Wat weet jij over wat er bij een groot taalmodel onder de motorkap gebeurt?

De presentatie start om 10:00

Te vroeg? Doe de ‘zelf’test! (5 min)
Vraag daarna naar de antwoorden

Joint Support Command
Ministry of Defence

Large Language Models

What happens “under the hood”?

Dr. Ir. Henrike Krijnsen

DOSCO
NLDA bibliotheek

28 February 2024

Before we start...

Urgent questions can be asked anytime

Otherwise, please wait until a subsection ends

Please turn off any sound on any device

Source: Pixabay

CC BY NC

Outline

A. General introduction to AI	75 min
B. Basics of ANNs	60 min
C. From ANNs towards LLMs	60 min
D. Further reading	n.a.

CC BY NC

A. General introduction to AI

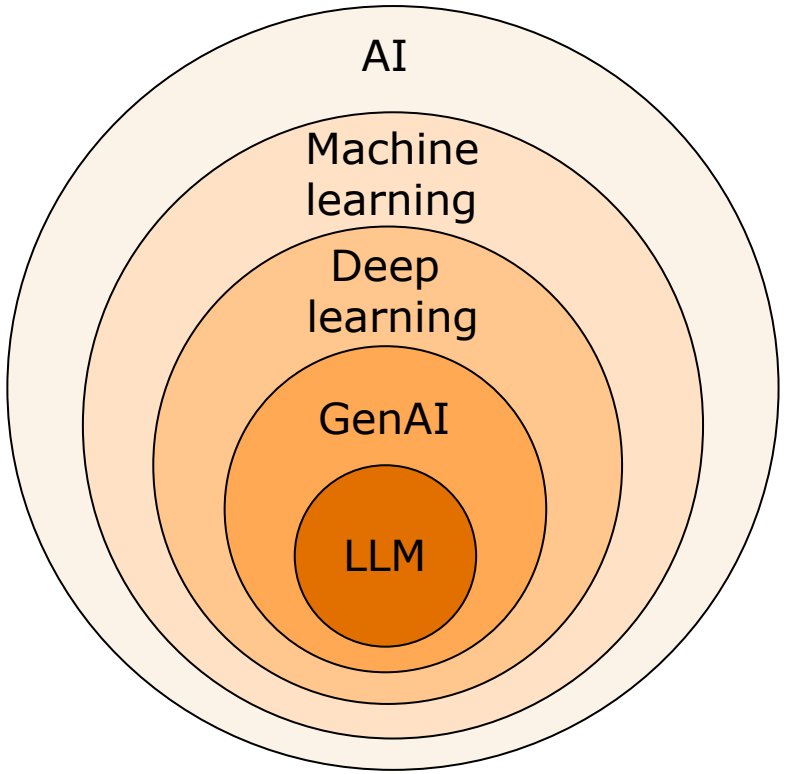
1. Moving from AI to ML, DL, GenAI and finally to LLMs

   AI - Why NOW?

