Generatieve AI

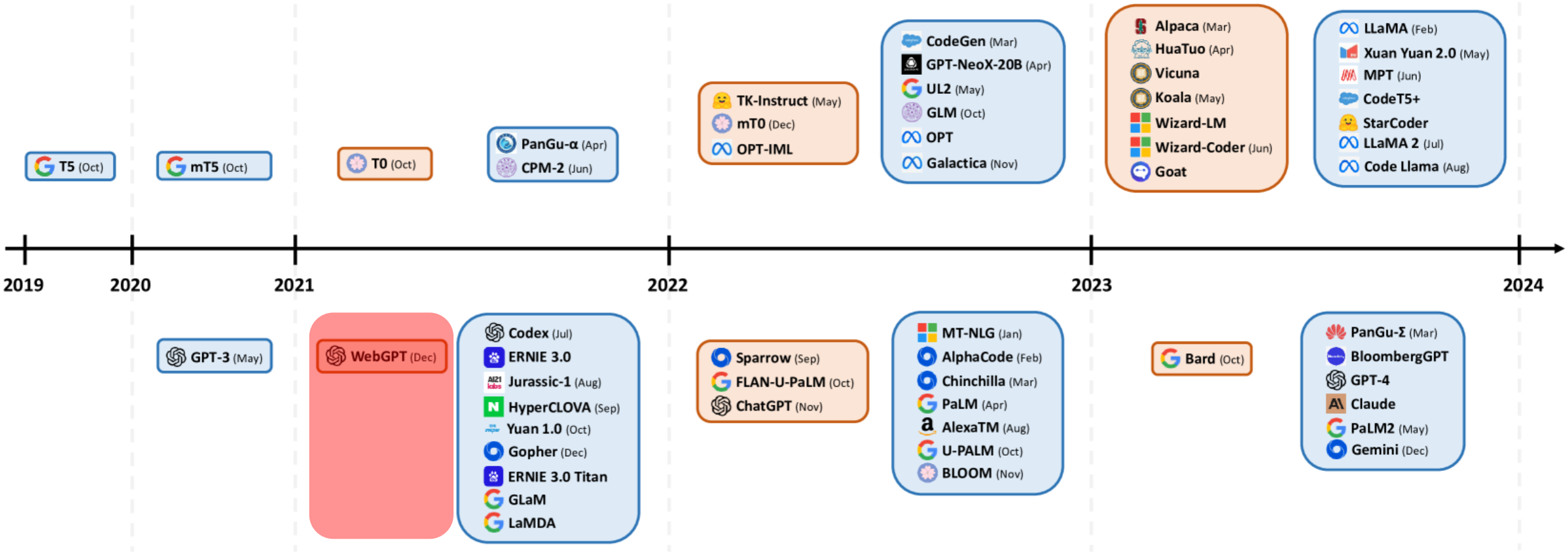
HOGESCHOOL ROTTERDAM

"Hoe kan het gebruik van Gen-AI valide en veilig worden toegepast?"

hr.nl/ai

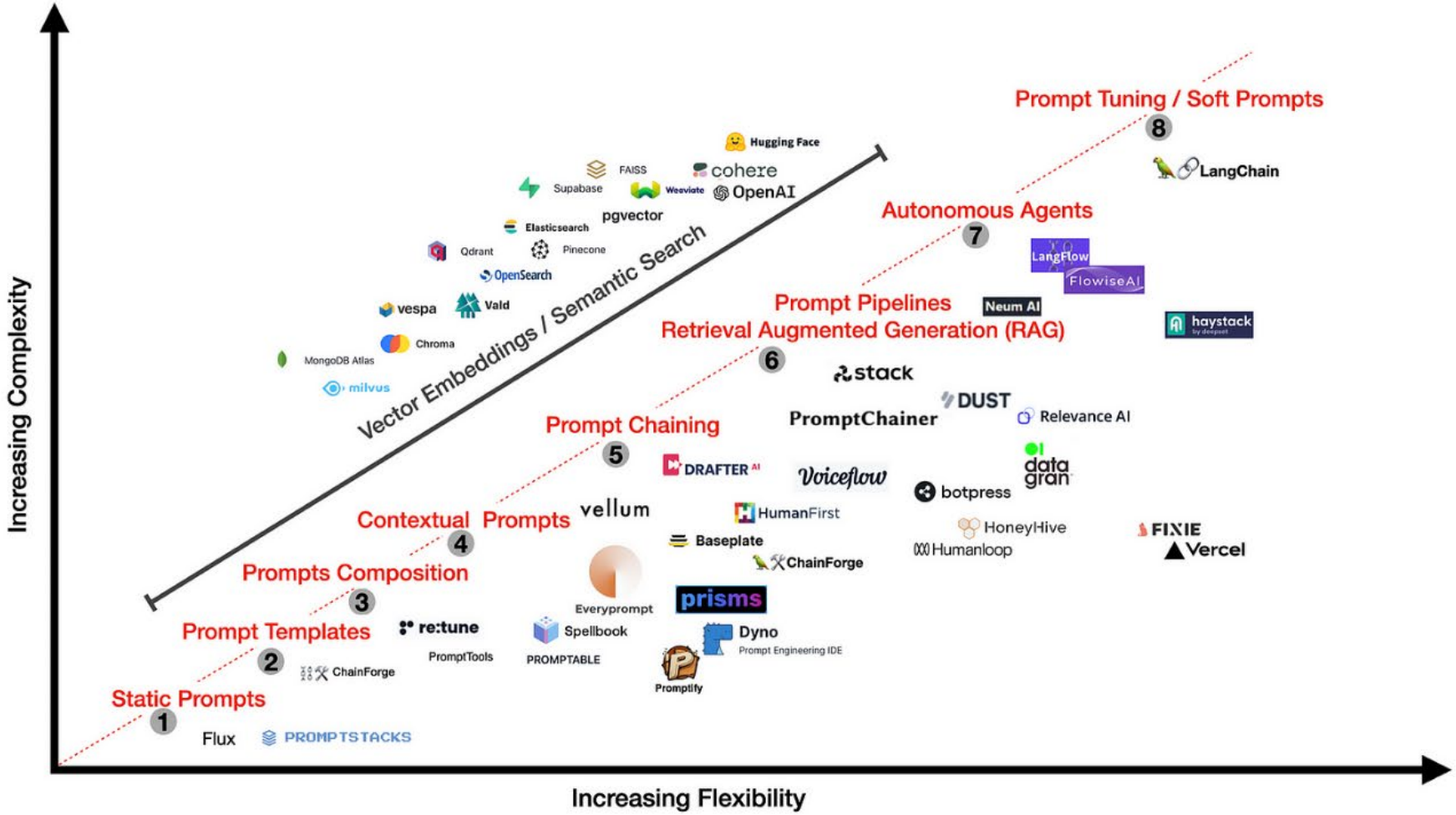
Ontstaansgeschiedenis + evolutie van grote taal modellen {LLM}

HOGESCHOOL ROTTERDAM



https://arxiv.org/html/2307.06435v9

LLM implementaties



https://blogs.novita.ai/exploring-architectural-structures-and-functional-capacities-of-llms/

Misvattingen die publieke organisaties ervan weerhouden grote taal modellen {LLM} *lees generatieve AI* optimaal te benutten

Misvatting 1:
Mijn organisatie beschikt niet over de juiste tools en platforms om betrouwbare AI te ontwikkelen

Misvatting 2:
AI is te gecompliceerd en veel te kostbaar voor mijn organisatie om vanaf nul op te bouwen

Misvatting 3:
Mijn organisatie beschikt niet over de juiste vaardigheden en expertise om AI te ontwikkelen

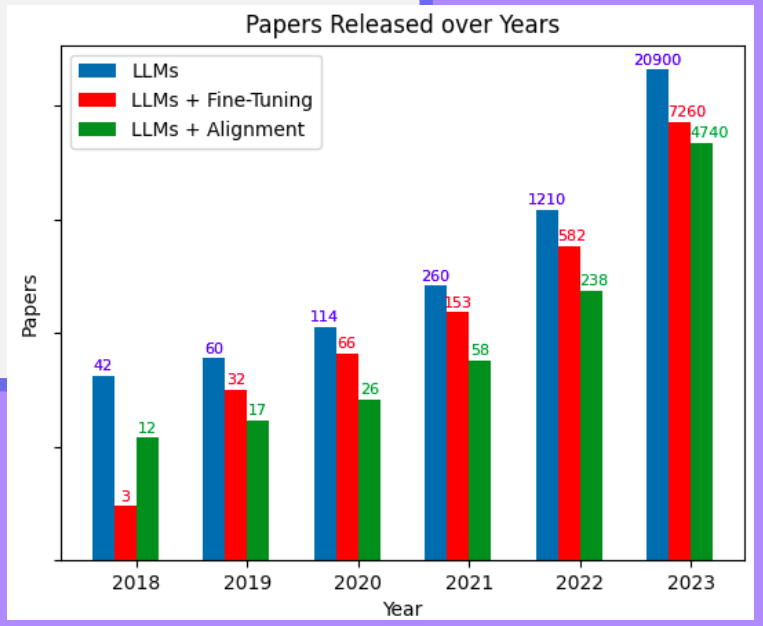
Misvatting 4:
De AVG staat het toepassen van AI in het onderwijs in de weg

HOGESCHOOL ROTTERDAM

CONTEXT:

Wat zijn grote taal modellen eigenlijk?

Waarom heeft Generatieve AI zo'n enorme impact op onderwijs & onderzoek



Defining AI is
nonsensical

REPORT / STUDY | Publication 18 December 2018

A definition of Artificial Intelligence: main capabilities and scientific disciplines

This document expands the definition of Artificial Intelligence (AI) as defined in the Commission Communication on AI. It clarifies certain aspects of AI as a scientific discipline and as a technology, with the aim to avoid misunderstanding, to achieve a shared common knowledge of AI that can be fruitfully used also by non - AI experts, and to provide useful details that can be used in the discussion on both the AI ethics guidelines and the AI policies recommendations.

Downloads

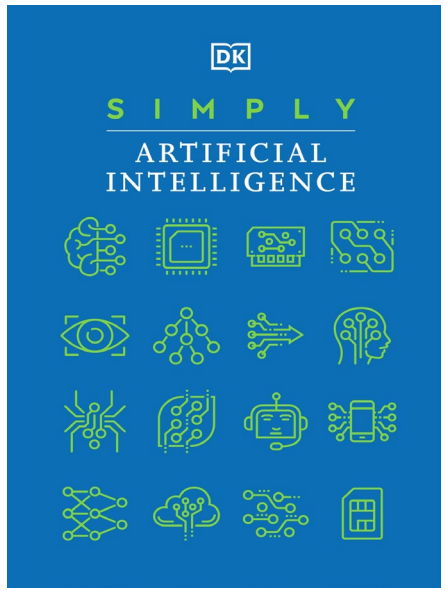
AI Definition.pdf

Download

Related topics

Artificial intelligence

Advanced Digital Technologies



DK | Penguin Random House

DK LONDON
Senior Editor Chauvnay Dunford
Senior Designer Mark Cavanagh
Editors Daniel Bryne, Elizabeth Dowsett, Steve Setford, Alison Sturgeon, Andrew Szudek
Managing Editor Gareth Jones
Senior Managing Art Editor Lee Griffiths
Production Editor Robert Dunn
Senior Production Controller Rachel Ng
Jacket Design Development Manager Sophia M.T.T.
Jacket Designer Akiko Kato
Associate Publishing Director Liz Wheeler
Art Director Karen Self
Publishing Director Jonathan Metcalf

First American Edition, 2023
Published in the United States by DK Publishing
1745 Broadway, 20th Floor, New York, NY 10019

23 24 24 26 10 9 8 7 6 5 4 3 2 1
001-334439-Mar/2023

Published in Great Britain
by Dorling Kindersley Limited

A catalog record for this book is available from the Library of Congress.
ISBN 978-0-7440-7682-0

Printed and bound in China

For the curious
www.dk.com

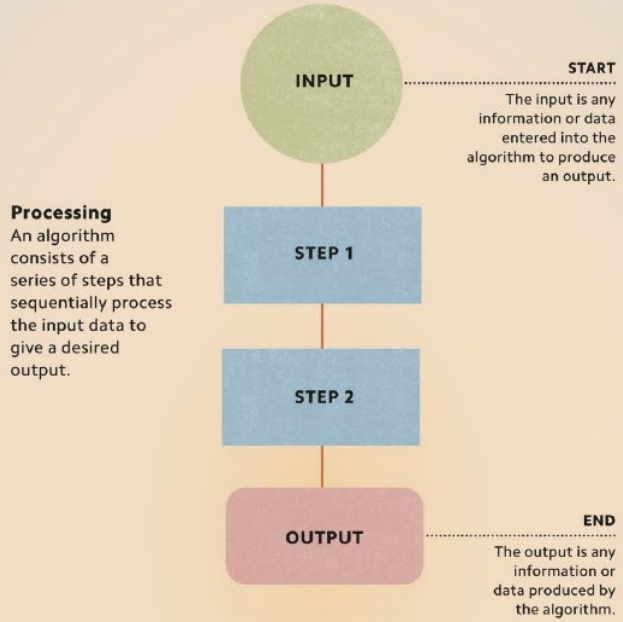
This book was made with Forest Stewardship Council™ certified paper—one small step in DK's commitment to a sustainable future. For more information go to www.dk.com/our-green-pledge

FSC MIX Paper | Supporting responsible forestry FSC® C018179

CONSULTANT
Hilary Lamb is an award-winning science and technology journalist, editor, and author. She studied physics at the University of Bristol and science communication at Imperial College London before spending five years as a staff magazine reporter. She has worked on previous DK titles, including *How Technology Works*, *The Physics Book*, and *Simply Quantum Physics*.

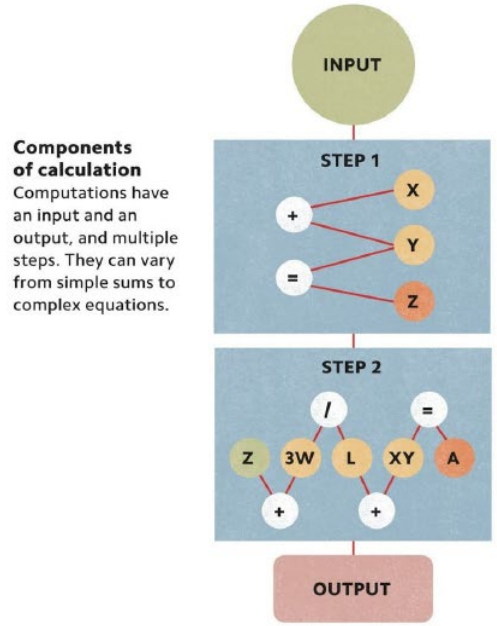
EDITORIAL CONSULTANT
Joel Levy is a writer who specializes in science and the history of science. His writing explores both mainstream science and weird technology, and his books include *The Infinite Tortoise: The Curious Thought Experiments of History's Great Thinkers*, *Gothic Science: The Era of Ingenuity and the Making of Frankenstein*, and *Reality Ahead of Schedule: How Science Fiction Inspires Science Fact*.

CONTRIBUTOR
Dr. Claire Quigley is a computing scientist who has previously worked for the Universities of Cambridge and Glasgow. She has contributed to the creation of coding activities for the BBC and Virgin Media, and written for several previous DK titles, including *Help Your Kids with Computer Science*, *Computer Coding Python Games for Kids*, and *Computer Coding Python Projects for Kids*.



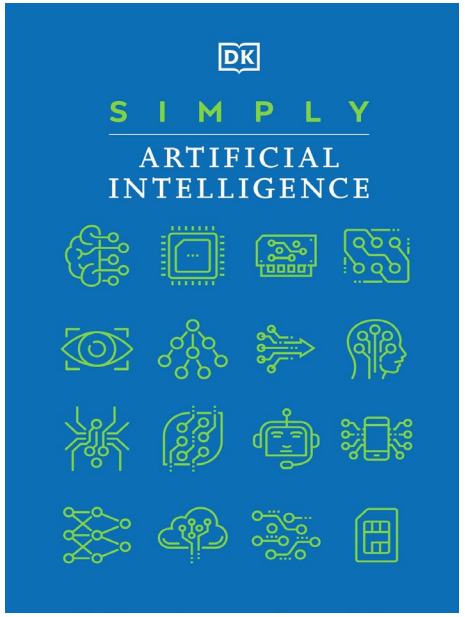
STEP BY STEP

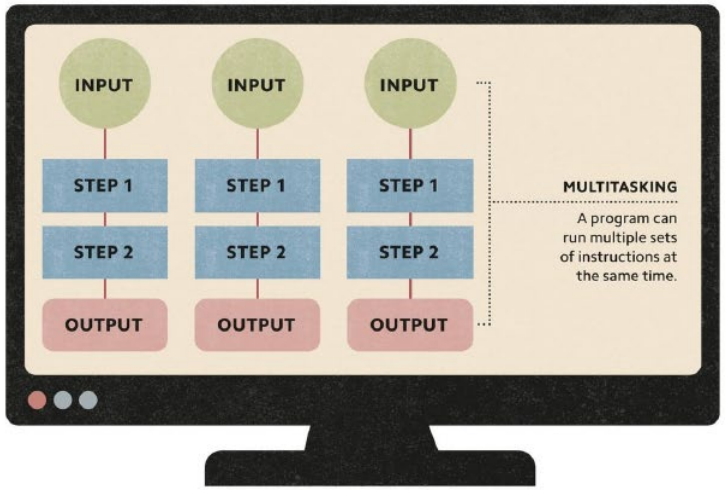
An algorithm is a sequence of instructions for accomplishing a task. It takes an input, such as information or data, and processes it in a series of steps to produce a desired result, or output. The task or process can range from a simple calculation, or following a recipe to make a meal, to solving complex mathematical equations. An algorithm is an example of what mathematicians call an "effective method," which means it has a finite number of steps and produces a definite answer, or output.



ALGORITHMS IN ACTION

A computation is a calculation that follows the steps of an algorithm (see opposite). The most straightforward example of computation is arithmetic calculation. For example, if you add together a pair of three-digit numbers in your head, you follow a series of steps, or an algorithm, to achieve this calculation. Computations use symbols to represent numbers, but symbols can represent almost anything else (see p.36). With the right symbols and the right algorithms, immensely complex computation becomes possible.



