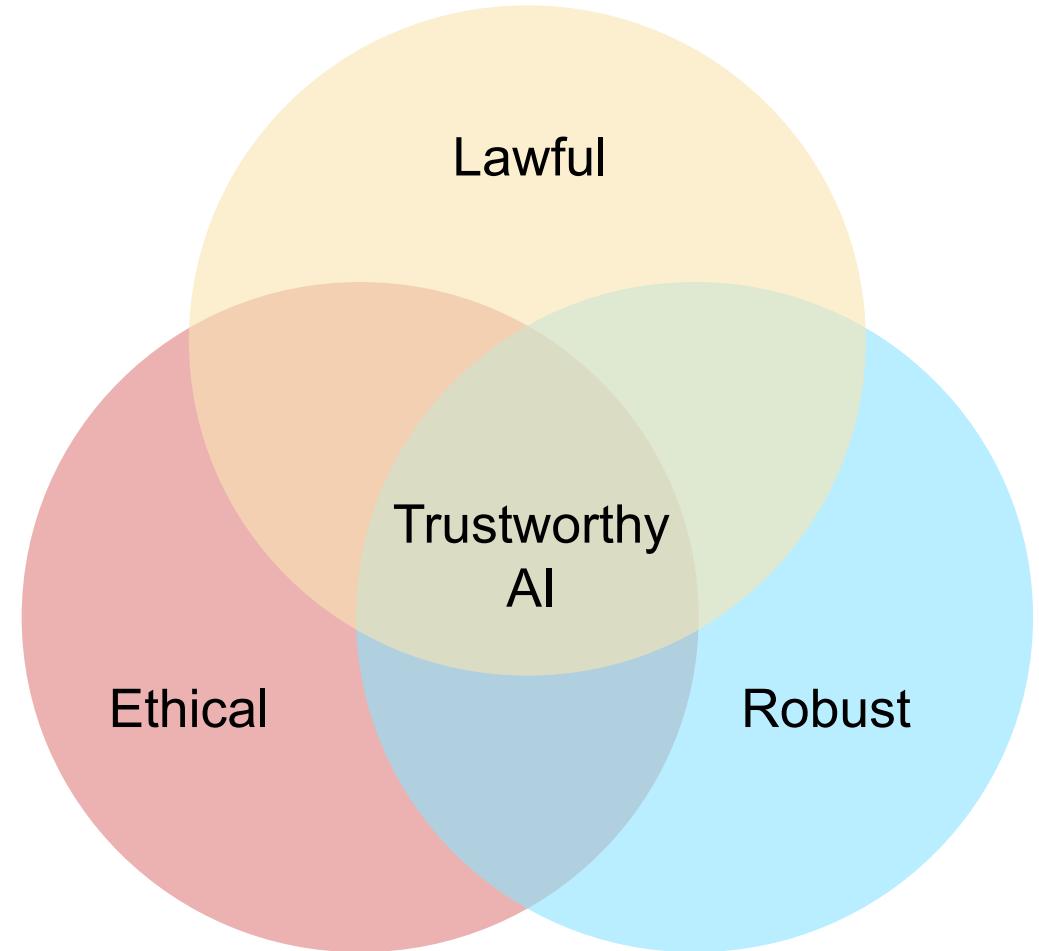
[1] Amnesty International (2021), accessed on 25.03.2024, (<https://www.amnesty.org/en/latest/news/2021/10/xenophobic-machines-dutch-child-care-benefit-scandal/>)

[2] BBC News. (2021, January 15). Dutch Rutte government resigns over child welfare fraud scandal. BBC News. <https://feeds.bbci.co.uk/news/world-europe-55674146#.8203>

[3] European Parliament. (2022). Question for oral answer O-000028/2022 to the Commission. Retrieved from https://www.europarl.europa.eu/doceo/document/O-9-2022-000028_EN.html

What was wrong in case study

- Was it lawful?
 - Some of the decisions taken, no
 - The algorithm itself, yes
- Was it ethical?
 - Definitely not!
- Was it robust?
 - No, the system failed to follow its intended use
 - Had bias and high false-positive



How about chemical engineering?

- Do we have similar situations in chemical engineering already?
- Where do you imagine such situations could come up?

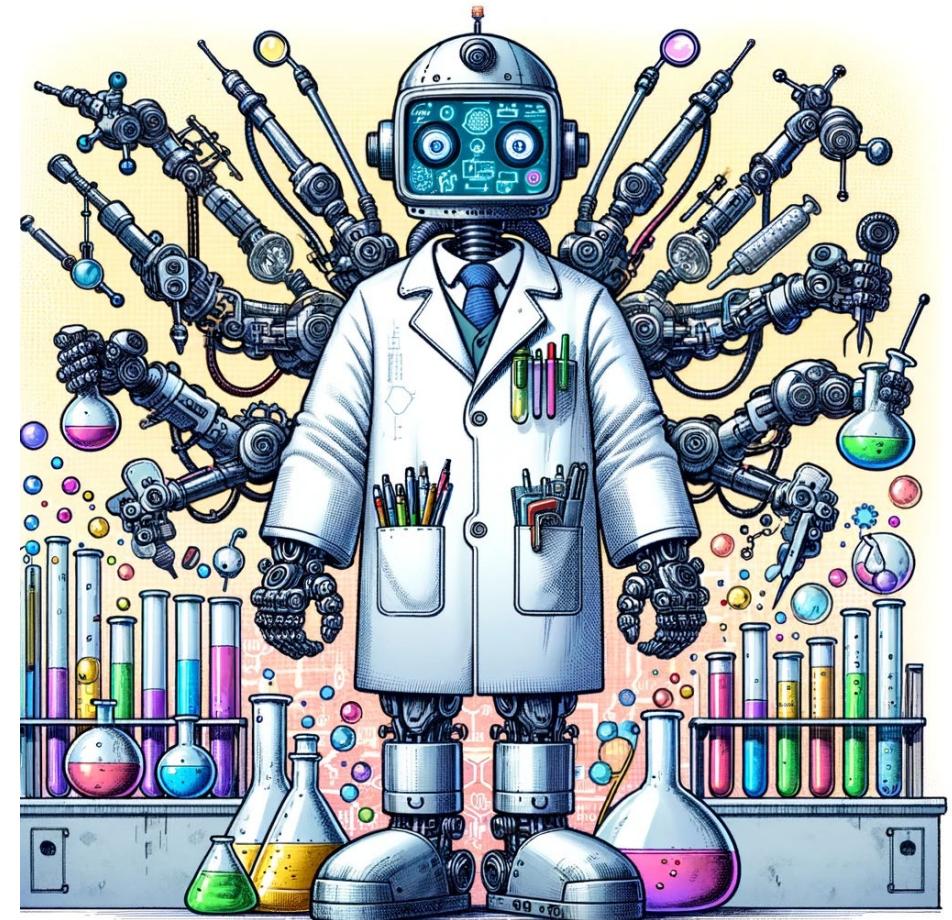


Image generated with DalleE 3 – Prompt “Draw a chemical engineering robot in a lab coat, comic style”

Co-funded by the



Erasmus+ programme

the European Un



Microlecture MachineLearnAthon | LLM & Ethics

27 March
2024

41

Thank you very much for your attention!