   Plenary session 1
2. Search engine versus reasoning engine
3. Plenary session 2

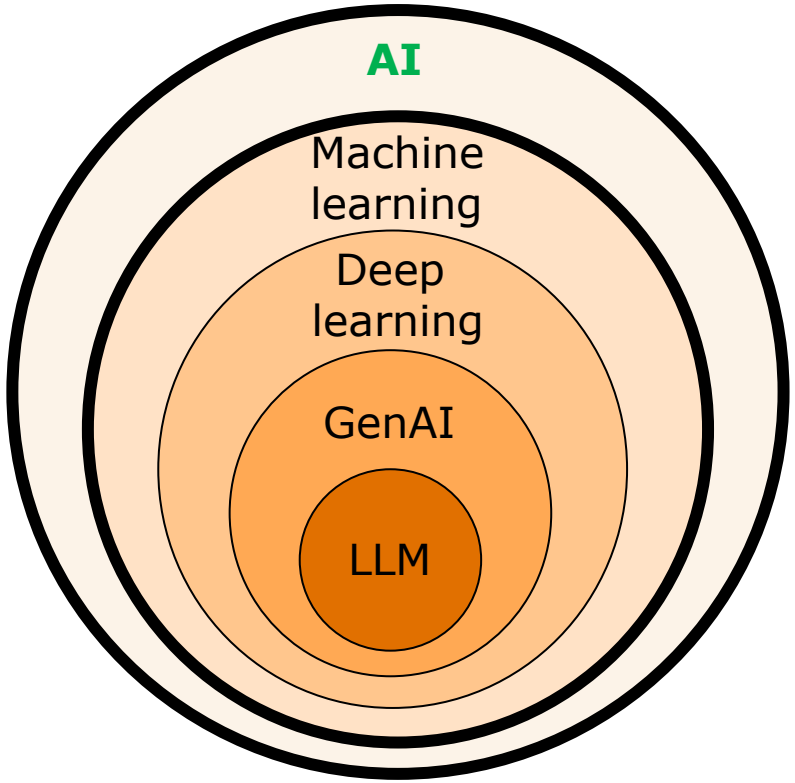
Source: Pixabay

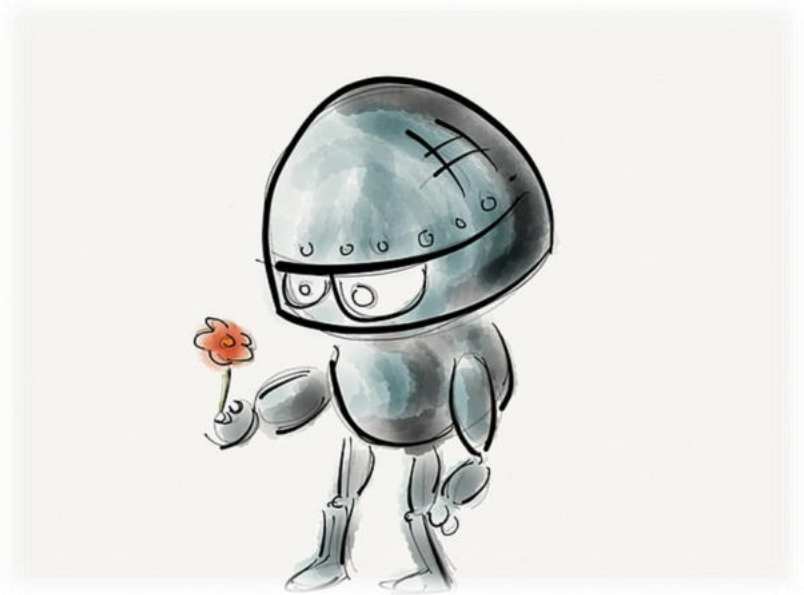
(A1) Moving from AI to ML, DL, GenAI and finally to LLMs



Human programmers **wrote a set of instructions (an algorithm)** that reviews large volumes of existing data to define the model itself. As such, the human programmers do **not** build the model, they programmed the algorithm that builds the model

CC BY NC





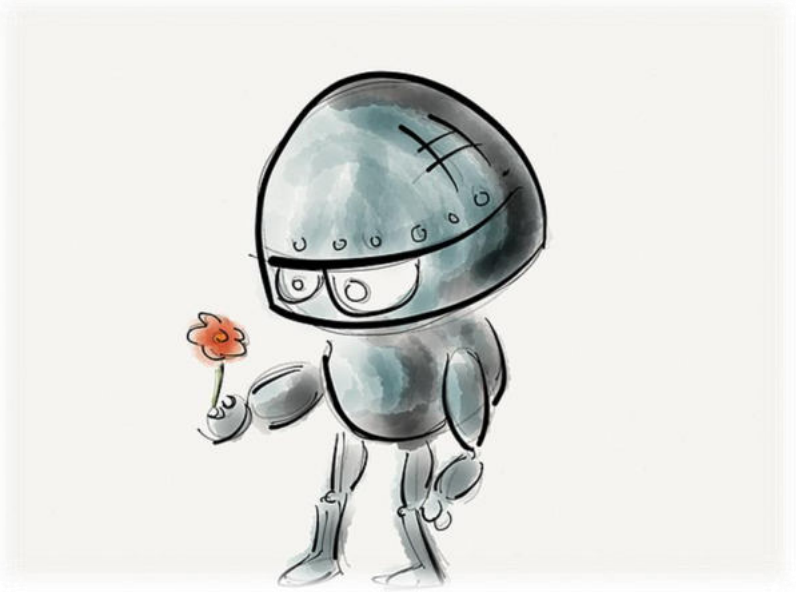
Source: Pixabay

CC BY NC

Weak versus strong AI

Expert systems: good at ① task only (1970's)
based on symbolic systems = matching patterns and symbols