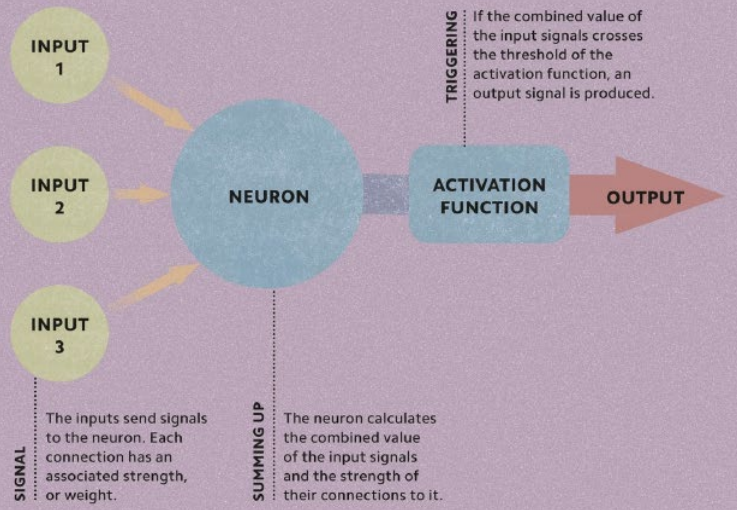


INSTRUCTING COMPUTERS

A program is a sequence of instructions written in code that enables a computer to perform one or more tasks. Charles Babbage (see opposite) imagined the first program. He was inspired by the design of a certain silk loom, which had parts that moved up or down in response to a pattern of holes punched into a card. Babbage recognized that these holes could store instructions to operate the cogs and levers of a machine he was designing: the "Analytical Engine". Modern computers work on the same principle, following sequences of instructions, which are usually written in binary code (see p.13).

ARTIFICIAL NEURONS

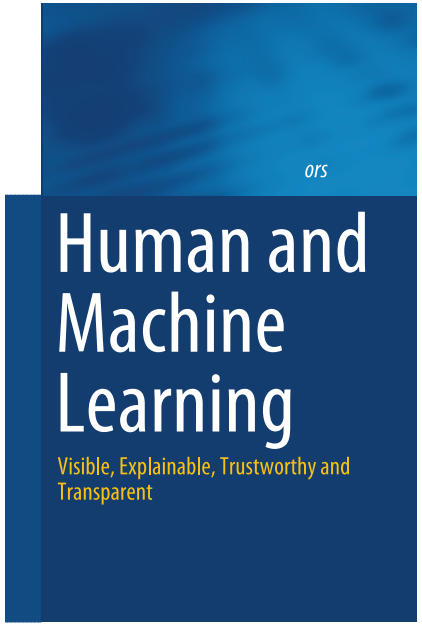
Each of the 86 billion neurons in the human brain is effectively a tiny processor, receiving electrical signals (inputs) from other neurons and sending out signals of its own (outputs). McCulloch and Pitts (see opposite) realized that neurons can act as logic gates—devices that can switch on and off (see p.13), depending on the input. The scientists described an imaginary neuron called a "threshold logic unit". This neuron works by first adding the values of its inputs (signals from other neurons) and then multiplying that value by a variable called a "weight" (see p.78)—this is the strength of a connection between neurons. If the input signals exceed a certain value (see p.79), the neuron is triggered to send an output signal. This triggering is called the "activation function".



Defining what characterizes AI *gives Insight*

{Human-in-the-Loop}

HOGESCHOOL ROTTERDAM



AI = ML + TD + HITL

Artificial Intelligence: in contrast to natural intelligence, it is *the ability of computer systems to perform tasks or actions that would normally require a human*

Machine Learning: *the ability of computer systems to use algorithms and statistical models to perform tasks without explicit instruction, through patterns and inferences*

Training Data: *the data used to train a machine learning algorithm to perform a task in supervised machine learning*

Human in the Loop: *the involvement of a human in training a machine learning algorithm*

AI-taxonomie is complex

HOGESCHOOL ROTTERDAM

https://www.researchgate.net/publication/359424818_Designing_Neural_Networks_Through_Sensory_Ecology_Biology_to_the_rescue_of_AI_Produced_by_Living-Lab_AiRA_Hub_voor_Data_Responsible_AI_Hogeschool_Rotterdam_Lunch-Lezing_Creating-010_FEB_2022

Representations are localized and knowledgeable to humans

STRONG AI Human-level Intelligence

Good-Old-Fashioned Artificial Intelligence

GOFAI symbolic → Reasoning & Logic → Acting Humanly / Thinking Humanly

AI Capability Levels

Top-down approach

Bottom-up approach

Rule driven

Data driven

WEAK AI (narrow) single purpose Intelligence

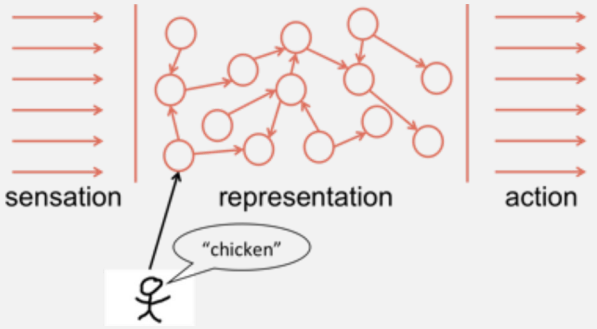
CI sub-symbolic → Machine Learning & Self-organizing → Connect-ionist models / Embodied (embedded) Artificial lifeEm / Distributed Parallel Repre-sentation

Modern Computational Intelligence

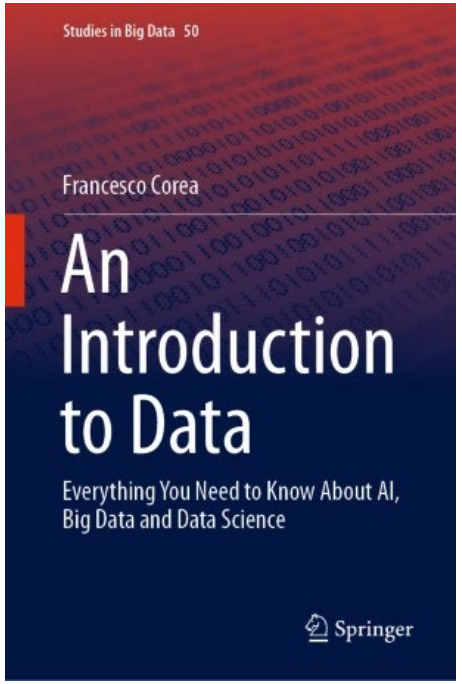
Representations are distributed not readable for humans

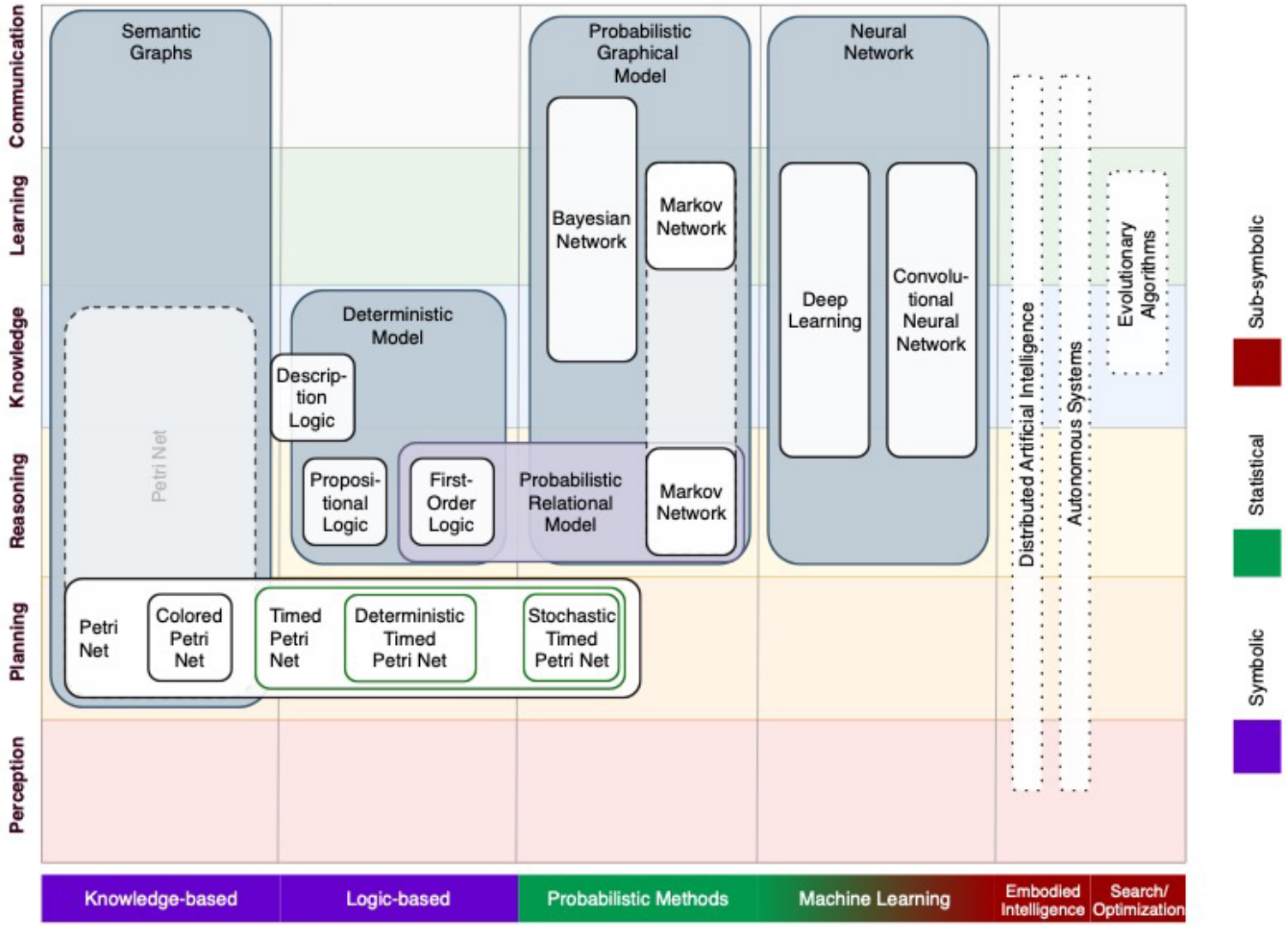
Aoccdrnig to rscheearch at Cmabrigde Uinervtisy, it deosn't mttaer in waht oredr the ltteers in a wrod are, the olny iprmoetnt tihng is taht the frist and lsat ltteer be at the rghit pclae. The rset can be a toatl mses and you can sitll raed it wouthit porbelm. Tihs is bcuseae the huamn mnid deos not raed ervey lteter by istlef, but the wrod as a wlohe.

SUB-SYMBOLISCH



{01}
Fundamentals





Artificial Intelligence:
How knowledge is created, transferred, and used

Trends in China, Europe, and the United States

HOGESCHOOL ROTTERDAM

Foreword

Defining AI: new approaches help with AI ontologies

Prof. Enrico Motta,
Professor of Knowledge Technologies, The Open University, United Kingdom

"Disciplines do not exist per se. They emerge because of a collective construction process, whereby a community of researchers comes together, formulating and sharing common objectives, methods, and conceptualizations. Hence, disciplines are essentially about research communities. As these evolve, so do the associated disciplines. Thus, attempts at characterizing disciplines are in my view more successful if they follow a bottom-up approach, focusing less on top-down definitions than on identifying the relevant body of work.

Given this premise, I am very happy to endorse this report produced by the Elsevier team, which provides an operational characterization of the field of AI, in terms of 600,000 documents and over 700 field-specific keywords. This is an impressive piece of work that, to my knowledge, provides the most comprehensive characterization of AI outputs produced so far. Crucially, in contrast with manually developed taxonomies of research areas, which inevitably end up reflecting the specific viewpoints of the experts involved in the process, this characterization is data-driven, using machine learning and text mining techniques to classify documents and identify the relevant keywords. Thus, in my view, the report enjoys greater validity, providing a more objective reflection of the variety of existing contributions to the AI field.

In addition to its scientific value, there is also no doubt that this report will be a very valuable practical resource for people who wish to explore this space. For example, it will be very interesting to use this comprehensive characterization of the AI field to get a better understanding of key trends and topics, especially when the relevant body of work may be spread across different research communities, as it is the case, fo instance, with work in the highly importar of AI ethics.

On a personal level, this work is also very exciting for me because it provides the ba for interesting new research. One of my main research areas concerns the use of A technologies to develop innovative solutio can help people to make sense of the dyn of scientific research. Within this broad co my team has developed an original appro to the automatic generation of taxonomie research areas and, for example, it would l extremely interesting to investigate to wh extent these different methods can cover t research space and to what extent they ca combined to improve accuracy. This is just example of the many interesting possibilit further research opened up by this work.

In sum, this is not just an excellent piece c work, but also the start of a very interestin of research. I congratulate the Elsevier tea their tremendous work and I look forward further developments in this space."

2.2 Seven AI research clusters

We aimed to provide more depth to our subsequent analyses by structuring AI into research areas, using an unsupervised clustering technique.[58] This approach maps the keywords of all perspectives into clusters and illustrates their connections, based on co-occurrence within the documents. Co-occurrence indicates that those clusters do not stand alone, but strongly relate to each other, e.g., neural networks in a computer vision document. How do capabilities connect with each other and to application fields? The resulting graph illustrates the subfields of AI (Figure 2.2) and their connections through co-occurrence in scholarly publications. On the Elsevier AI Resource Center,[59] the graph is interactive, allowing users to browse individual connections and clusters, by region and over time.

As shows in Figure 2.2, AI seems to cluster around the areas of Search and Optimization, Fuzzy Systems, Natural Language Processing and Knowledge Representation, Computer Vision, Machine Learning and Probabilistic Reasoning, Planning and Decision Making, and Neural Networks. Societal application fields, such as self-driving cars or robotics, are embedded into Planning and Decision Making as they have fewer underlying publications. The clusters seem to focus on statistics-based AI. Knowledge-based capabilities, such as "Ontologies or Semantics," do not form a cluster on their own, but are embedded in other clusters, predominantly in "Natural Language Processing and Knowledge Representation." Further research might investigate the sensitivity of this approach to the number of keywords and related publications in terms of normalized proportions over time. The strong growth of publications in recent years within the learning system field might outweigh knowledge-based approaches from more than 15 years ago.

Figure 2.2 illustrates the breadth of industry keywords (green), especially in the areas of "Fuzzy Systems" and "Computer Vision," whereas specific research keywords appear in "Neural Networks," teaching keywords in "Search and Optimization," and media keywords in fields such as "Planning and Decision Making" and "Natural Language Processing and Knowledge Representation."[60] The relatively low proportion of media-driven keywords could indicate that these are not key AI research fields, or that they are still in their research infancy, representing only a fraction of AI documents.

The online interactive graph[61] allows the exploration of connections and co-occurrences through time. For instance, it shows the intensification of the two clusters "Machine Learning and Probabilistic Reasonin" and "Neural Networks." It also reveals that the clusters "Deep Learning" in 2003 and "Swarm Intelligence" in 2000 have no co-occurring keywords but grow to become visible nodes on the graph in more recent years. Co-occurrences illustrate that certain learning system and neural network approaches are predominantly used in specific application fields, like "Recurrent Neural Networks" with "Natural Language Processing and Knowledge Representation." They also show that the keyword "Convolutional Neural Networks" is linked with "Computer Vision" and "Collaborative Filtering" with "Recommender Systems." Some connections indicate potential hierarchical relations, such as "Artificial Intelligence" co-occurring with the keyword "Neural Network," and further co-occurring with specific forms of neural networks.

58 Louvain clustering:
https://perso.uclouvain.be/vincent.blondel/research/louvain.html
"The Louvain method is a simple, efficient and easy-to-implement method for identifying communities in large networks...The method is a greedy optimization method that attempts to optimize the 'modularity' of a partition of the network...The original idea for the method is due to Etienne Lefebvre who first developed it during his Master thesis at UCL (Louvain-la-Neuve) in March 2005...The method was first published in: 'Fast unfolding of communities in large networks,'Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, Etienne Lefebvre, Journal of Statistical Mechanics: Theory and Experiment 2008 (10), P10008 (12pp) doi: 10.1088/1742-5468/2008/10/P10008. arXiv: http://arxiv.org/abs/0803.0476."

59 Elsevier. Artificial Intelligence Resource Center.
https://www.elsevier.com/connect/ai-resource-center.

60 Learn more about these in Science Direct Topic Pages:
https://www.sciencedirect.com/topics/index
Fuzzy Systems: https://www.sciencedirect.com/topics/chemical-engineering/fuzzy-systems; Speech Recognition:
https://www.sciencedirect.com/topics/neuroscience/speech-recognition;
Computer Vision: https://www.sciencedirect.com/topics/food-science/computer-vision-technology (specific application field) or Face recognition:
https://www.sciencedirect.com/topics/neuroscience/face-recognition;
Learning Systems: https://www.sciencedirect.com/topics/chemical-engineering/learning-systems;
Neural Networks: https://www.sciencedirect.com/topics/veterinary-science-and-veterinary-medicine/neural-network-software.

61 Elsevier. Artificial Intelligence Resource Center.
https://www.elsevier.com/connect/ai-resource-center.

The AI research field clusters around seven main research areas.

In summary, our co-occurrence analysis reveals a sub-structure of the AI field, determined by its document corpus and keyword selection. Those might influence the weighting and share of subfields, such as knowledge-based fields. Further research might explore normalized approaches to compare subfields per year and investigate the sensitivity of keywords to the structure of the field. In chapter 3 we will use these clusters to present global and regional trends in AI.

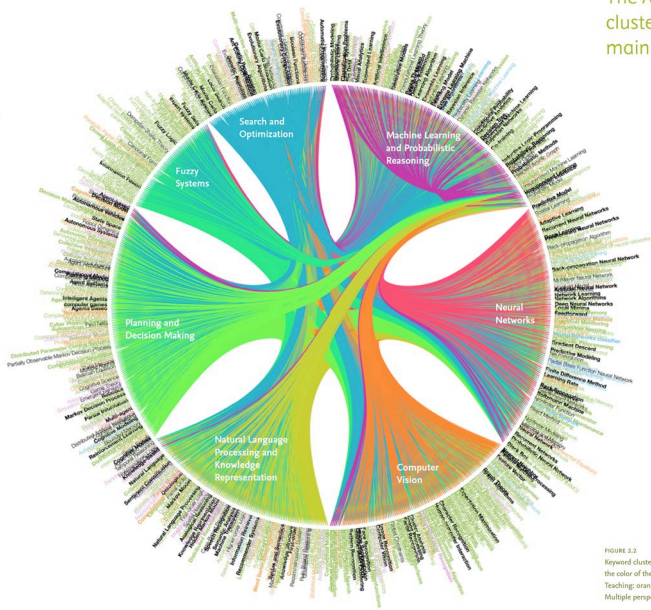


FIGURE 2.2
Keyword clusters and co-occurrences in the AI field, 2017; the color of the keyword represents its originating perspective: Teaching: orange, Research: blue, Industry: green, Media: pink, Multiple perspectives: black. source: Scopus.

HOGESCHOOL ROTTERDAM

software

algorithms in decision making

fuzzy boundaries

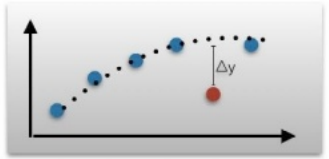
rule-based decision making

if condition fulfilled *then*
activity 1
else
activity 2

boolean data
(yes or no)

Examples:
- phone notification
- time- or threshold-based alarms
- simple pattern matching

statistical reasoning

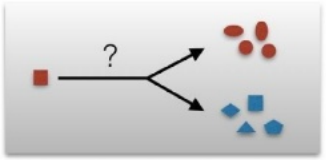


simple regression

numerical data
allowing for curve fitting

Examples:
- extra- and interpolation
- outlier detection
- predictive maintenance

machine learning

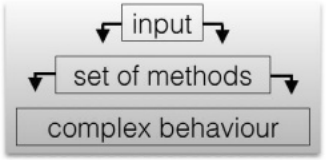


classification tasks

arbitrary data
that needs to be abstracted into numbers

Examples:
- identification of relevant features from large input datasets
- quality control using various metrics

artificial intelligence



dynamic adaptation to novelty

autonomous selection of best methodology when presented with **arbitrary data**

Examples:
- autonomous vehicles
- human-like conversational skills
- intelligent digital assistant

every programmer

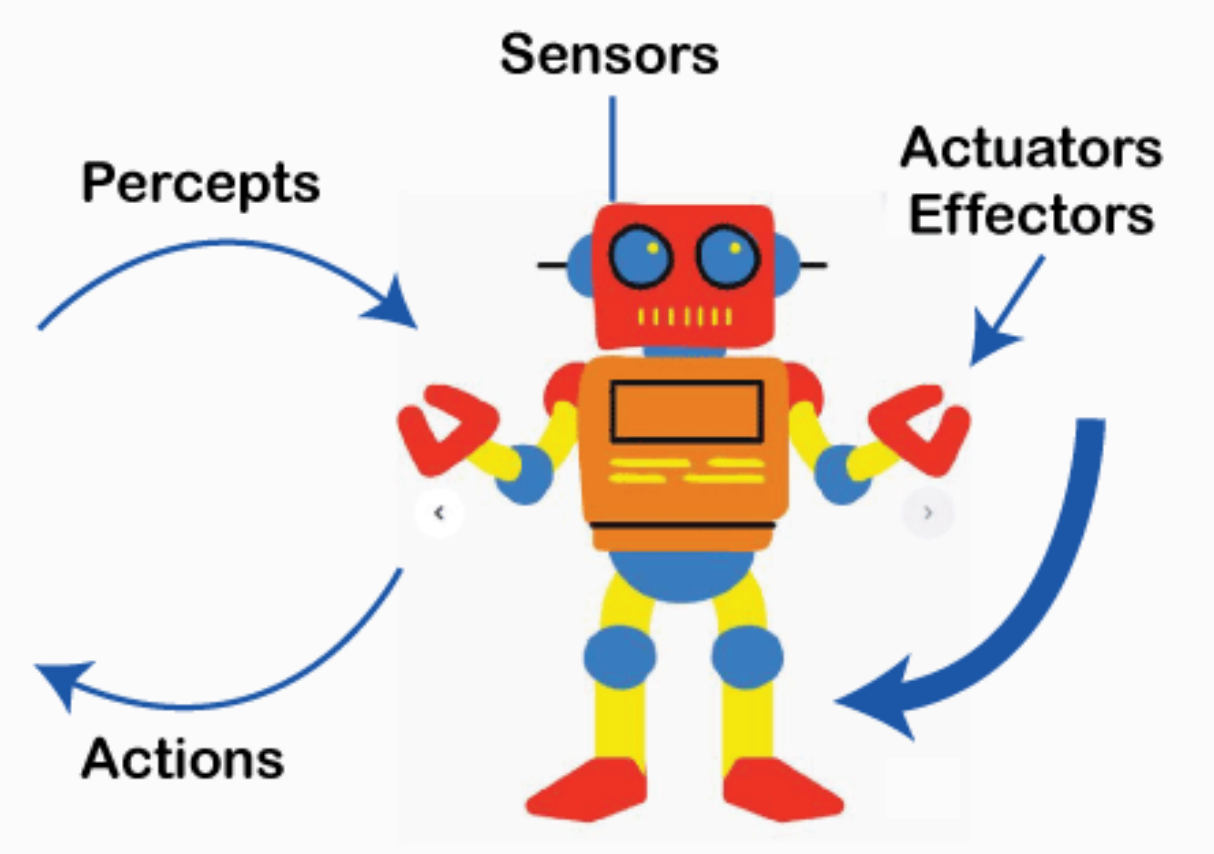
data science types

complex systems specialists

multimodal agent

HOGESCHOOL ROTTERDAM

EXTERNAL INPUT



Generatieve AI according to Google

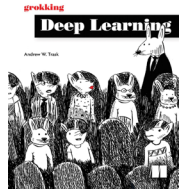
Neurale netwerk modellen die machinaal hebben geleerd op basis van bestaande multimodale content (trainen van het model) nieuwe inhoud te creëren, zoals tekst, afbeeldingen, muziek en code.

Generatieve AI zijn geen informatiedatabases of deterministische informatiezoeksystemen, omdat het voorspellingssystemen zijn.

https://ai.google/static/documents/google-about-generative-ai.pdf

HOGESCHOOL ROTTERDAM

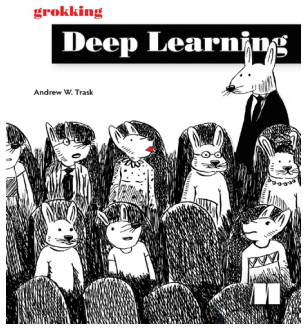
What is machine learning?



> A field of study that gives computers the ability to learn without being explicitly programmed.
>
> —Attributed to Arthur Samuel

Given that deep learning is a subset of machine learning, what is machine learning? Most generally, it is what its name implies. Machine learning is a subfield of computer science wherein *machines learn* to perform tasks for which they were *not explicitly programmed*. In short, machines observe a pattern and attempt to imitate it in some way that can be either direct or indirect.

Machine learning ~= Monkey see, monkey do



HOGESCHOOL ROTTERDAM

Supervised machine learning

Supervised learning transforms datasets.

Supervised learning is a method for transforming one dataset into another. For example, if you had a dataset called Monday Stock Prices that recorded the price of every stock on every Monday for the past 10 years, and a second dataset called Tuesday Stock Prices recorded over the same time period, a supervised learning algorithm might try to use one to predict the other.

If you successfully trained the supervised machine learning algorithm on 10 years of Mondays and Tuesdays, then you could predict the stock price on any Tuesday in the future given the stock price on the immediately preceding Monday. I encourage you to stop and consider this for a moment.

Supervised machine learning is the bread and butter of applied artificial intelligence (also known as narrow AI). It's useful for taking *what you know* as input and quickly transforming it into *what you want to know*. This allows supervised machine learning algorithms to extend human intelligence and capabilities in a seemingly endless number of ways.

The majority of work using machine learning results in the training of a supervised classifier of some kind. Even unsupervised machine learning (which you'll learn more about in a moment) is typically done to aid in the development of an accurate supervised machine learning algorithm.

For the rest of this book, you'll be creating algorithms that can take input data that is observable, recordable, and, by extension, *knowable* and transform it into valuable output data that requires logical analysis. This is the power of supervised machine learning.

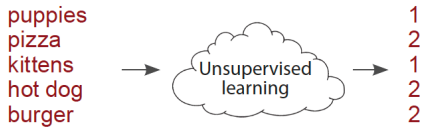
Unsupervised machine learning

Unsupervised learning groups your data.

Unsupervised learning shares a property in common with supervised learning: it transforms one dataset into another. But the dataset that it transforms into is *not previously known or understood*. Unlike supervised learning, there is no "right answer" that you're trying to get the model to duplicate. You just tell an unsupervised algorithm to "find patterns in this data and tell me about them."

For example, *clustering a dataset into groups* is a type of unsupervised learning. Clustering transforms a sequence of *datapoints* into a sequence of *cluster labels*. If it learns 10 clusters, it's common for these labels to be the numbers 1–10. Each datapoint will be assigned to a number based on which cluster it's in. Thus, the dataset turns from a bunch of datapoints into a bunch of labels. Why are the labels numbers? The algorithm doesn't tell you what the clusters are. How could it know? It just says, "Hey scientist! I found some structure. It looks like there are groups in your data. Here they are!"

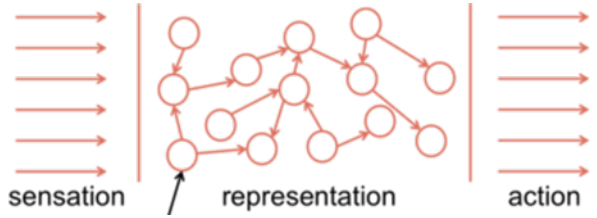
I have good news! This idea of clustering is something you can reliably hold onto in your mind as the definition of unsupervised learning. Even though there are many forms of unsupervised learning, *all forms of unsupervised learning can be viewed as a form of clustering*. You'll discover more on this later in the book.



Check out this example. Even though the algorithm didn't tell what the clusters are named, can you figure out how it clustered the words? (Answer: 1 == cute and 2 == delicious.) Later, we'll unpack how other forms of unsupervised learning are also just a form of clustering and why these clusters are useful for supervised learning.

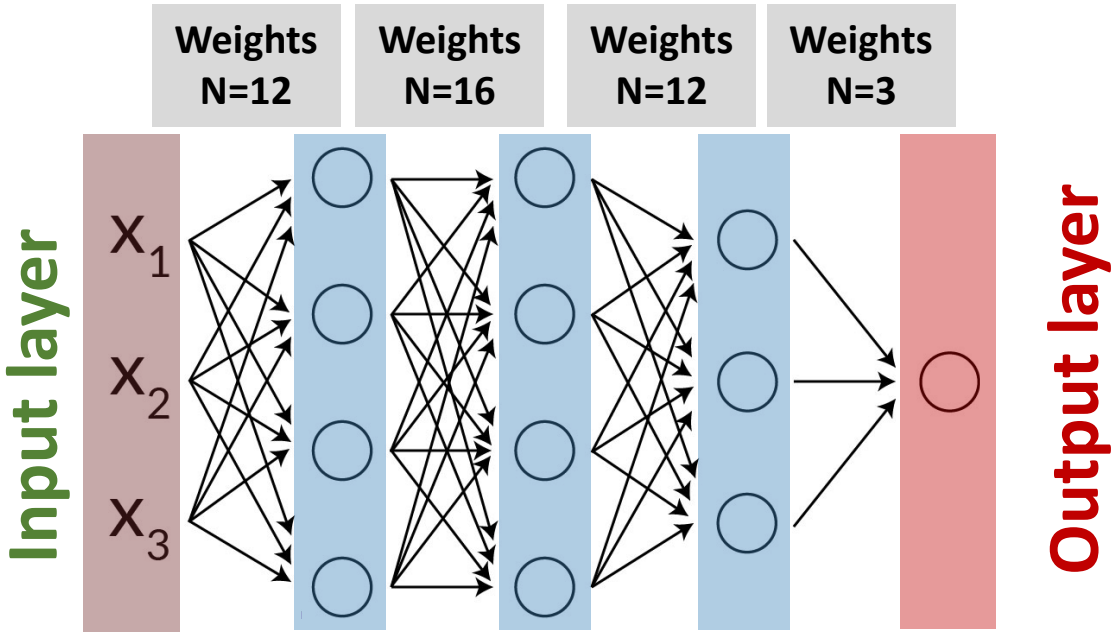
{NN Layers}

SUB-SYMBOLISCH



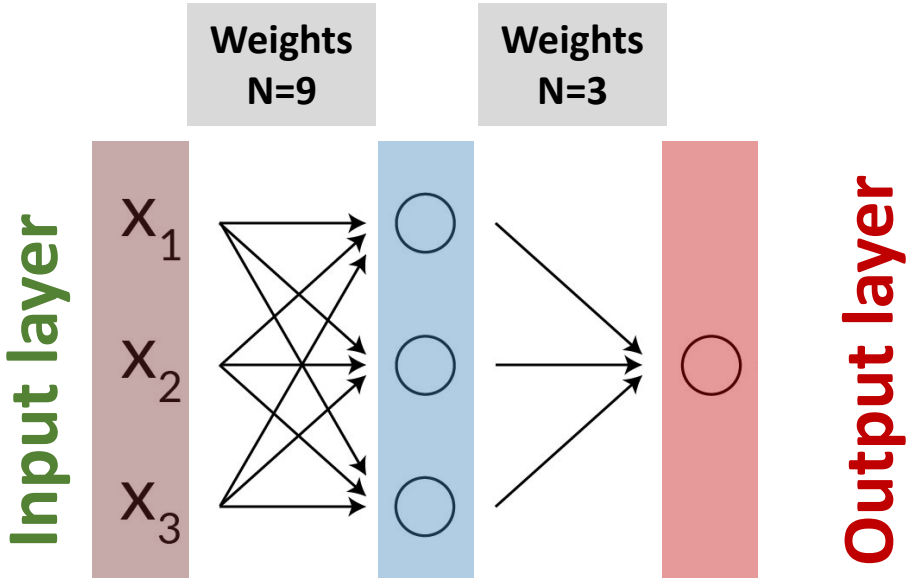
HOGESCHOOL ROTTERDAM

Neural Network {NN} Layer Architecture



hidden layers N>2

Deep network



Hidden layer N=1

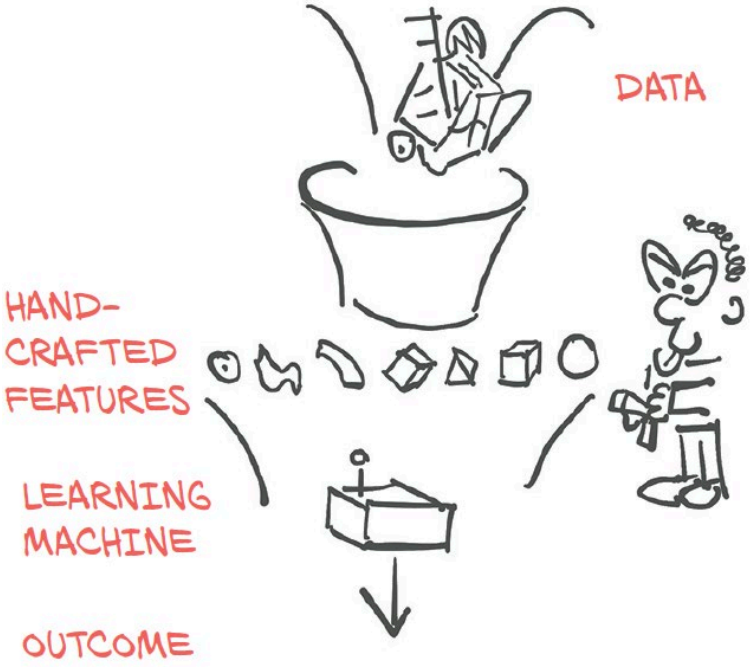
Shallow network

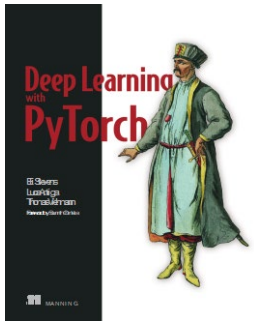
{AI Paradigm-shift}

HOGESCHOOL ROTTERDAM

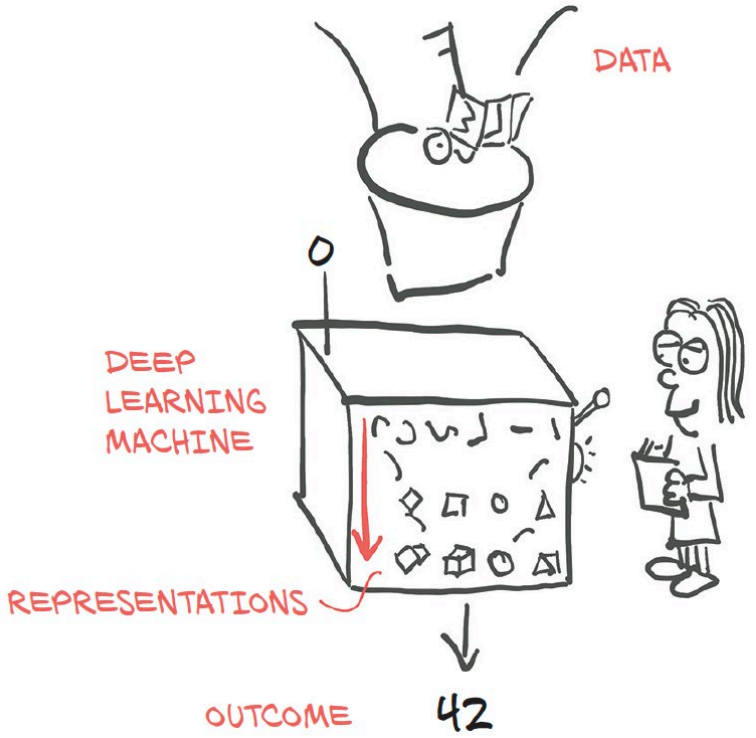
More data, parameters & computing power | Less human-in-the-loop