Weak AI



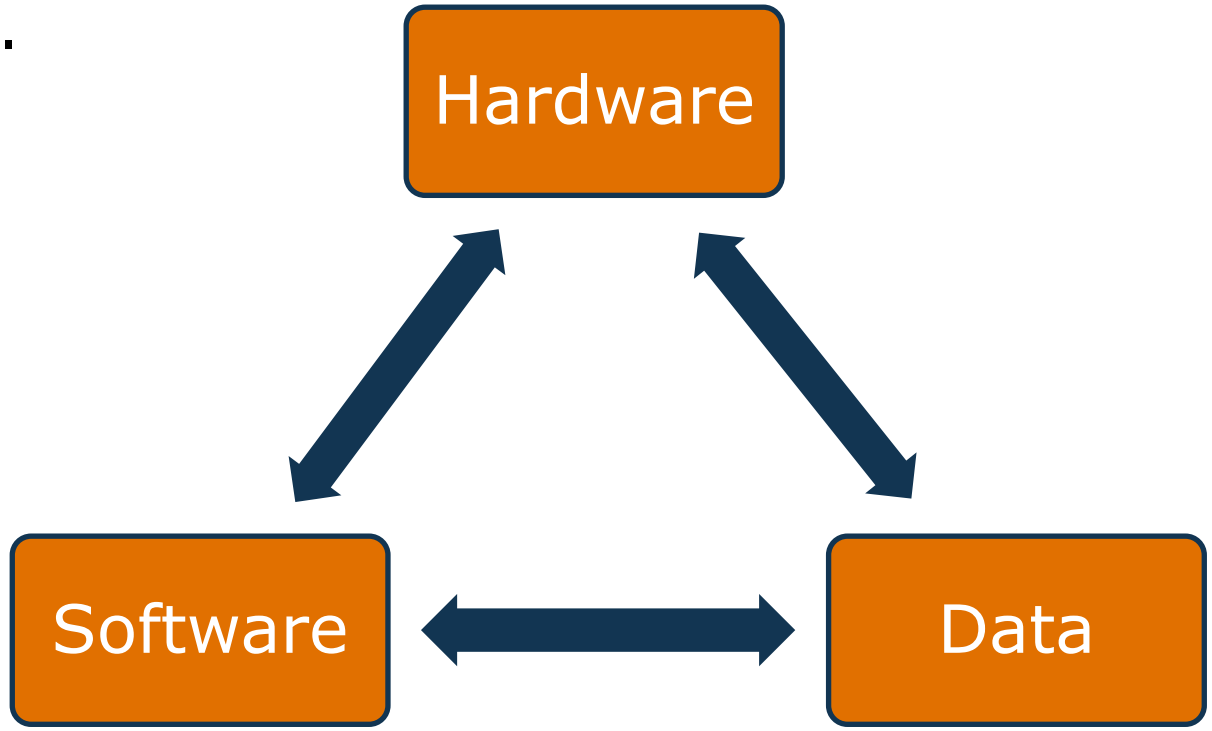
Strong AI

Source: Pixabay

CC BY NC

(A2) AI – Why now?

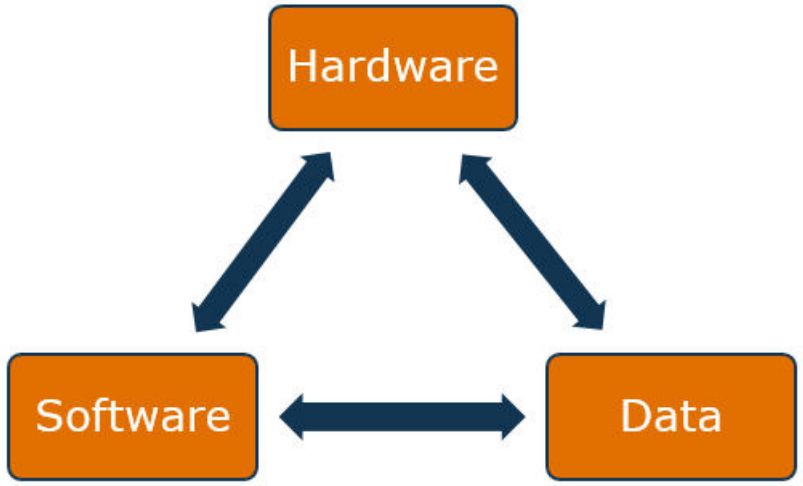
Availability of ...