Machine Learning Paradigm {ML}





Deep Learning Paradigm {DL}

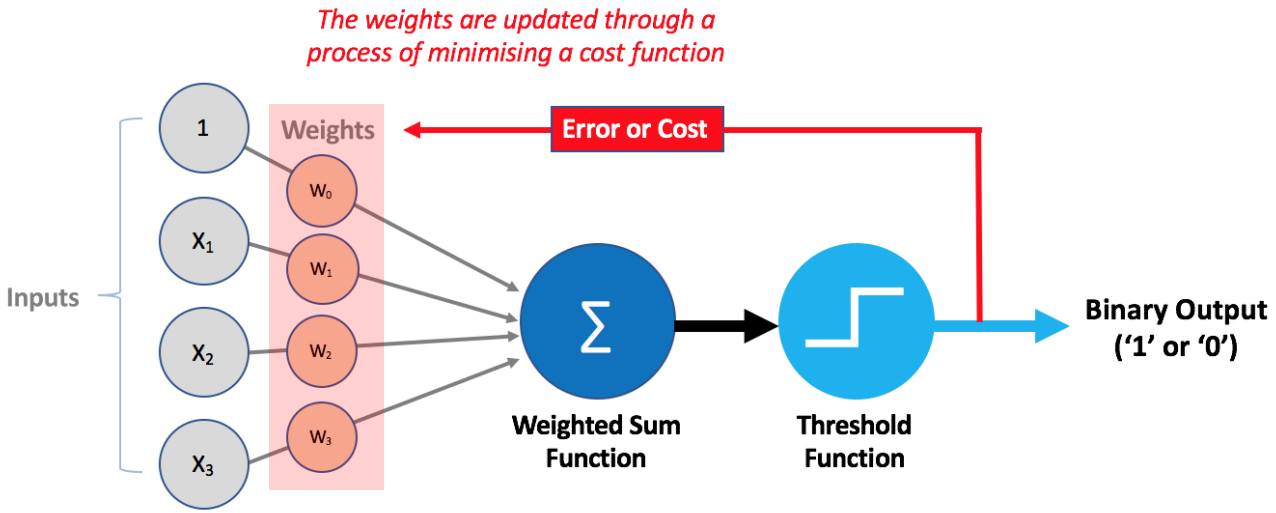


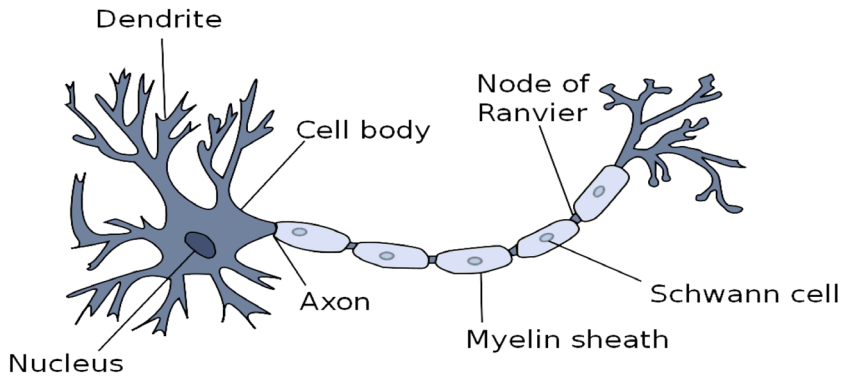
{Artificial Neurons}

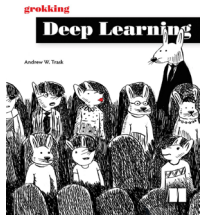
HOGESCHOOL ROTTERDAM
CREATING

Deep Neural Nets {DNNs} harbor vast amounts of "artificial neurons" ➔smallest computational unit⬅

Names for Artificial Neurons
{unit}
{cell}
{node}
{perceptron}







HOGESCHOOL ROTTERDAM

Backpropagation in code

You can learn the amount that each weight contributes to the final error.

At the end of the previous chapter, I made an assertion that it would be important to memorize the two-layer neural network code so you could quickly and easily recall it when I reference more-advanced concepts. This is when that memorization matters.

The following listing is the new learning code, and it's essential that you recognize and understand the parts addressed in the previous chapters. If you get lost, go to chapter 5, memorize the code, and then come back. It will make a big difference someday.

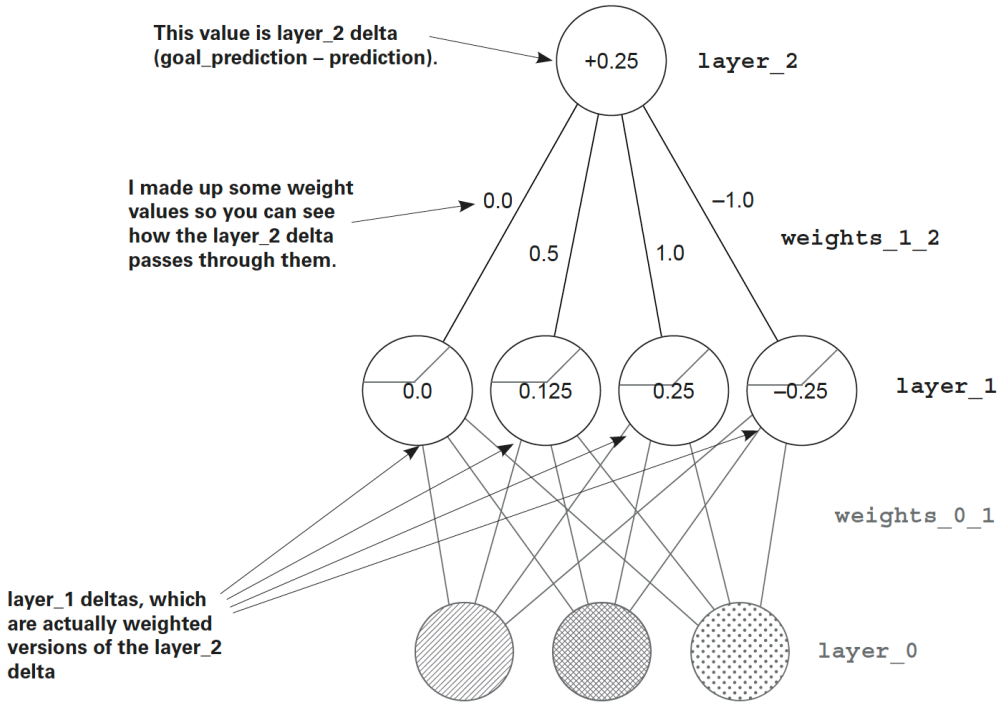
```
import numpy as np

np.random.seed(1)

def relu(x):
    return (x > 0) * x

def relu2deriv(output):
    return output>0

alpha = 0.2
hidden_size = 4

weights_0_1 = 2*np.random.random((3,hidden_size)) - 1
weights_1_2 = 2*np.random.random((hidden_size,1)) - 1

for iteration in range(60):
   layer_2_error = 0
   for i in range(len(streetlights)):
      layer_0 = streetlights[i:i+1]
      layer_1 = relu(np.dot(layer_0,weights_0_1))
      layer_2 = np.dot(layer_1,weights_1_2)

      layer_2_error += np.sum((layer_2 - walk_vs_stop[i:i+1]) ** 2)

      layer_2_delta = (walk_vs_stop[i:i+1] - layer_2)
      layer_1_delta=layer_2_delta.dot(weights_1_2.T)*relu2deriv(layer_1)

      weights_1_2 += alpha * layer_1.T.dot(layer_2_delta)
      weights_0_1 += alpha * layer_0.T.dot(layer_1_delta)

   if(iteration % 10 == 9):
      print("Error:" + str(layer_2_error))
```

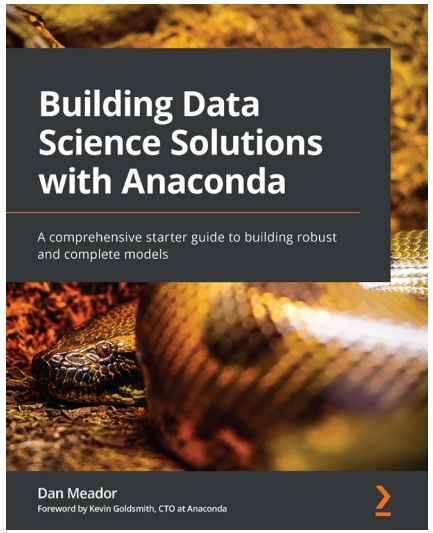
Returns x if x > 0; returns 0 otherwise

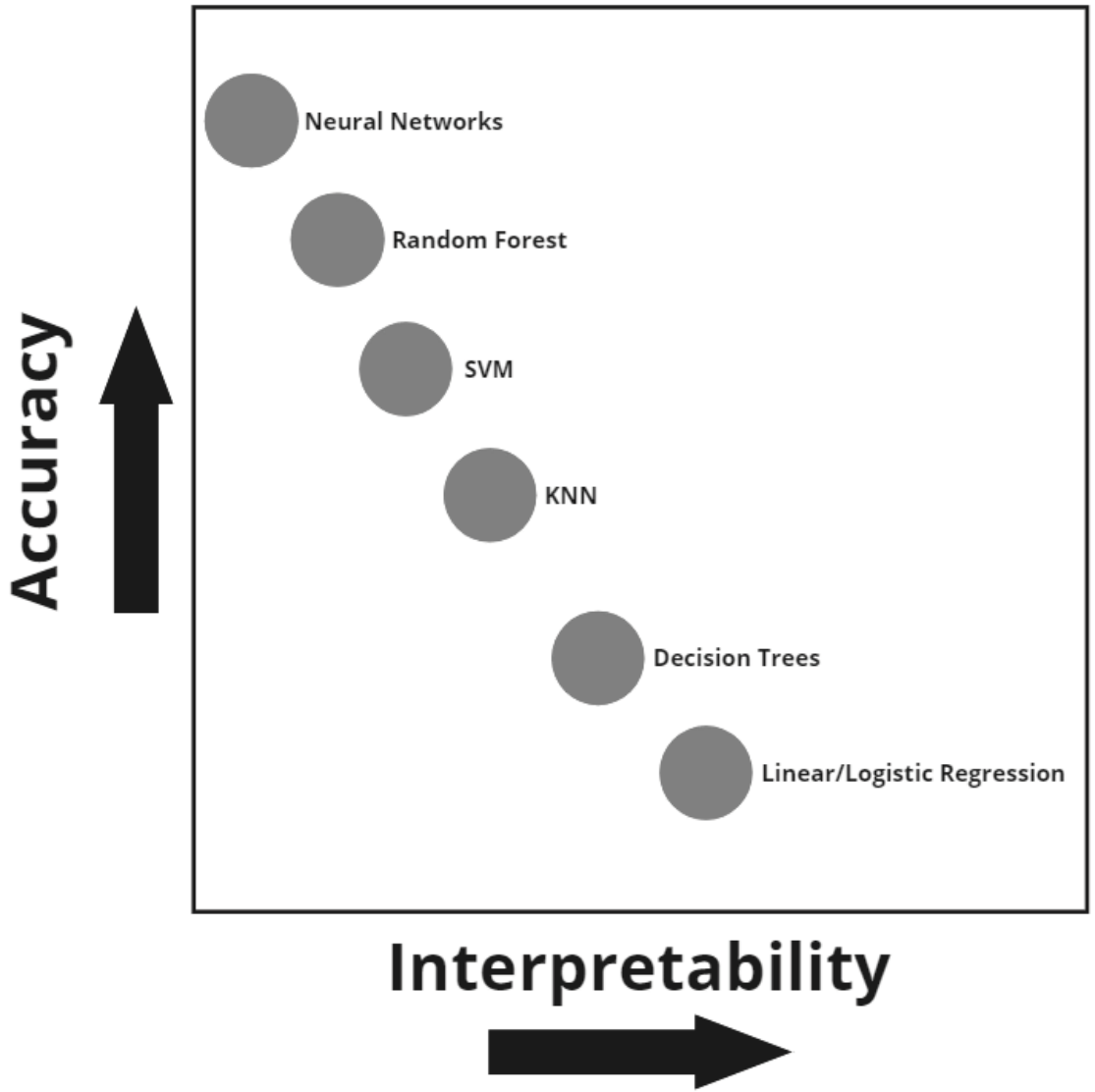
Returns 1 for input > 0; returns 0 otherwise

This line computes the delta at layer_1 given the delta at layer_2 by taking the layer_2_ delta and multiplying it by its connecting weights_1_2.

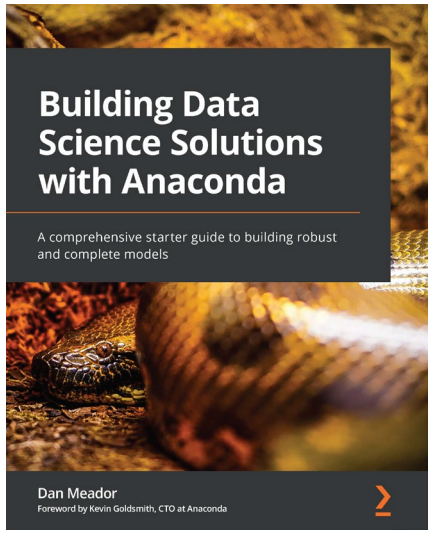
Believe it or not, the only truly new code is in bold. Everything else is fundamentally the same as in previous pages. The `relu2deriv` function returns 1 when `output` > 0; otherwise, it returns 0. This is the *slope* (the *derivative*) of the `relu` function. It serves an important purpose, as you'll see in a moment.

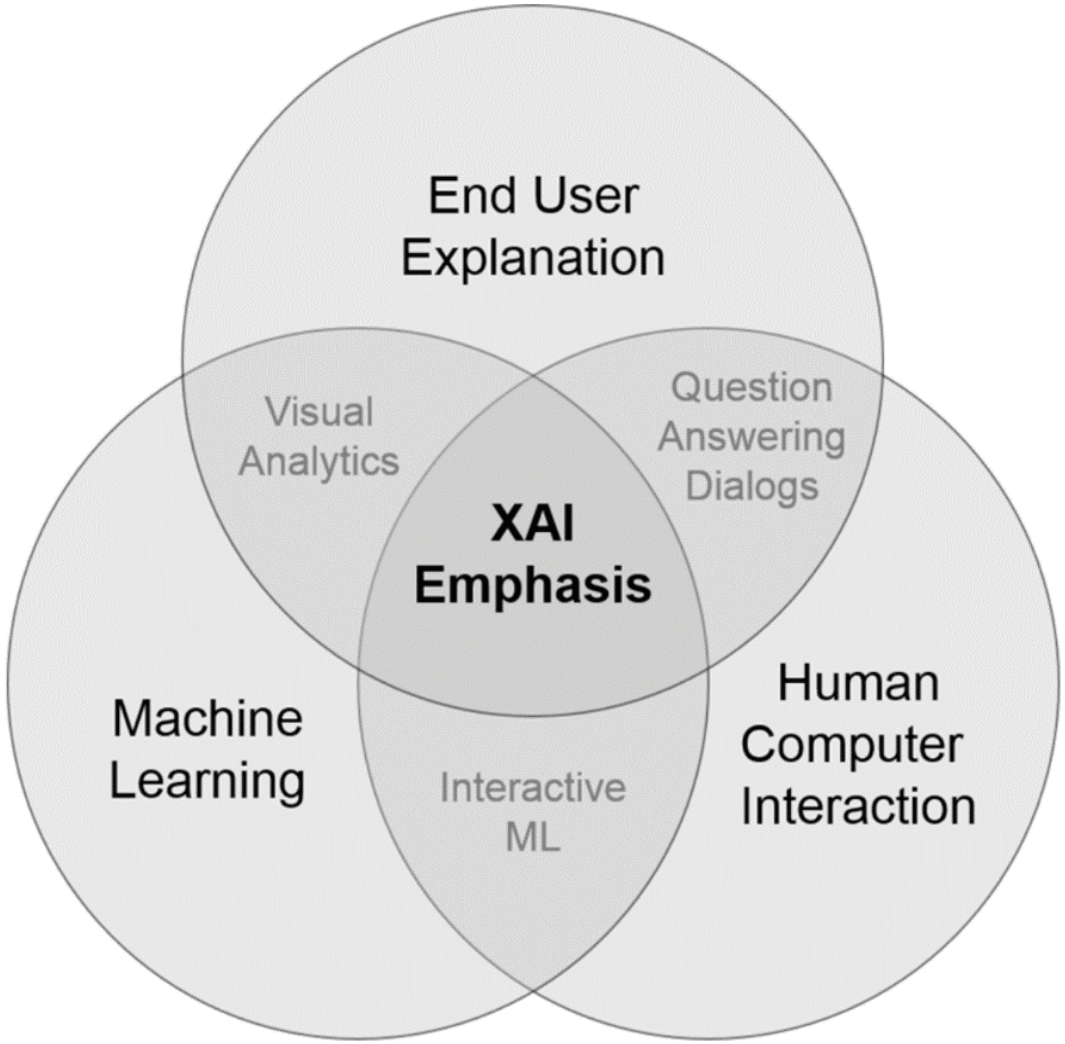






HOGESCHOOL ROTTERDAM



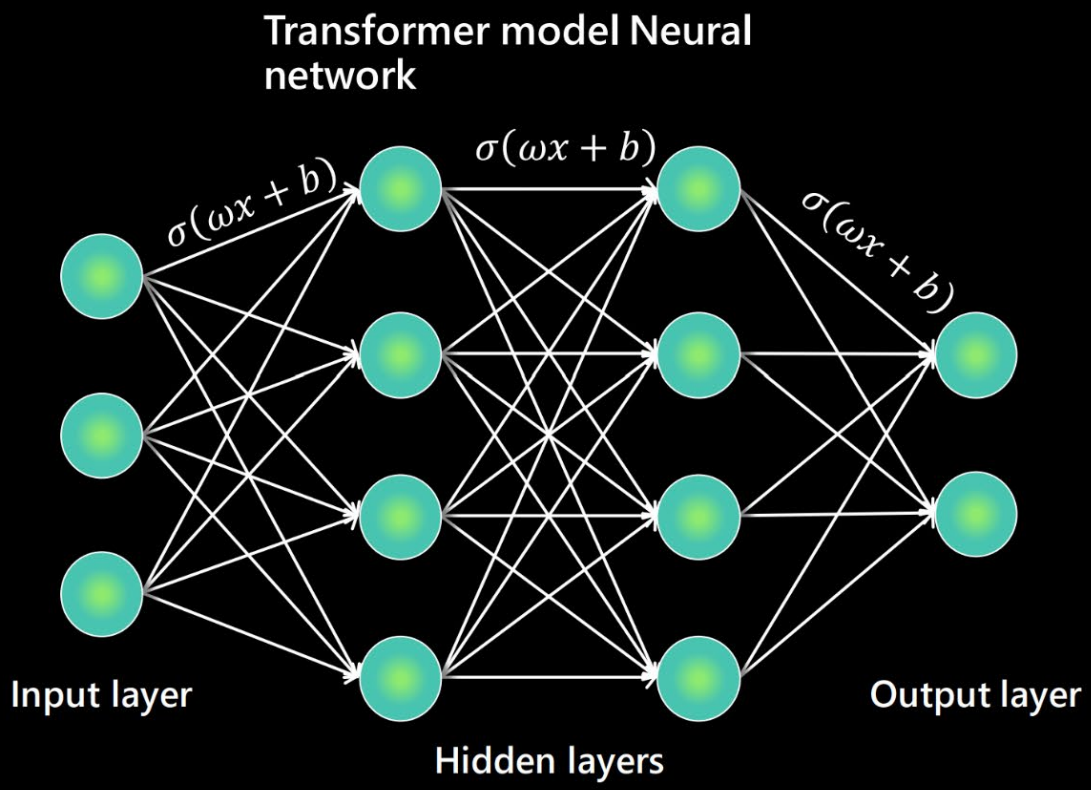


HOGESCHOOL ROTTERDAM

Deep Neural Networks build itself

HOGESCHOOL ROTTERDAM

How large are they?