CC BY NC

AI – Why now? **Hardware**

- Fast, powerful, cheap CPUs
- Fast, powerful, cheap GPUs
- Cheap storage capabilities
- Available online High Performing Computing power



CC BY NC

AI – Why now? **Software**

- Plenty of tools and algorithms enabling machine learning
- Transformer model (2017)
- Attention head algorithm (2017)

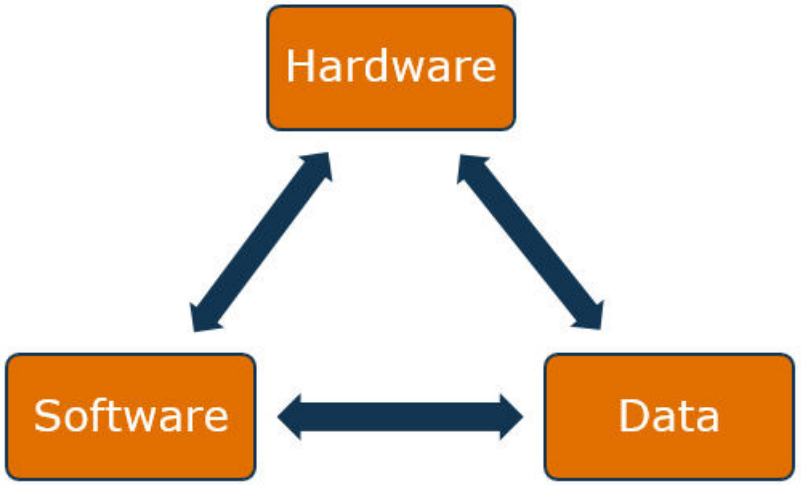
Still, this does not mean the general public has access to all of these tools

- Closed versus **Open** AI tools and software
  AI dataset / model / application – rights differ

CC BY NC

AI – Why now? **Data**

- Enormous amount of texts and data
- Internet of Things



CC BY NC

AI types and their current applications

AI types	Applications
• Reactive machines	autonomous cars
• Limited memory	weather forecast
• Theory of mind	customer assistance through chatbots
• Narrow AI	customer suggestions
• Supervised training	identification of objects, generation
• Unsupervised training	fraude detection
• Reinforcement training	playing games

Source: Pixabay

CC BY NC

Supervised training and NLP

“Ba”	“Ball”
“Baw”	“Ball”
“Bow”	“Ball”
“foBow”	“Football”
“foeBowl”	“Football”
“Footbowl”	“Football”
“Football”	“Football”

CC BY NC

Source: Pixabay

(Un)Supervised training

Supervised training – there is always some kind of tutor, providing feedback

Unsupervised training takes place through observations

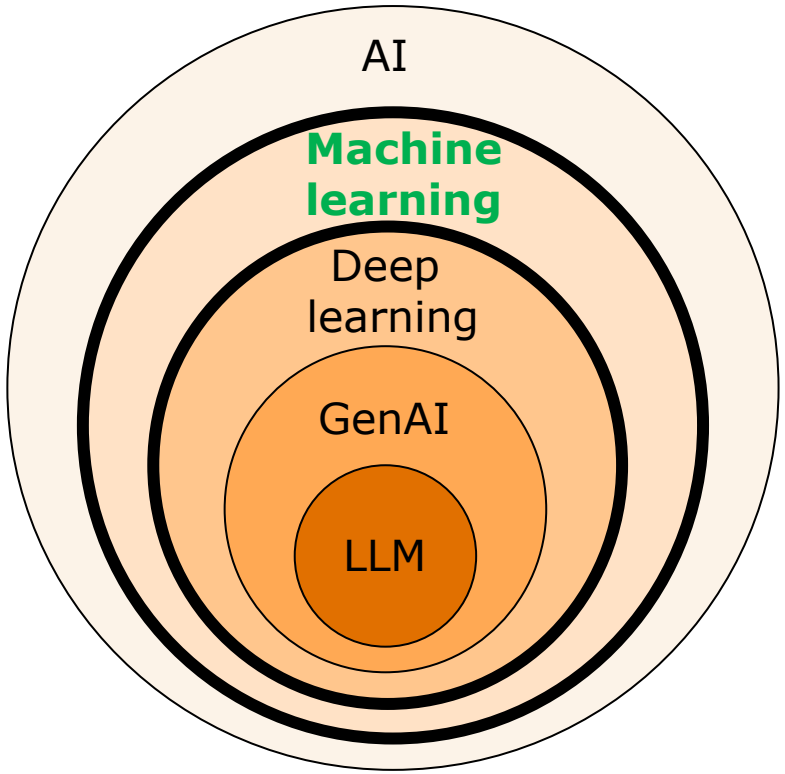
CC BY NC

Fun fact: AI also tries to learn from toddlers

Research by NY University
n=1

CC BY NC

Machine learning (ML)



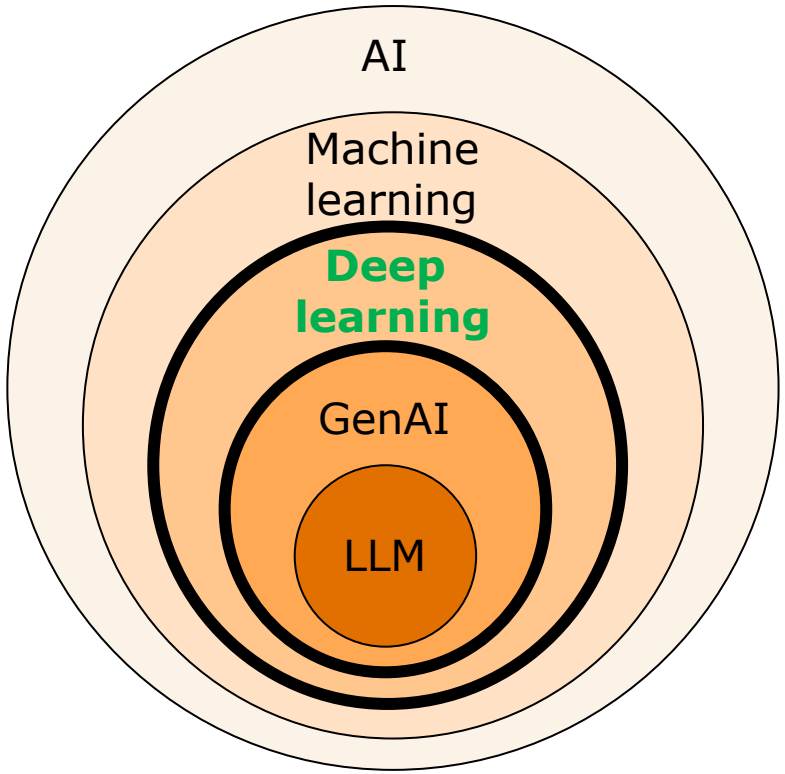
1959 - First application: checkers game

1950s - Too little digital info on other domains

1990s - Introduction of the internet
Availability of loads of digital training data, such as dog images
Creation of computer systems designed to mimic our brain

CC BY NC

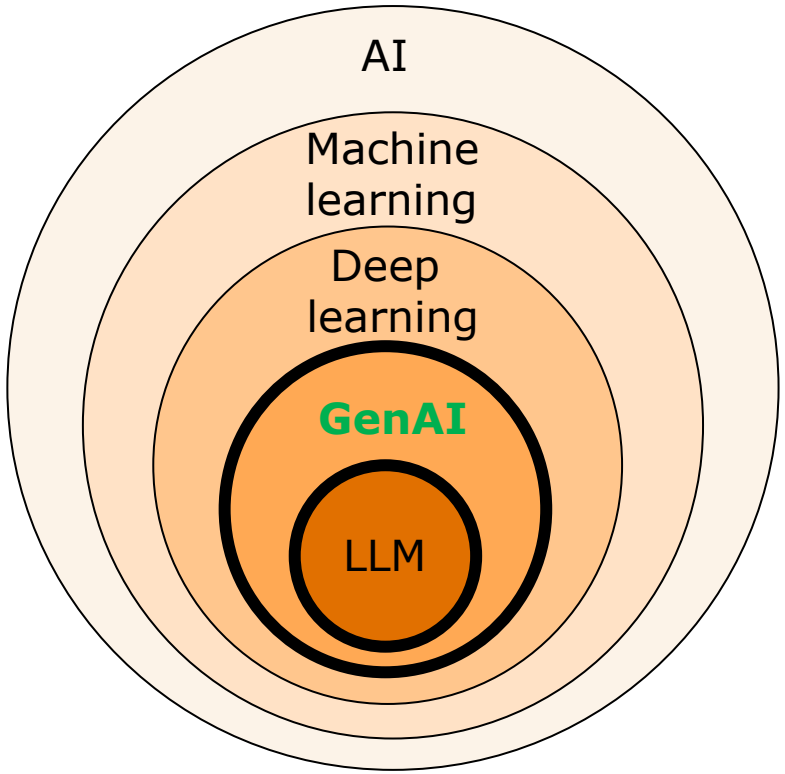
Deep learning (DL)