Function: weight * input plus bias

BERT Large - 2018
345M

GPT2 - 2019
1.5B

GPT3 - 2020
175B

Turing Megatron NLG 2021
530B

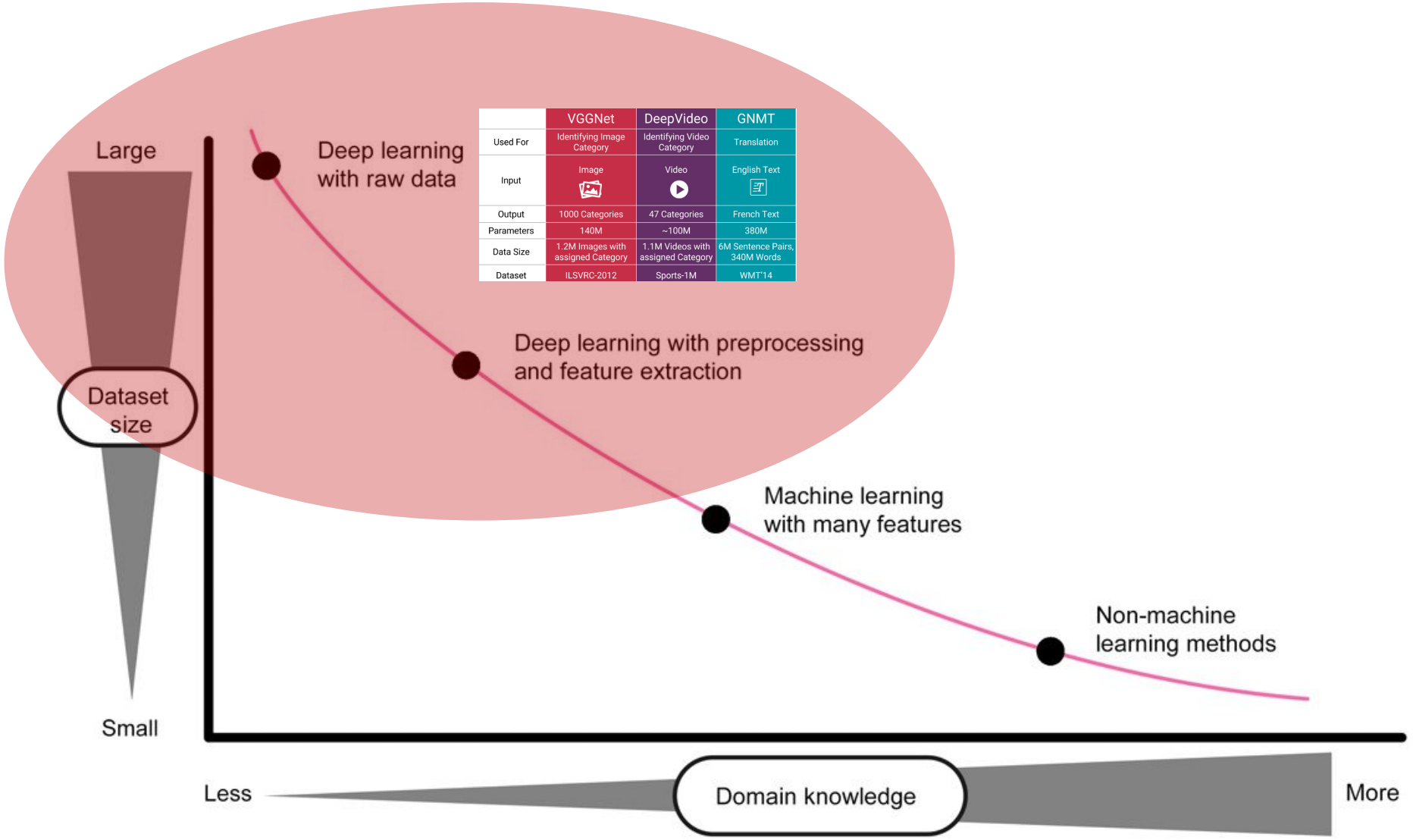
GPT4 – 2023
1.4T (estimated)

{Big-data}

Big-data is needed to avoid hand-crafted feature extraction

HOGESCHOOL ROTTERDAM

CREATING



	VGGNet	DeepVideo	GNMT
Used For	Identifying Image Category	Identifying Video Category	Translation
Input	Image	Video	English Text
Output	1000 Categories	47 Categories	French Text
Parameters	140M	~100M	380M
Data Size	1.2M Images with assigned Category	1.1M Videos with assigned Category	6M Sentence Pairs, 340M Words
Dataset	ILSVRC-2012	Sports-1M	WMT'14

A Unified Approach to Interpreting Model Predictions

Scott M. Lundberg
Paul G. Allen School of Computer Science
University of Washington
Seattle, WA 98105
slund1@cs.washington.edu

Su-In Lee
Paul G. Allen School of Computer Science
Department of Genome Sciences
University of Washington
Seattle, WA 98105
suinlee@cs.washington.edu

Abstract

Understanding why a model makes a certain prediction can be as crucial as the prediction's accuracy in many applications. However, the highest accuracy for large modern datasets is often achieved by complex models that even experts struggle to interpret, such as ensemble or deep learning models, creating a tension between *accuracy* and *interpretability*. In response, various methods have recently been proposed to help users interpret the predictions of complex models, but it is often unclear how these methods are related and when one method is preferable over another. To address this problem, we present a unified framework for interpreting predictions, SHAP (SHapley Additive exPlanations). SHAP assigns each feature an importance value for a particular prediction. Its novel components include: (1) the identification of a new class of additive feature importance measures, and (2) theoretical results showing there is a unique solution in this class with a set of desirable properties. The new class unifies six existing methods, notable because several recent methods in the class lack the proposed desirable properties. Based on insights from this unification, we present new methods that show improved computational performance and/or better consistency with human intuition than previous approaches.

1 Introduction

The ability to correctly interpret a prediction model's output is extremely important. It engenders appropriate user trust, provides insight into how a model may be improved, and supports understanding of the process being modeled. In some applications, simple models (e.g., linear models) are often preferred for their ease of interpretation, even if they may be less accurate than complex ones. However, the growing availability of big data has increased the benefits of using complex models, so bringing to the forefront the trade-off between accuracy and interpretability of a model's output. A wide variety of different methods have been recently proposed to address this issue [5, 8, 9, 3, 4, 1]. But an understanding of how these methods relate and when one method is preferable to another is still lacking.

Here, we present a novel unified approach to interpreting model predictions.[1] Our approach leads to three potentially surprising results that bring clarity to the growing space of methods:

1. We introduce the perspective of viewing *any* explanation of a model's prediction as a model itself, which we term the *explanation model*. This lets us define the class of *additive feature attribution methods* (Section 2), which unifies six current methods.

[1] https://github.com/slundberg/shap

31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

https://proceedings.neurips.cc/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf

{Top-down}

Top-down Encoding Capacity increases by adding hidden layers

HOGESCHOOL ROTTERDAM

What are the limits of deep learning?

The much-ballyhooed artificial intelligence approach boasts impressive feats but still falls short of human brainpower. Researchers are determined to figure out what's missing.

M. Mitchell Waldrop, *Science Writer*

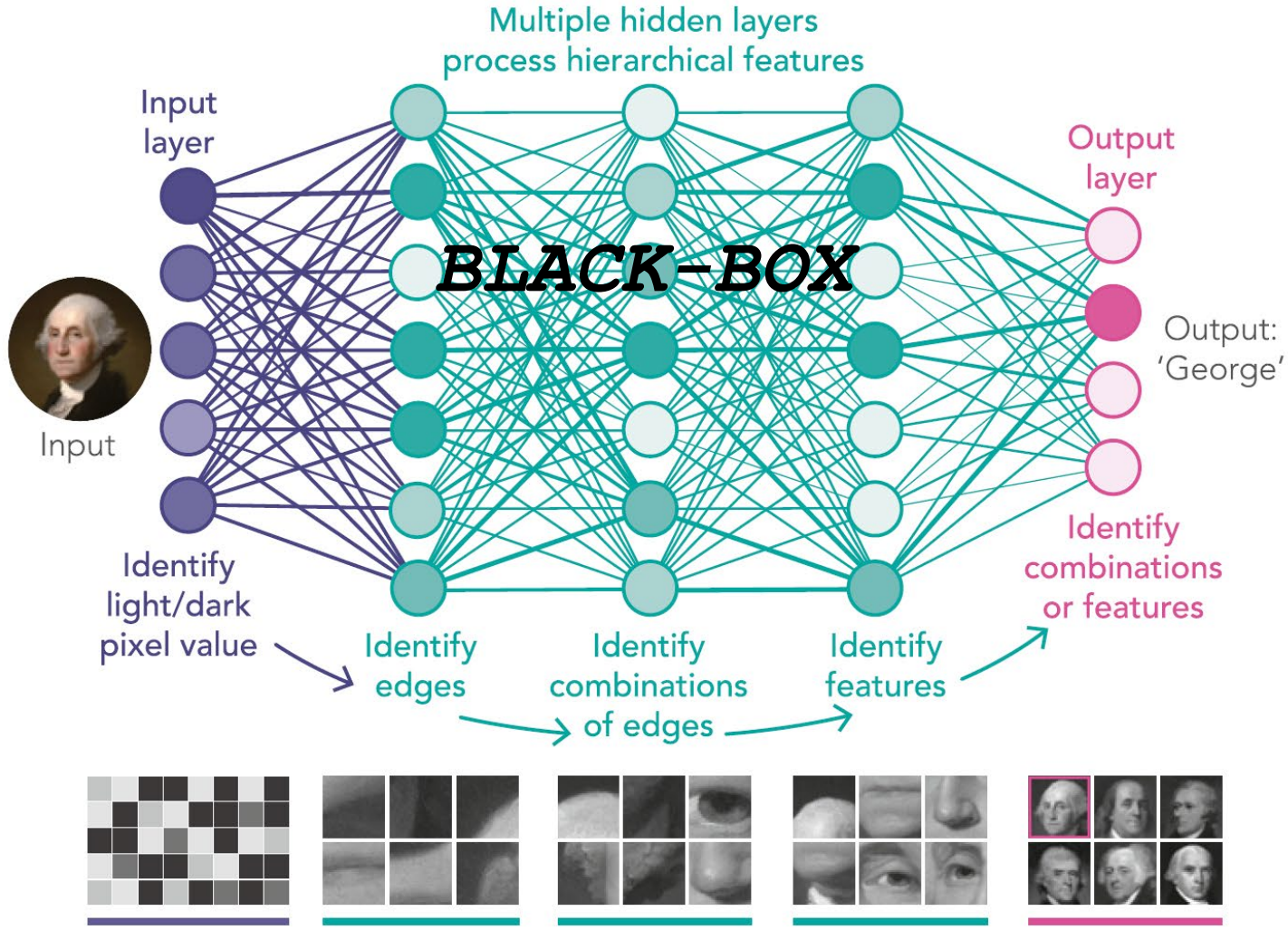
There's no mistaking the image: It's a banana—a big, ripe, bright-yellow banana. Yet the artificial intelligence (AI) identifies it as a toaster, even though it was trained with the same powerful and oft-publicized deep-learning techniques that have produced a white-hot revolution in driverless cars, speech understanding, and a multitude of other AI applications. That means the AI was shown several thousand photos of bananas, slugs, snails, and similar-looking objects, like so many flash cards, and then drilled on the answers until it had the classification down cold. And yet this advanced system was quite easily confused—all it took was a little day-glow sticker, digitally pasted in one corner of the image.

This example of what deep-learning researchers call an "adversarial attack," discovered by the Google Brain team in Mountain View, CA (1), highlights just how far AI still has to go before it remotely approaches human capabilities. "I initially thought that adversarial examples were just an annoyance," says Geoffrey Hinton, a computer scientist at the University of Toronto and one of the pioneers of deep learning. "But I now think they're probably quite profound. They tell us that we're doing something wrong."

That's a widely shared sentiment among AI practitioners, any of whom can easily rattle off a long list of deep learning's drawbacks. In addition to its vulnerability

Apparent shortcomings in deep-learning approaches have raised concerns among researchers and the general public as technologies such as driverless cars, which use deep-learning techniques to navigate, get involved in well-publicized mishaps. Image credit: Shutterstock.com/MONOPOLY919.