2000	words were turned into features, and patterns could be recognized.
2000s	**deep** learning was introduced
2013	introduction of the first vectors to better learn languages

CC BY NC

GenAI



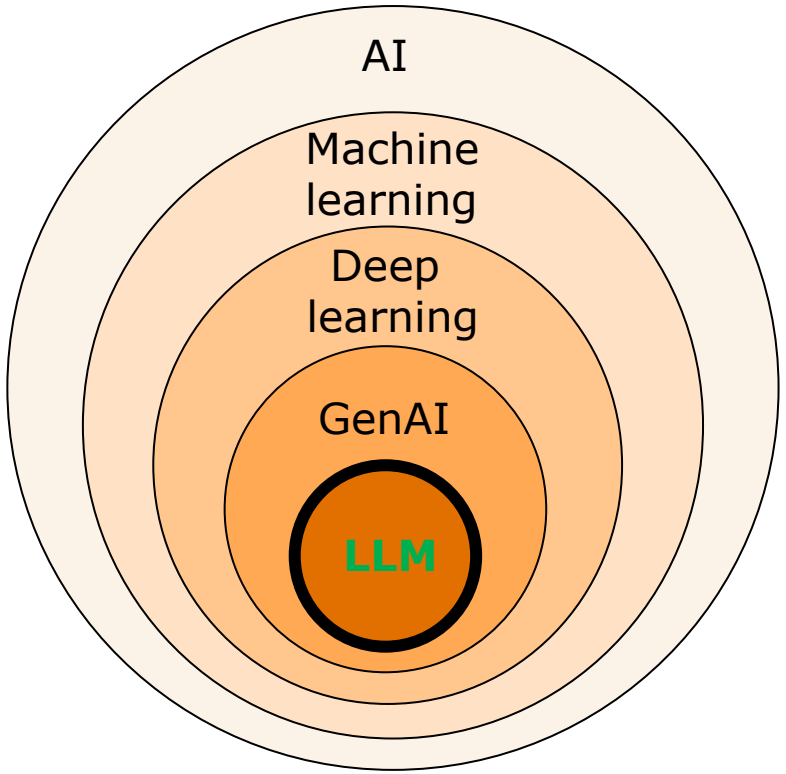
Generative AI: you *make* something.

Generative AI versus Discriminative AI

Definition of GenAI:
An AI system with a degree of **autonomy** aiming to create content, such as:
text, code, images, video, music, audio, translations

CC BY NC

LLM



Deep learning algorithm performing NLP tasks

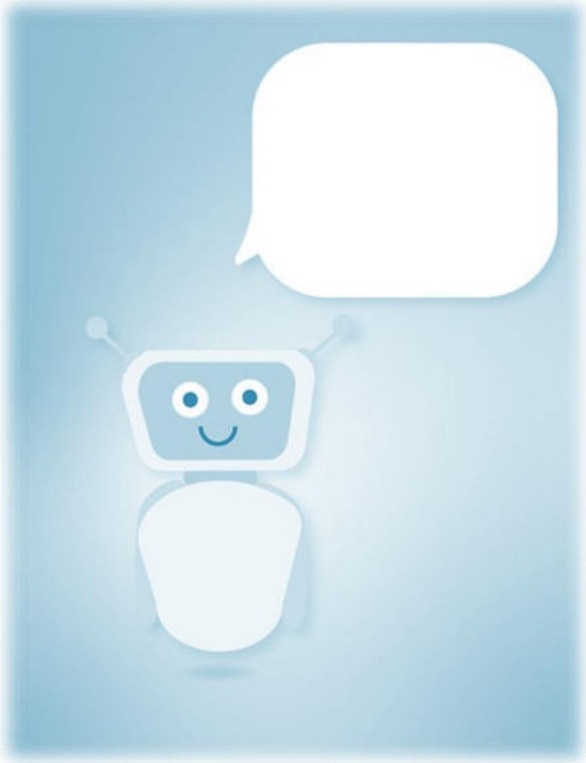
'Large' comes from massive data sets used for training

Thus, it enables recognition, translation, prediction and generation of text or other content

CC BY NC

Applications of (small) language models

- Phone Keyboard
- Email Predictions
- Search Engines
- Chatbots
- Automatic Translation
- Speech Recognition



Source: Pixabay

Based on conversation with CoPilot, 2024-01-30

CC BY NC

Large language models, ...