www.pnas.org/cgi/doi/10.1073/pnas.1821594116

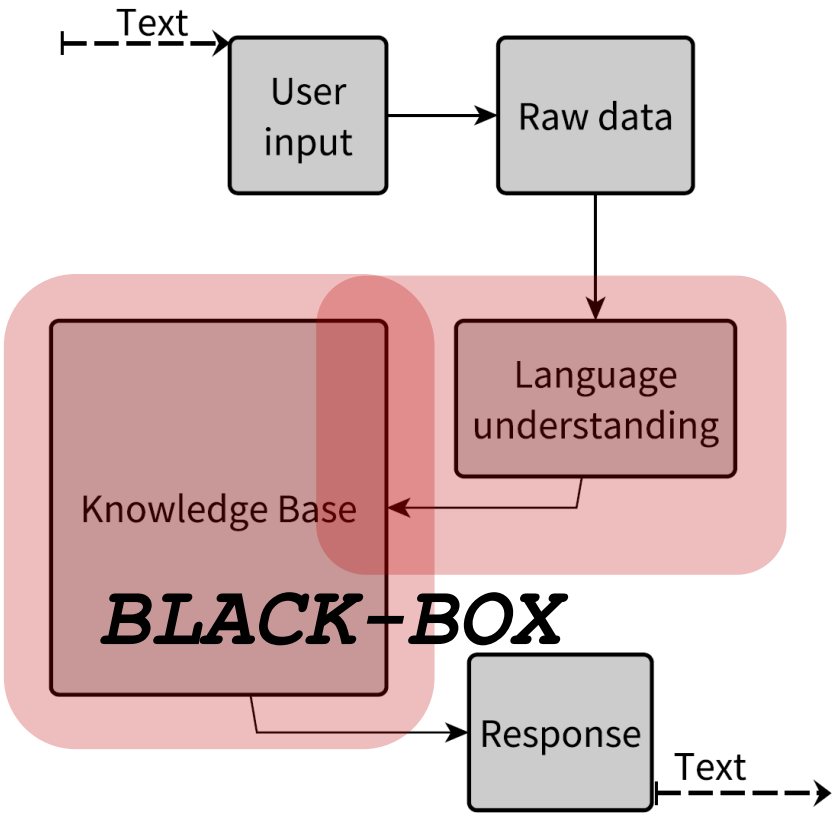


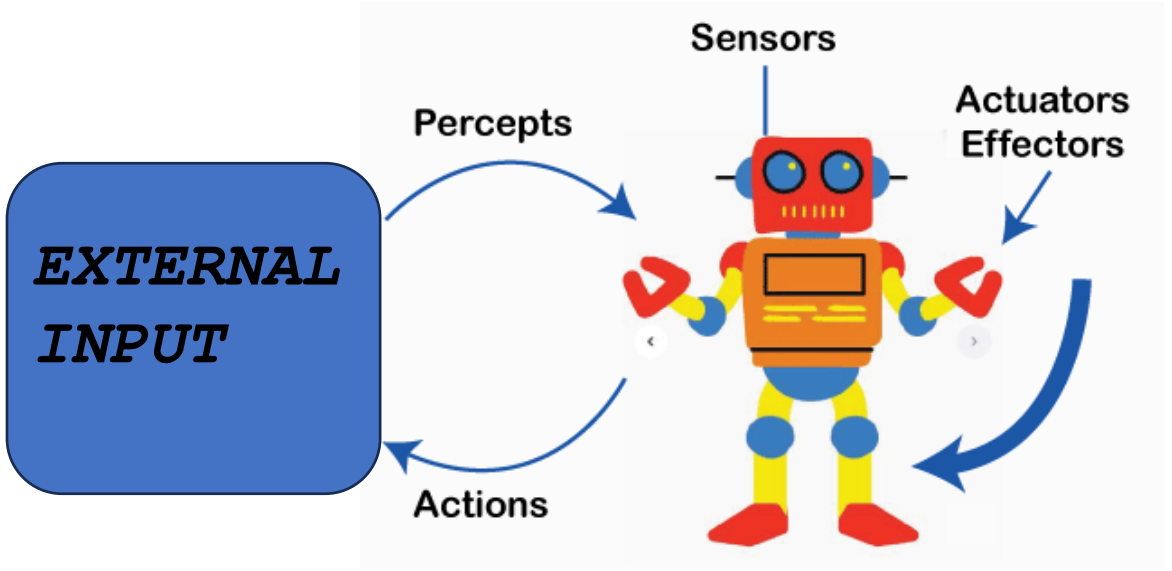
HOGESCHOOL ROTTERDAM

MACHINE LEARNING data-flow

DEEP LEARNING data-flow

ChatGPT is een Conversationele tekst-in/tekst-uit ChatBot



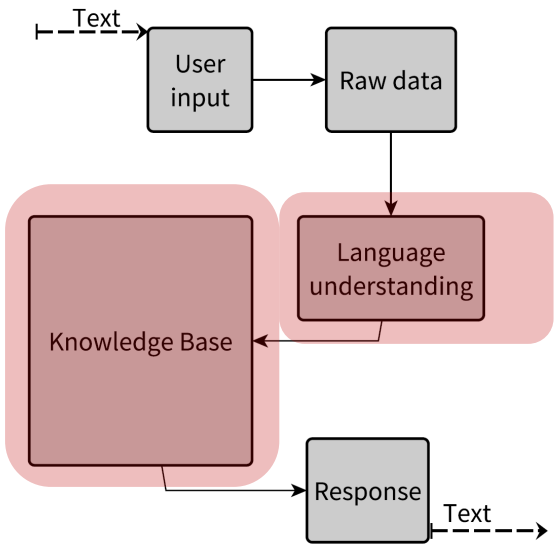


ROBOTIC AI-AGENT

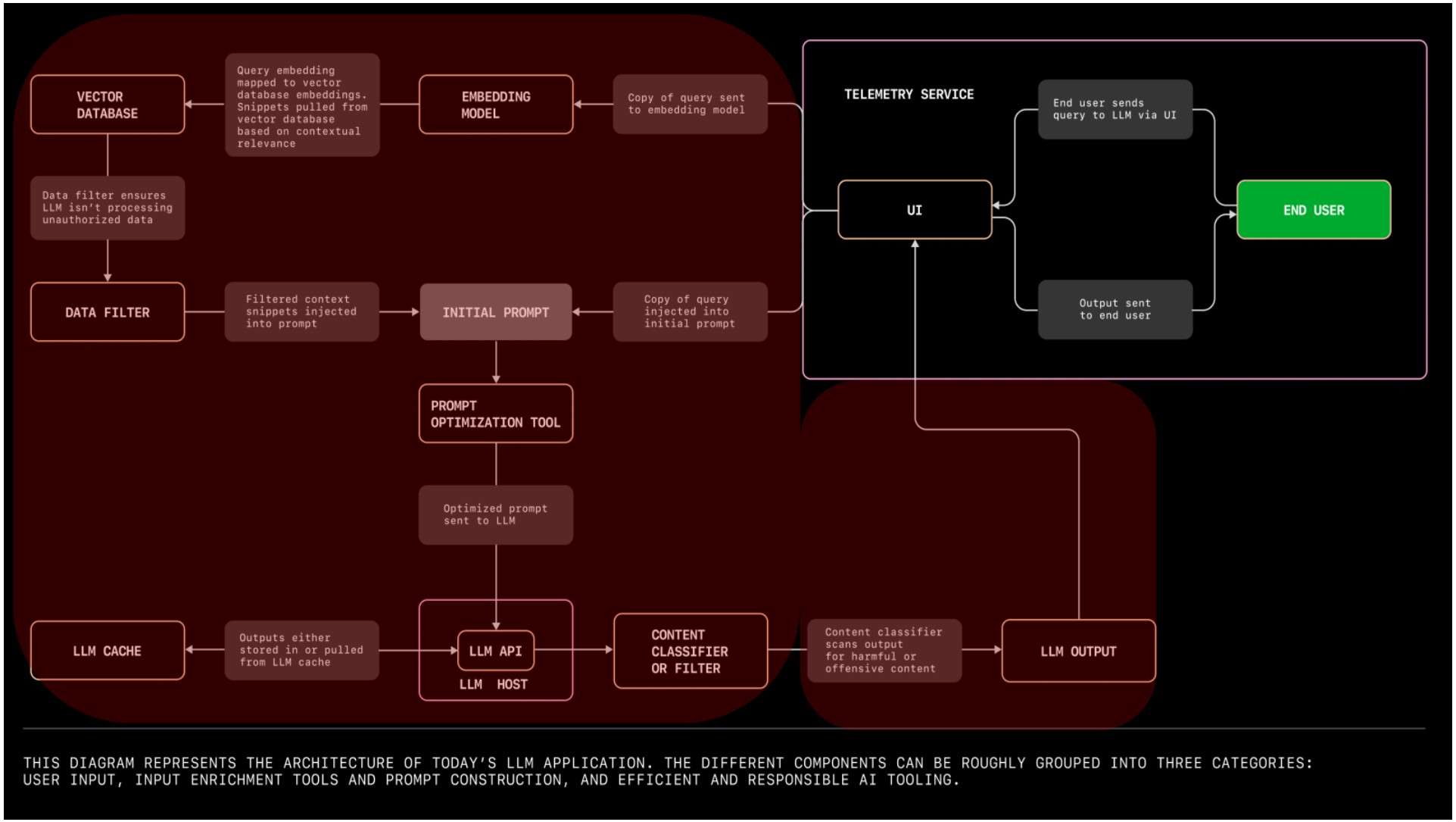
Sánchez-Díaz, X., Ayala-Bastidas, G., Fonseca-Ortiz, P., & Garrido, L. (2018). A knowledge-based methodology for building a conversational chatbot as an intelligent tutor. https://doi.org/10.1007/978-3-030-04497-8_14

ChatGPT == 99% BLACKBOX + 1% user-interface

HOGESCHOOL ROTTERDAM



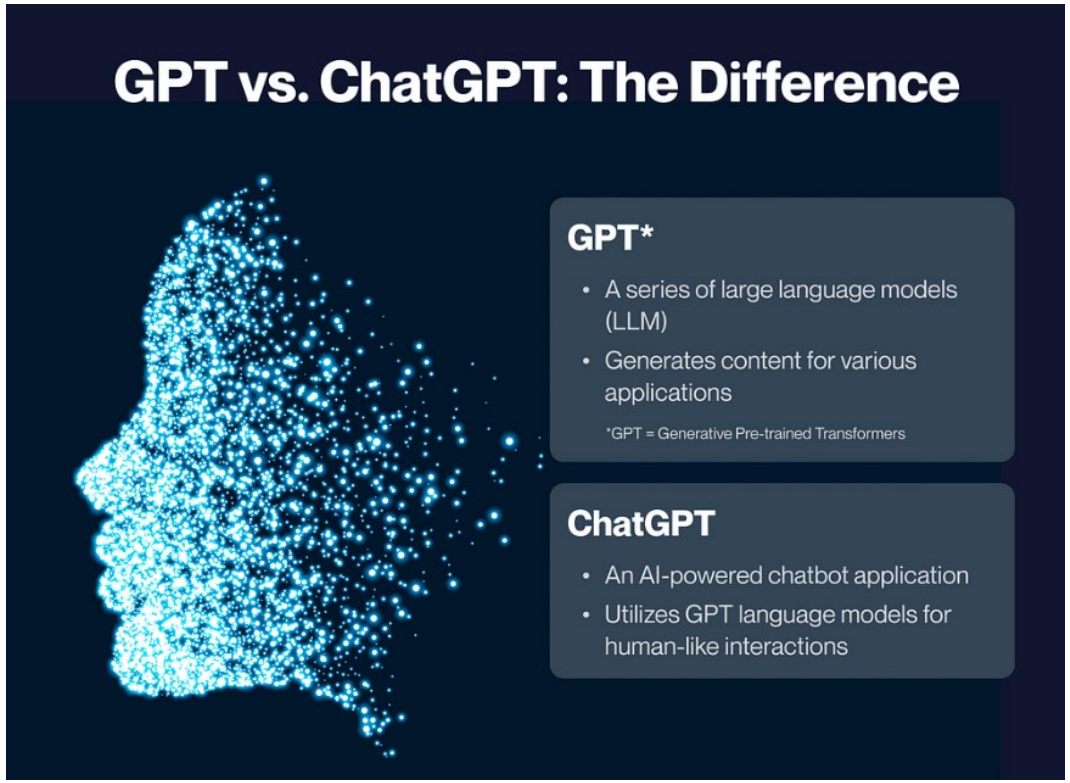
BLACK-BOX

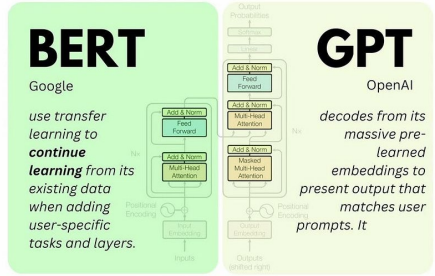


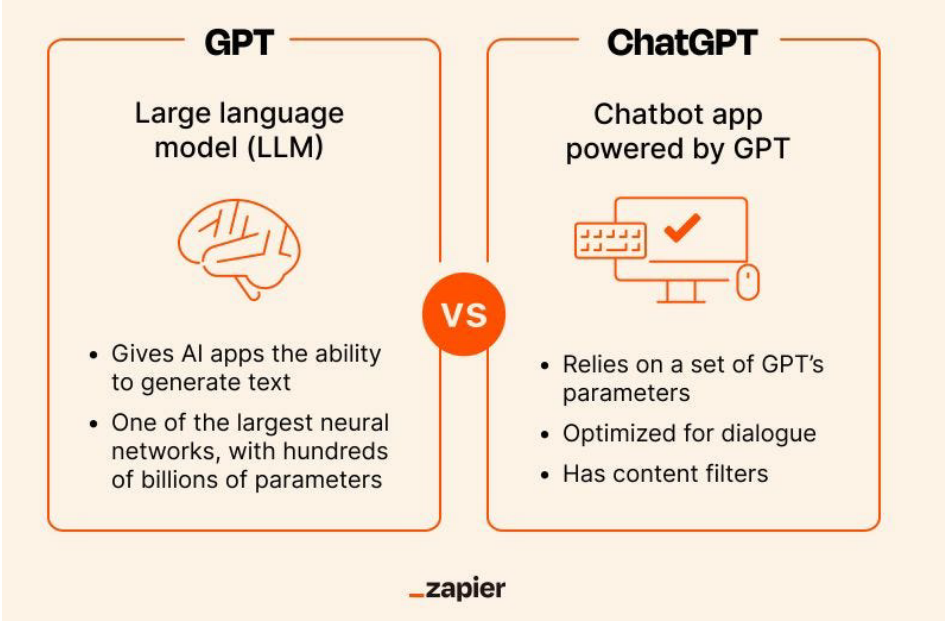
https://blogs.novita.ai/exploring-architectural-structures-and-functional-capacities-of-llms/

GPT {LLM} versus ChatGPT

HOGESCHOOL ROTTERDAM

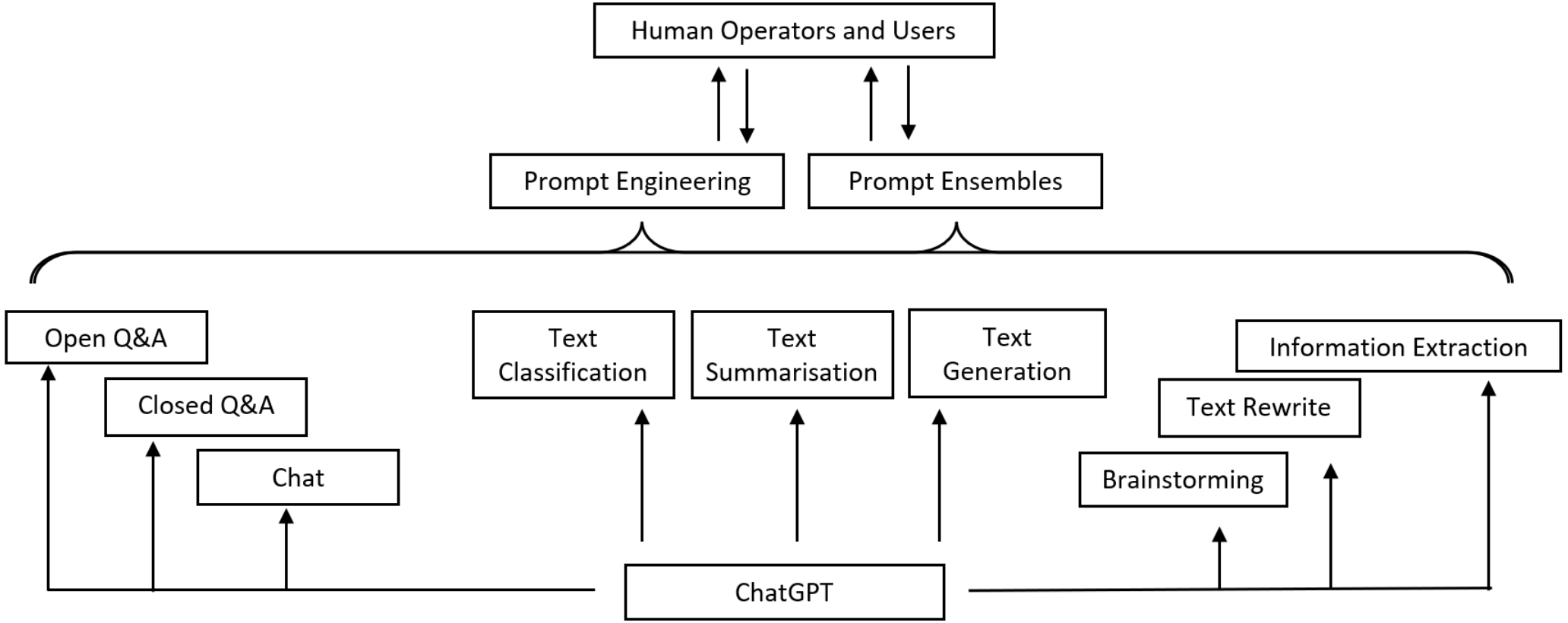






https://blogs.novita.ai/what-is-the-difference-between-llm-and-gpt/

ChatBot Use-Cases



Conferences > 2023 IEEE International Confe...

ChatGPT and Generative AI Guidelines for Addressing Academic Integrity and Augmenting Pre-Existing Chatbots

Publisher: IEEE Cite This PDF

Daswin De Silva ; Nishan Mills ; Mona El-Ayoubi ; Milos Manic ; Damminda Alahakoon All Authors

635 Full Text Views

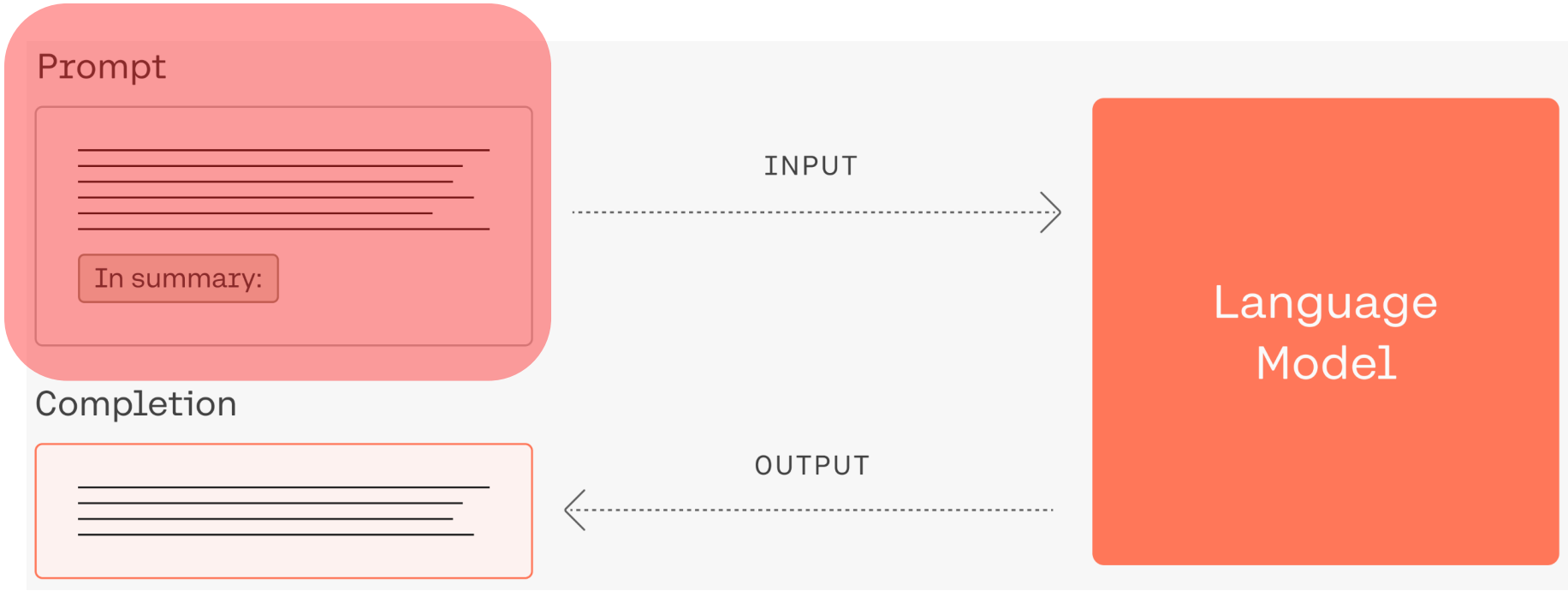
generieke train-dataset
machinaal leren
Black-Box
GEEN Human-in-the-loop
multimodaliteit
Commerciële belangen

creëert risico's
- veiligheid
- Privacy
- betrouwbaarheid &
- reproduceerbaarheid

Hoe kun je een chatbot Dialoog effectief & unbiased sturen

Met behulp van
prompt engineering
kun je de output van LLMs
Betrouwbaarder maken

Conversationele AI-agenten worden aangestuurd via "prompts"



https://docs.cohere.com/docs/prompt-engineering

https://docs.cohere.com/docs/introduction-to-large-language-models

PROMPT-ENGINEERING

Het creatieve proces van het schrijven van een effectief ***prompt-recept*** wordt in het Engels ***"prompt engineering"*** genoemd.

Het schrijven van prompt-recepten
---pseudo-Code---
is een talige manier van het programmeren van "bevroren" voorgetraind taalmodellen.

Prompt Taxonomie

openeinde spectrum

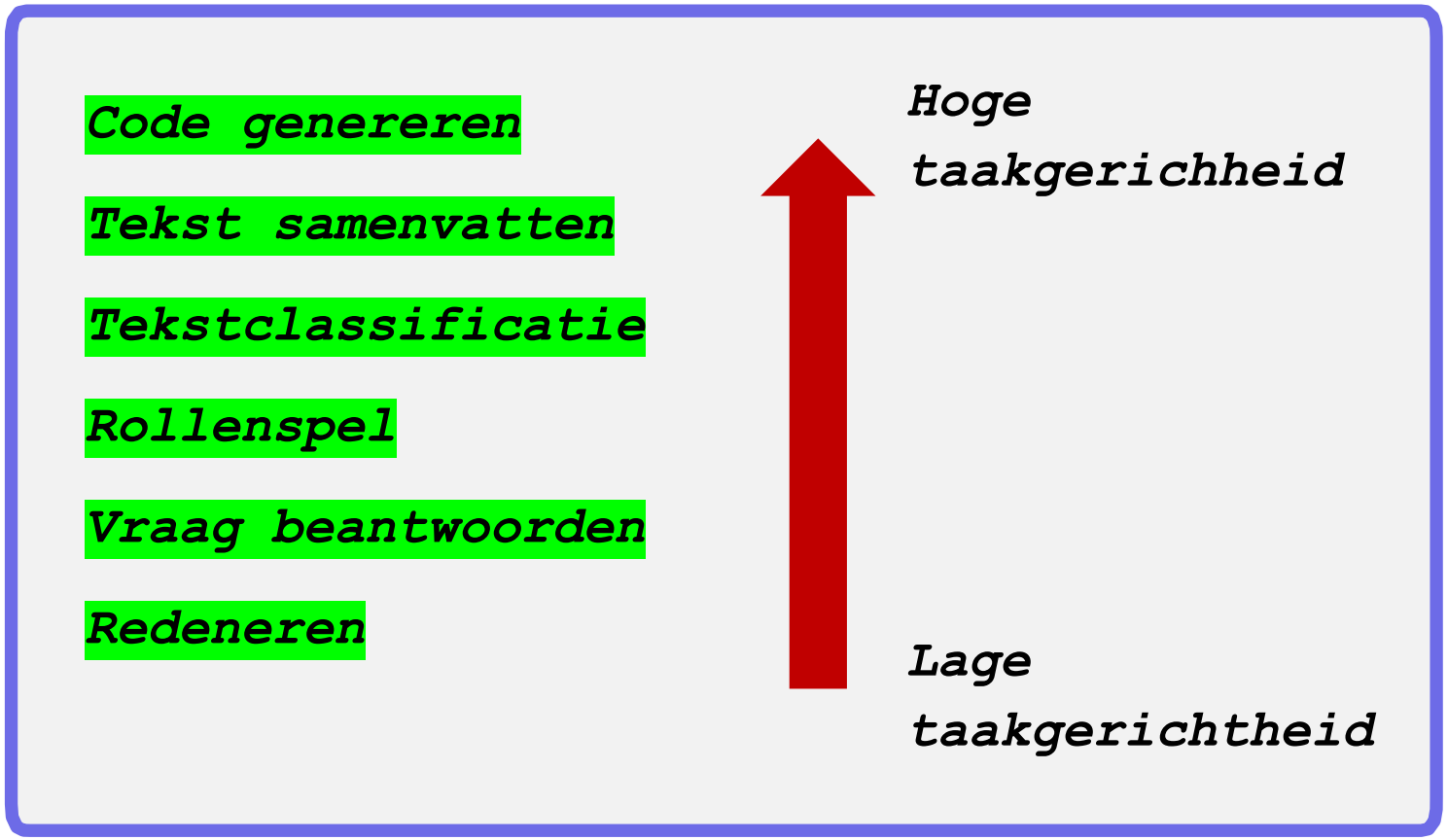
Uitkomst vertoont weinig variatie

PROMPT is COMPLEX

Uitkomst is onzeker vertoont zeer veel variatie

PROMPT is SIMPEL

Taakgerichtheid van prompts



Wat is het belang van Prompt Recepten Schrijven?