- Generate fluent and coherent texts on various subjects and styles.
- Answer complex questions that require more than one fact or source.
- Perform multiple tasks simultaneously.
- Learn new knowledge and skills from unstructured data such as books, websites, podcasts, etc. without the need for instructions.
- Mimic patterns. Mimicking allows reusing ‘stored’ patterns or tags.

However, their scope is limited to what has been trained.

CC BY NC

CC BY NC

Personal image by H. Krijnsen, 2023-12-12

LLM (Plenary session 1)

LLMs work by predicting the next word

I have a sad story to tell you
It may hurt your feelings a bit
Last night when I walked into my bathroom
I stepped in a big pile of sh...

Source: Pixabay

CC BY NC

LLM

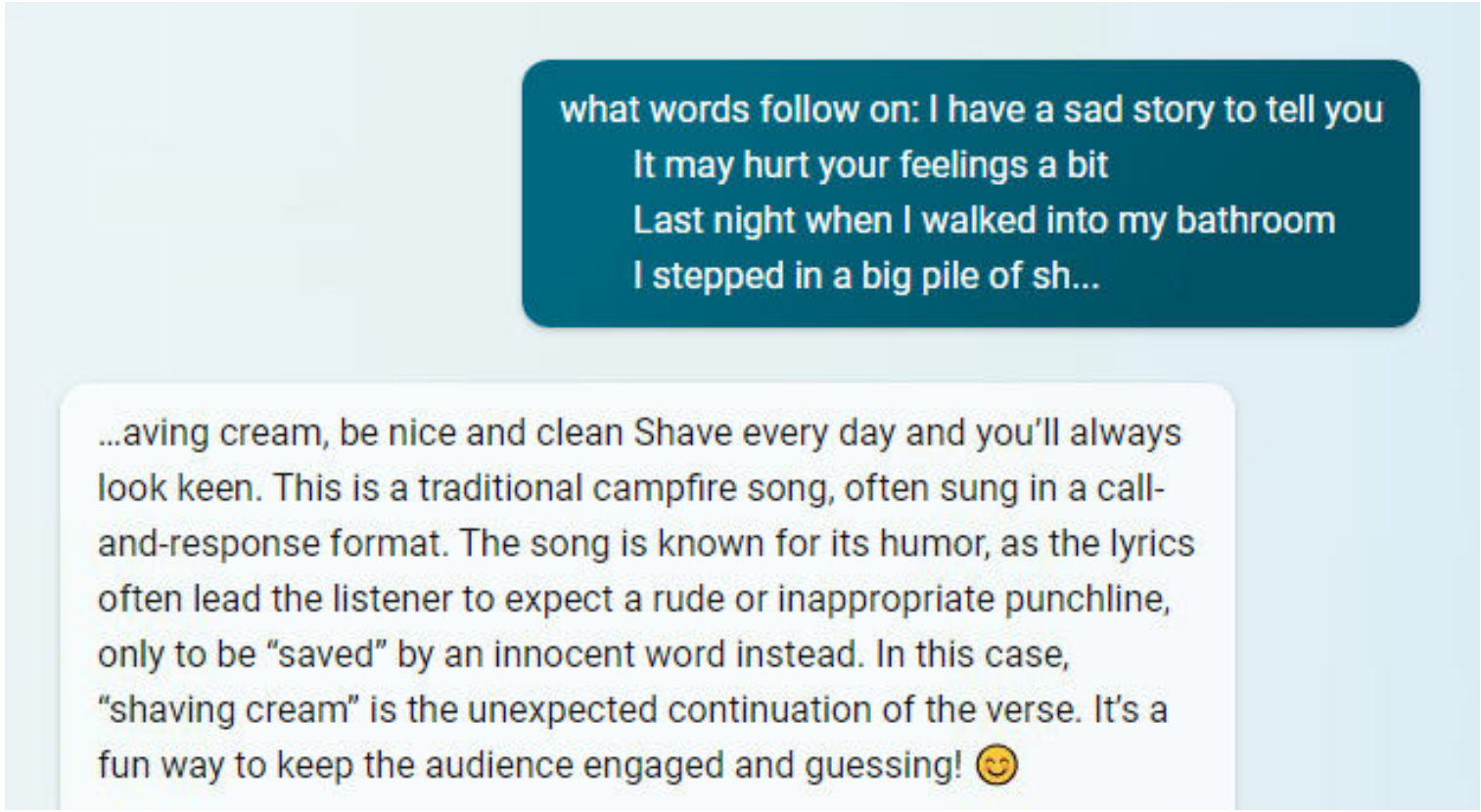
LLMs work by predicting the next word

I have a sad story to tell you
It may hurt your feelings a bit
Last night when I walked into my bathroom
I stepped in a big pile of s

CC BY NC

Source: Pixabay

LLM – Let’s put it to the test



Screenshot Copilot, generated 2024-01-22

CC BY NC

Large Language Model (LLM)

2017 – The transformer was invented

Attention heads were introduced – not just reading words from left to right but emphasis (“attention”) on specific words

Even words from previous sentences may be contributing

CC BY NC

(A2) Search engine versus reasoning engine

Traditionally, we are familiar with search engines, like Google, Bing

Since the introduction of GenAI we now also sometimes use the term ‘reasoning engines’

These ‘reasoning engines’
- do not just focus on the exact query
- provide more information, in more fluent language
- are able to combine information

CC BY NC

Search engine

Traditional search engine works by
1. Crawling (crawlers or spiders)
2. Indexing
3. Ranking (relevancy, authority, …)

As a result, traditional search engines:
- Are easy to use, free
- Provide comprehensive access
- Provide relevant results as a list, the user can make a selection
- Save time (reduce noise, fast access to information)