Sturen van de mate van taakgerichtheid door reduceren van variatie in het antwoord zodat de kans groter wordt dat de uitkomst correct is.

Prompt Recept Structuur

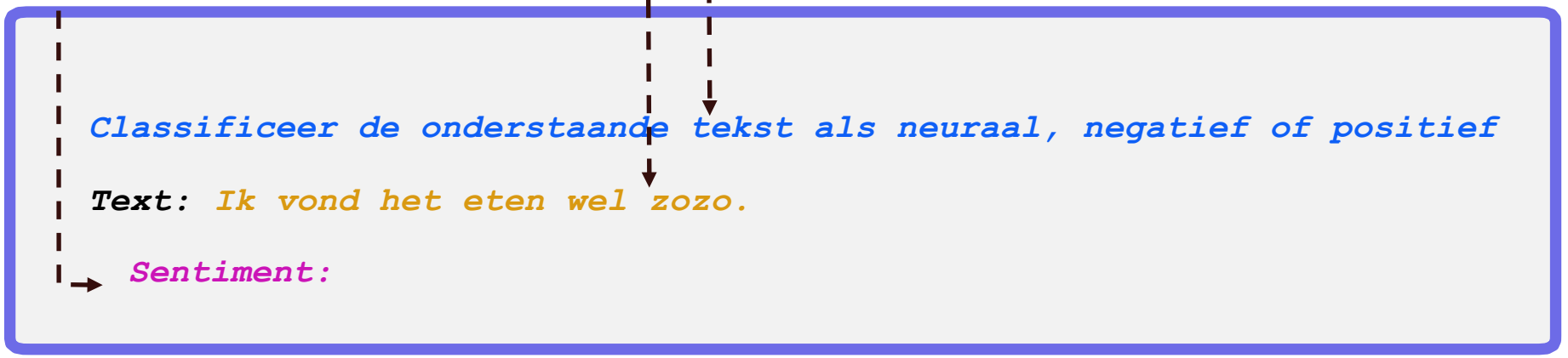
Een prompt is opgebouwd uit de volgende elementen:

Instructie(s)

Context

Invoergegevens

Uitvoer-indicator



Prompt recept ontwerptechnieken gebaseerd op fine-tuning van het onderliggende taal-model

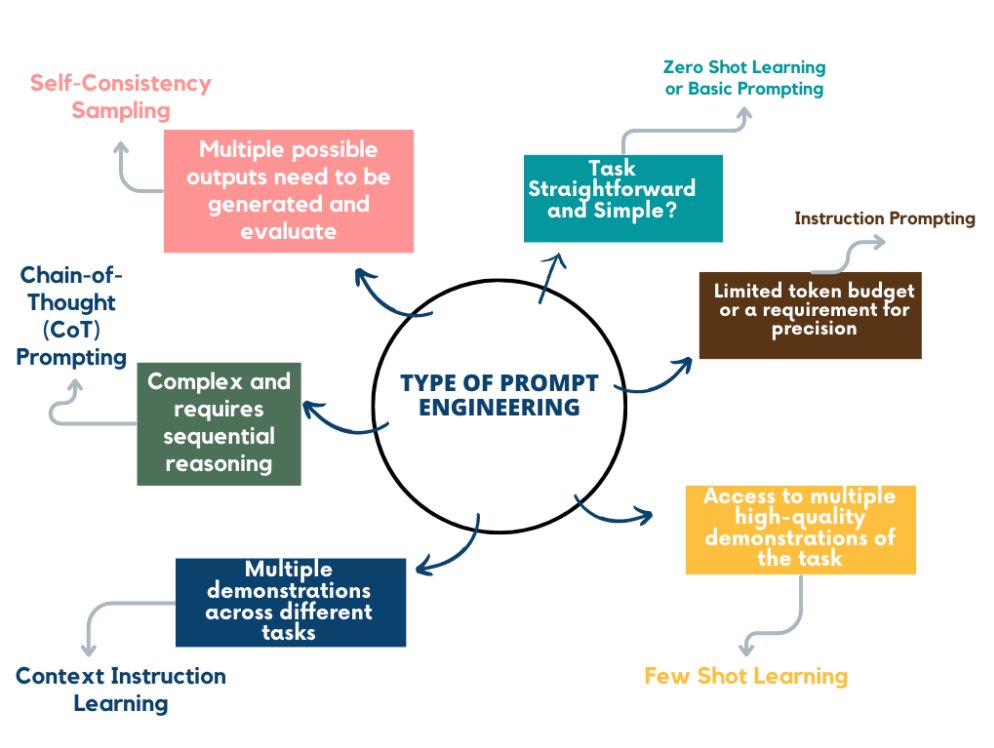
Few-shot prompts (In Context Leren)

Chain-of-thought (CoT) reasoning

Self-Consistency Sampling

Knowledge Generation Prompting

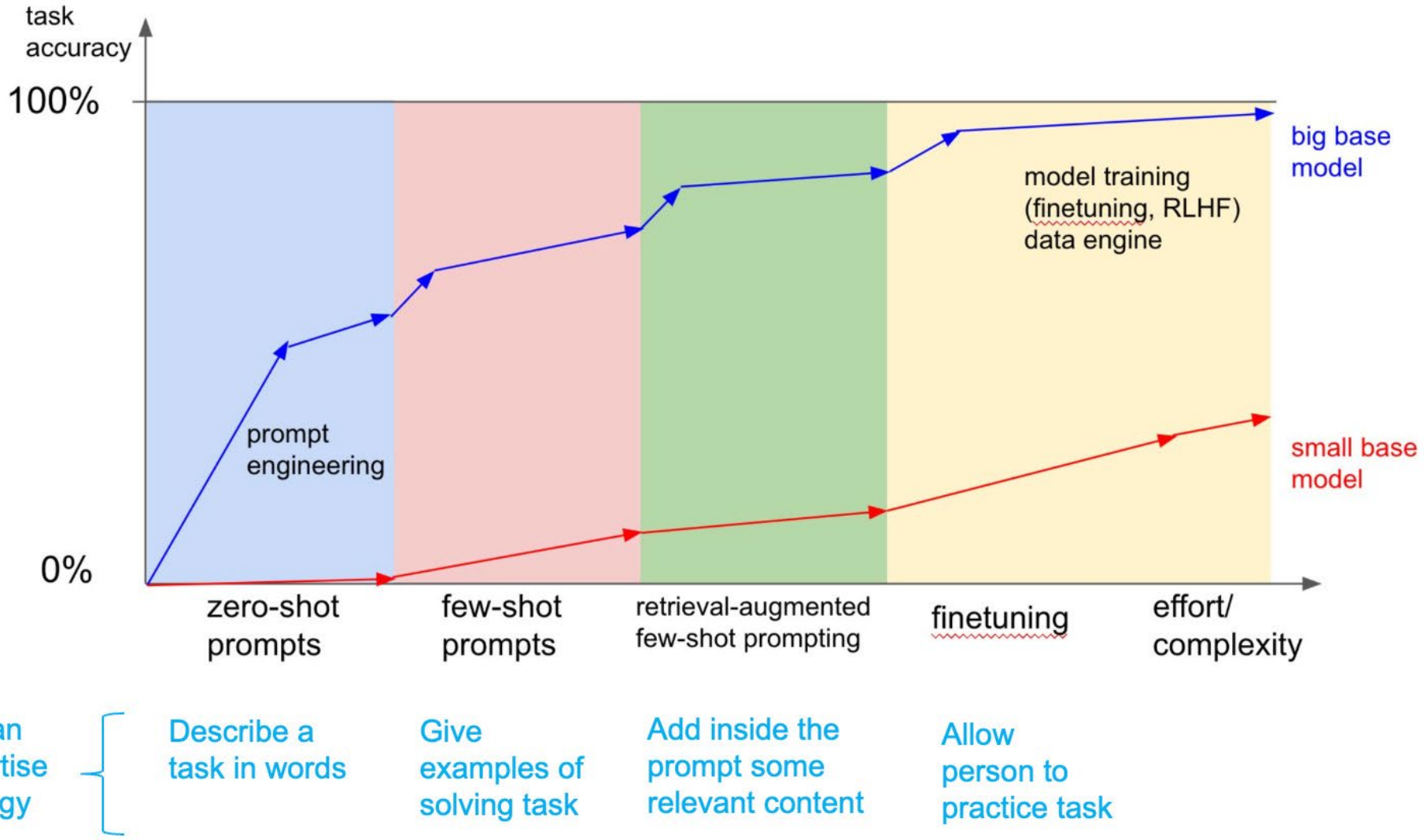
ReAct



https://tinyml.substack.com/p/your-go-to-guide-to-master-prompt

HOGESCHOOL ROTTERDAM

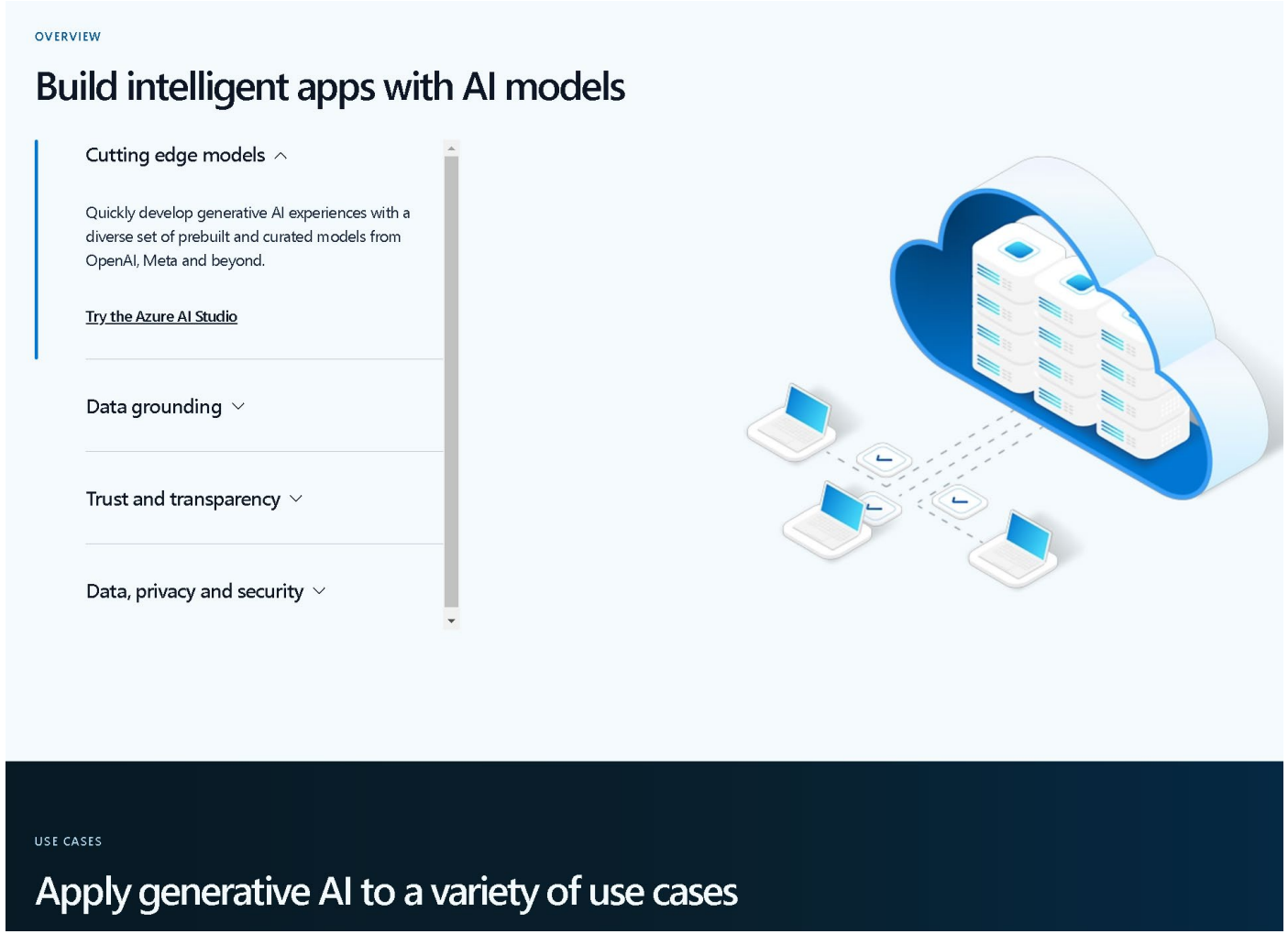
Building a PoC dialogue



https://www.anti-vc.com/p/fine-tuning-llms-learnings-from-the

Hoe bouw en test je veilig Talige Generatieve AI-technologie?

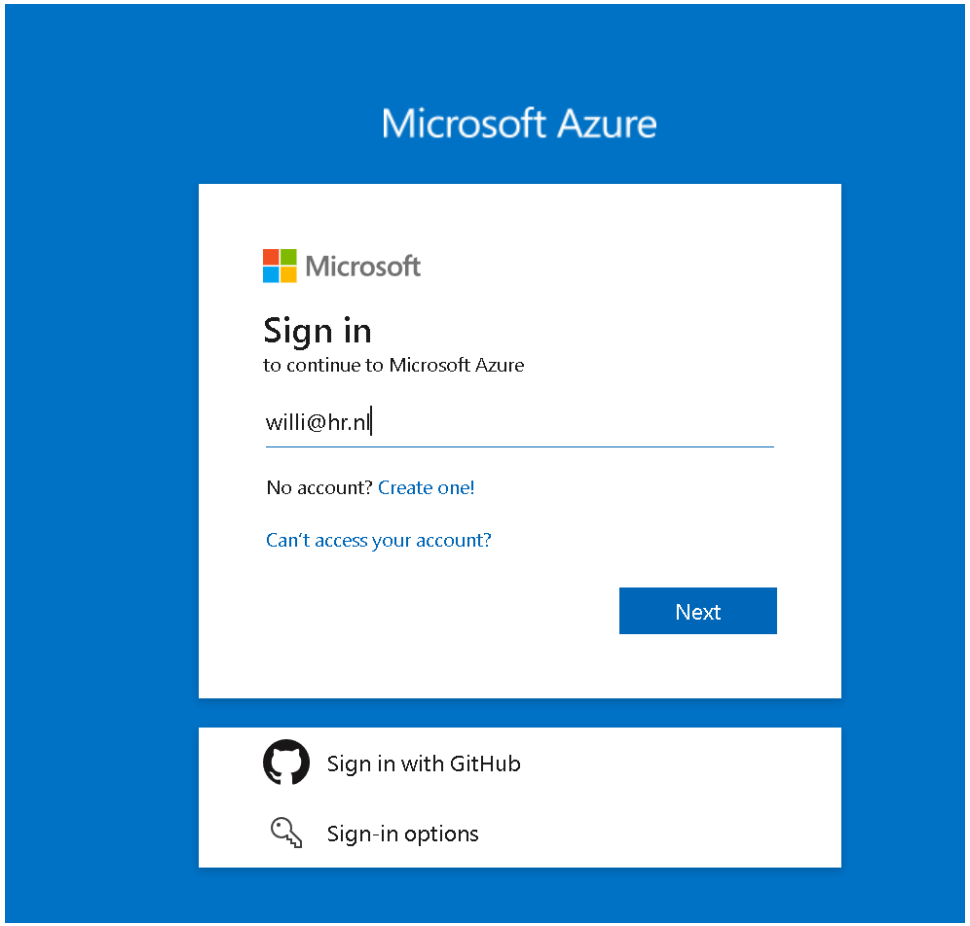
Azure OpenAI service



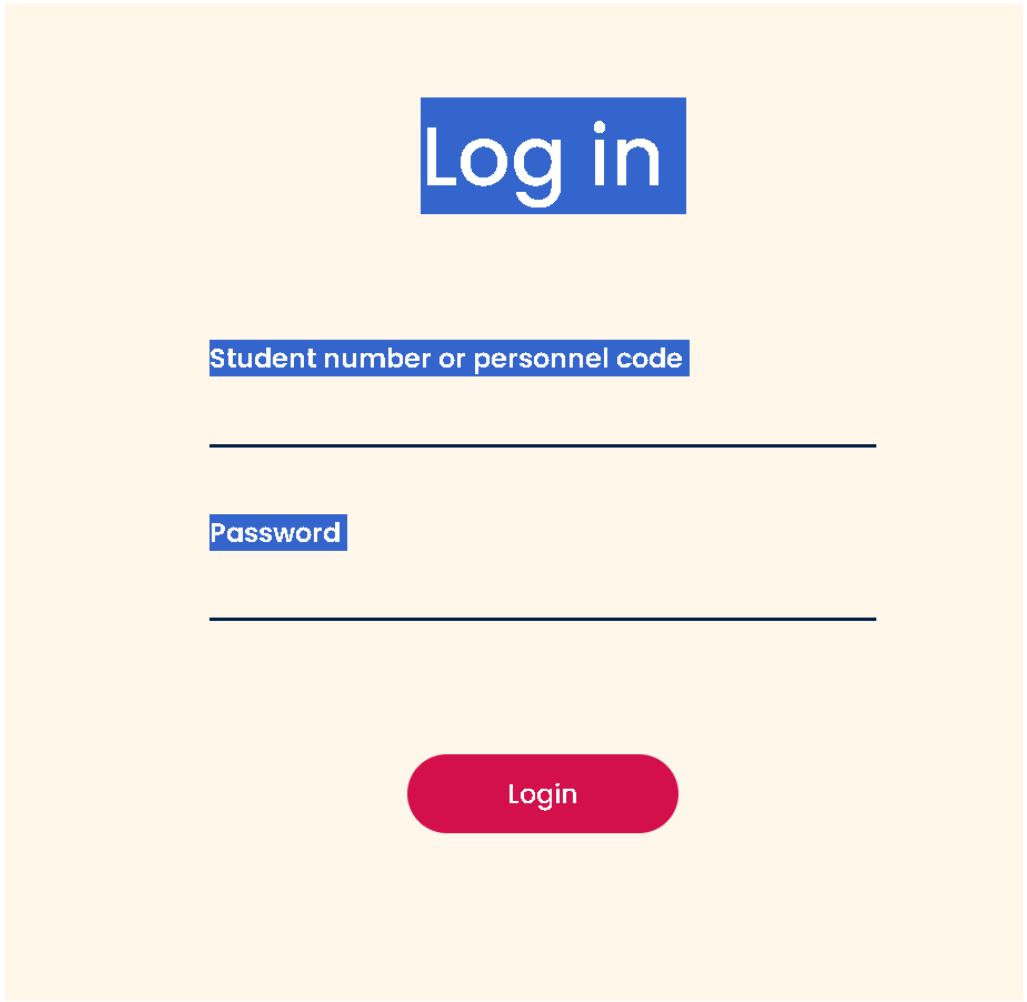
https://azure.microsoft.com/en-us/products/ai-services/openai-service