- Have advanced search options

CC BY NC

Source: Pixabay

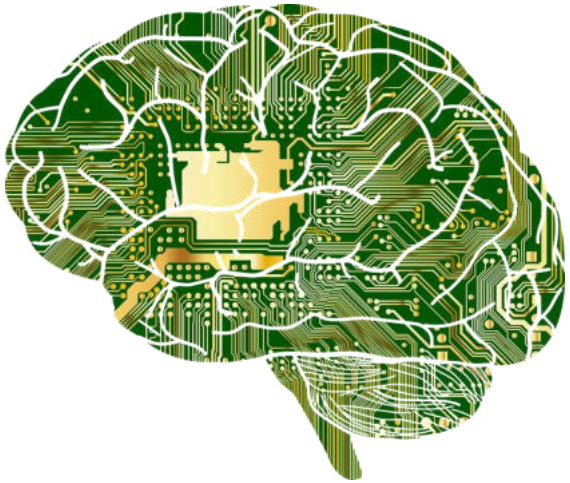
Reasoning engine (based on AI)

It is performing some kind of logical or rational ‘thinking’:
It is able to make summaries, draw conclusions, combine information, solve problems

It is ‘aware’ of the context

It provides a response using human-like speech:
the response is more conversational and more elaborate

Basically, there is one response



CC BY NC

Search engine versus reasoning engine

1. Reasoning engines are
- designed to interpret human language
- providing direct relevant responses
- maintaining the context and understand the intent of your question
- accessing other data sources like websites, databases for recent information

2. Search engines are
- good resources when you aim to do further exploration, but are
- not optimized for deeper questions (phrases),
- not truly understanding a query, they just match your keywords

CC BY NC

Reasoning engines versus search engines

Quite often you are not looking for a very factual answer, and aim to be more widely oriented, and the reasoning engine may be more helpful.

Validation: always verify and validate the results, check the sources.

Reasoning engines and search engine serve various purposes.

Already, the two types of engines are sometimes combined

CC BY NC

Main take-away messages (part A)

- AI is the ability of machines to mimic human intelligence
- ML, DL, genAI and LLMs are all techniques that allow machines to learn from data
- LLMs require the current availability of Hardware, Software and Data
  - The introduction of the transformer and attention head in 2017 are crucial for the current quality of LLMs
- Traditional search engines and Reasoning engines serve different purposes
- Reasoning engines are based on LLMs and are able to combine various tasks

CC BY NC

(A3) Task delegation (Plenary session 2)

What information specialists' tasks would you 'delegate' to AI?
For what reason?

10 min

CC BY NC

CC BY NC