HOGESCHOOL ROTTERDAM

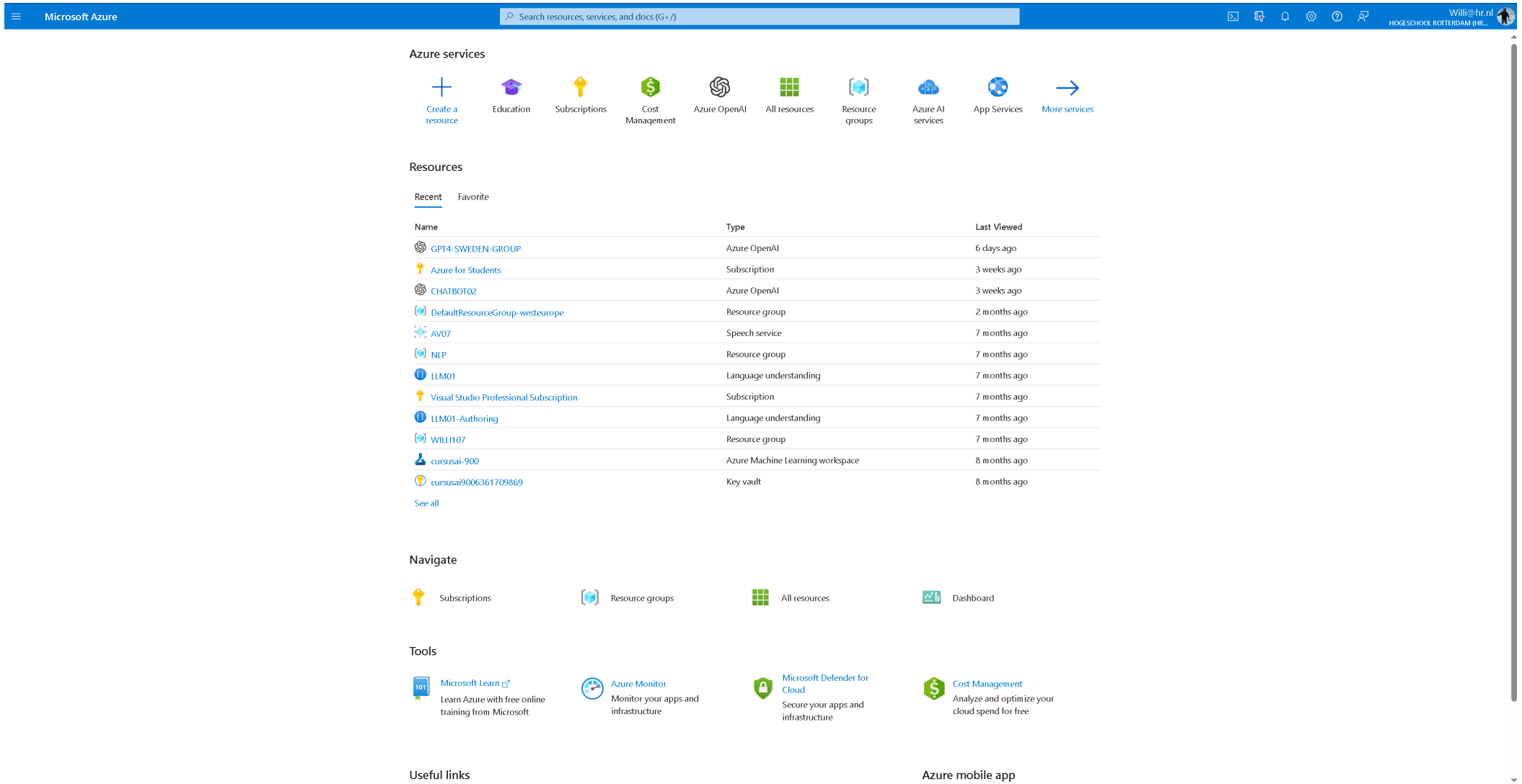
portal.azure.com



HOGESCHOOL ROTTERDAM



HOGESCHOOL ROTTERDAM



HOGESCHOOL ROTTERDAM

Home >

Create Azure OpenAI ...

Azure services

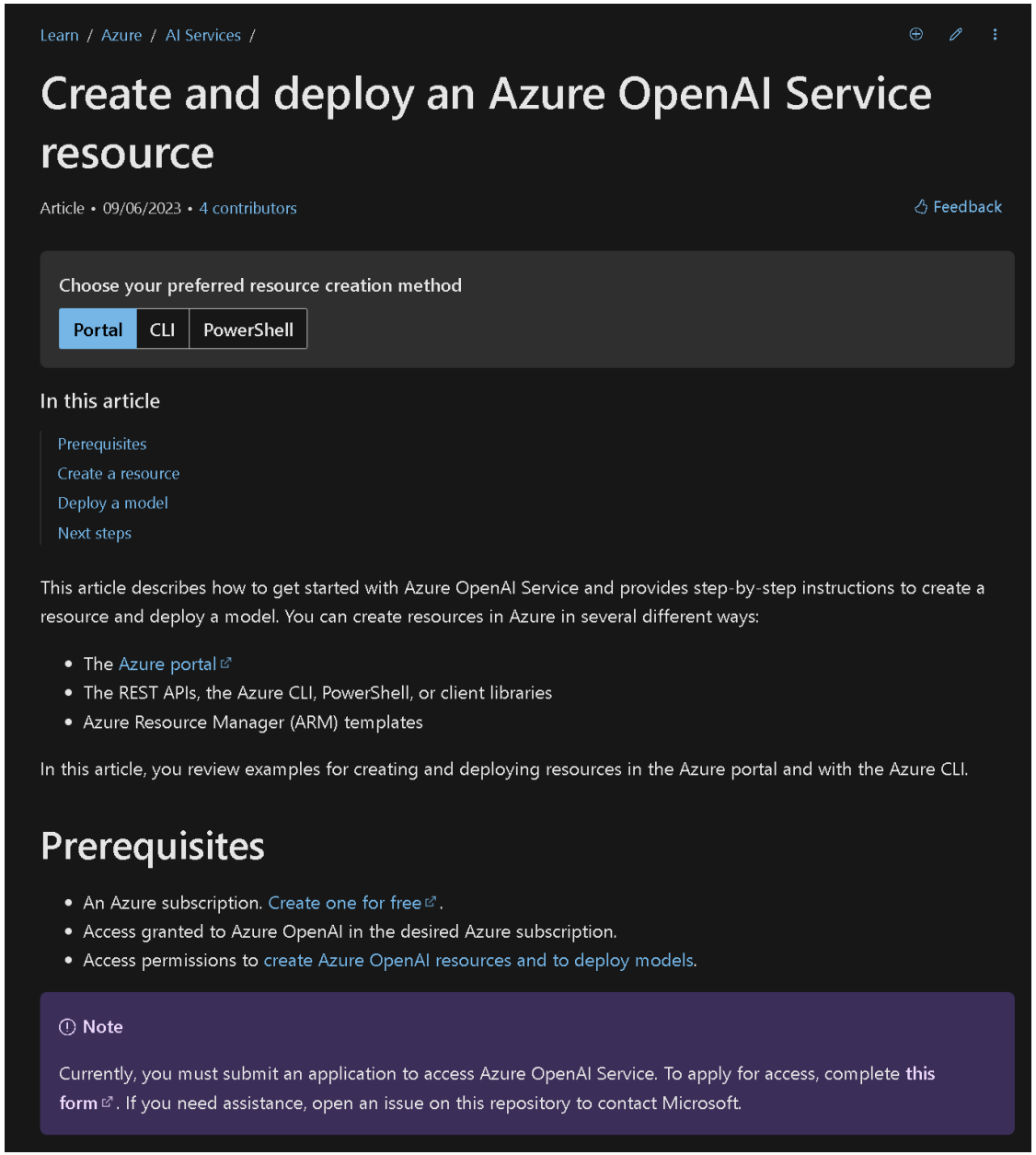
Resources

Recent | Favorite

Name	Type	Last Viewed
GPT4-SWEDEN-GROUP	Azure OpenAI	8 minutes ago
cursusai9006361709869	Key vault	9 minutes ago
Azure for Students	Subscription	9 minutes ago
Visual Studio Professional Subscription	Subscription	10 minutes ago
NLP	Resource group	11 minutes ago
CHATBOT02	Azure OpenAI	3 weeks ago
DefaultResourceGroup-westeurope	Resource group	2 months ago
AV07	Speech service	7 months ago
LLM01	Language understanding	7 months ago
LLM01-Authoring	Language understanding	7 months ago
WILLI107	Resource group	7 months ago
cursusai-900	Azure Machine Learning workspace	8 months ago

See all

Review the Azure OpenAI code of conduct



Learn / Azure / AI Services /

Create and deploy an Azure OpenAI Service resource

Article • 09/06/2023 • 4 contributors

Feedback

Choose your preferred resource creation method

Portal | CLI | PowerShell

In this article

This article describes how to get started with Azure OpenAI Service and provides step-by-step instructions to create a resource and deploy a model. You can create resources in Azure in several different ways:

- The Azure portal
- The REST APIs, the Azure CLI, PowerShell, or client libraries
- Azure Resource Manager (ARM) templates

In this article, you review examples for creating and deploying resources in the Azure portal and with the Azure CLI.

Prerequisites

- An Azure subscription. Create one for free.
- Access granted to Azure OpenAI in the desired Azure subscription.
- Access permissions to create Azure OpenAI resources and to deploy models.

Note

Currently, you must submit an application to access Azure OpenAI Service. To apply for access, complete this form. If you need assistance, open an issue on this repository to contact Microsoft.

https://learn.microsoft.com/en-us/azure/ai-services/openai/how-to/create-resource?pivots=web-portal

HOGESCHOOL ROTTERDAM

Request Access to Azure OpenAI Service

* Required

Please read all instructions carefully and complete form as instructed

Thank you for your interest in Azure OpenAI Service. **Please submit this form to register for approval to access and use Azure OpenAI's Limited Access text and code and/or DALL·E 2 text to image models (as indicated in the form). All use cases must be registered.** Azure OpenAI Service requires registration and is currently only available to approved enterprise customers and partners. Learn more about limited access to Azure OpenAI Service here.

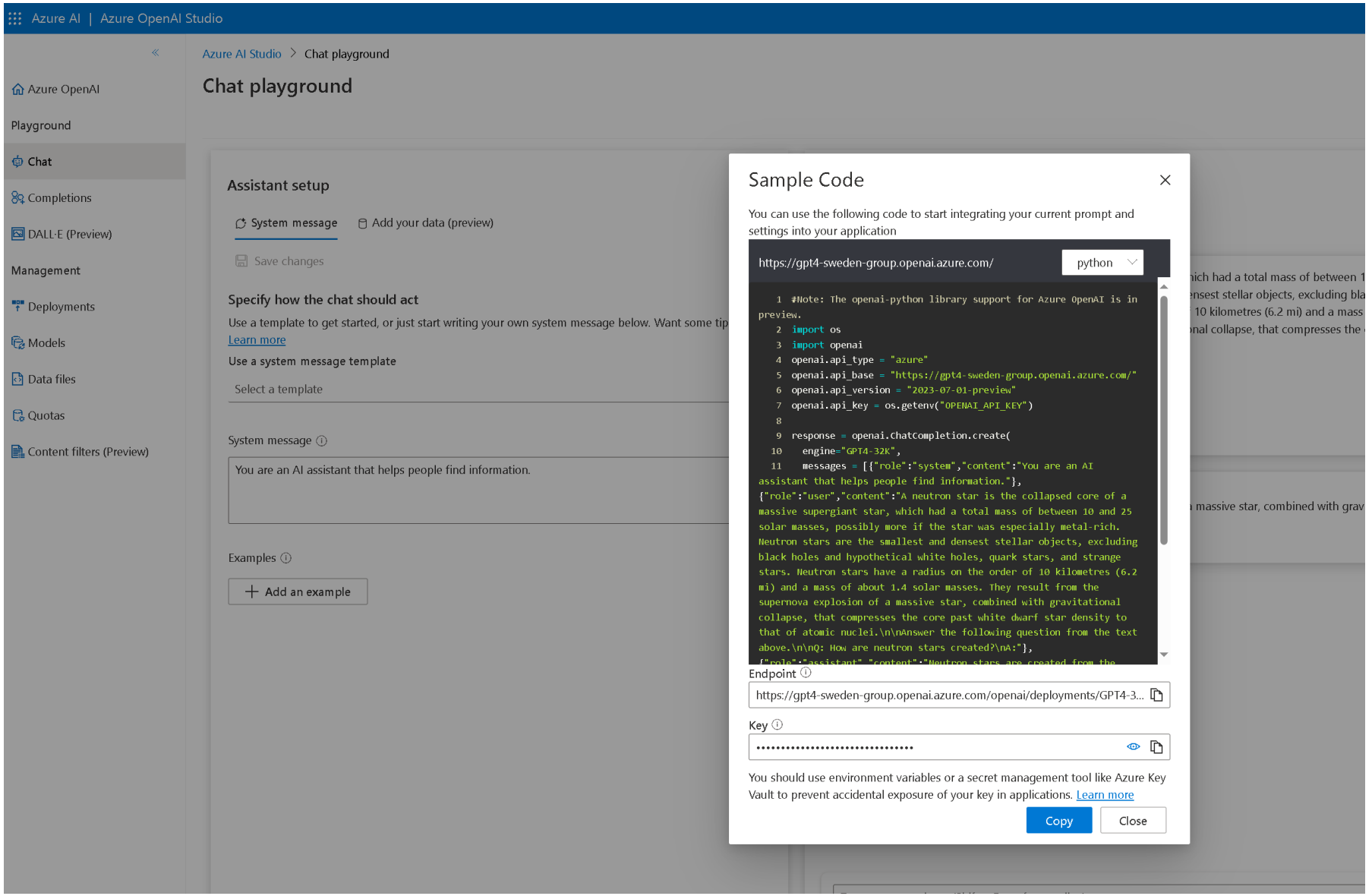
Limited access scenarios: When evaluating which scenarios to onboard, we consider who will directly interact with the application, who will see the output of the application, whether the application will be used in a high-stakes domain (e.g., medical), and the extent to which the application's capabilities are tightly scoped. In general, applications in high stakes domains will require additional mitigations and are more likely to be approved for applications with internal-only users and internal-only audiences. Applications with broad possible uses, including content generation capabilities, are more likely to be approved if 1) the domain is not high stakes and users are authenticated or 2) in the case of high stakes domains, anyone who views or interacts with the content is internal to your company.

Please be sure to visit the Azure OpenAI Service's transparency note, which provides information and guidelines for responsible use of the service as well as system limitations that may be applicable to your scenario.

If you are **a current Azure OpenAI customer** and would like to add additional use cases, please fill out the Azure OpenAI Additional Use Case form

https://customervoice.microsoft.com/Pages/ResponsePage.aspx?id=v4j5cvGGr0GRqy180BHbR7en2Ais5pxKtso_Pz4b1_xUOFA5Qk1UWDRBMjg0WFhPMklzTzhKQ1dWNyQlQCN0PWcu

HOGESCHOOL ROTTERDAM



HOGESCHOOL ROTTERDAM

Azure chat completions example (preview)

In this example we'll try to go over all operations needed to get chat completions working using the Azure endpoints.
This example focuses on chat completions but also touches on some other operations that are also available using the API. This example is meant to be a quick way of showing simple operations and is not meant as a tutorial.

```python
import os
import openai
openai.api_type = "azure"
openai.api_base = "https://gpt4-sweden-group.openai.azure.com/"
openai.api_version = "2023-07-01-preview"
openai.api_key = "ded218c778894f6da4d3c595c6904194"


#!setx AZURE_OPENAI_KEY "ded218c778894f6da4d3c595c6904194"
#!setx AZURE_OPENAI_ENDPOINT "https://gpt4-sweden-group.openai.azure.com/openai/deployments/GPT4-32K/chat/completions?api-version=2023-07-01-preview"

response = openai.ChatCompletion.create(
  engine="GPT4-32K",
  messages = [
        {"role": "system", "content": "You are a helpful assistant."},
        {"role": "user", "content": "Does Azure OpenAI support customer managed keys?"},
        {"role": "assistant", "content": "Yes, customer managed keys are supported by Azure OpenAI."},
        {"role": "user", "content": "Do other Azure AI services support this too?"}
  ],
  temperature=0.7,
  max_tokens=800,
  top_p=0.95,
  frequency_penalty=0,
  presence_penalty=0,
  stop=None)


print(response)
print(response['choices'][0]['message']['content'])

```

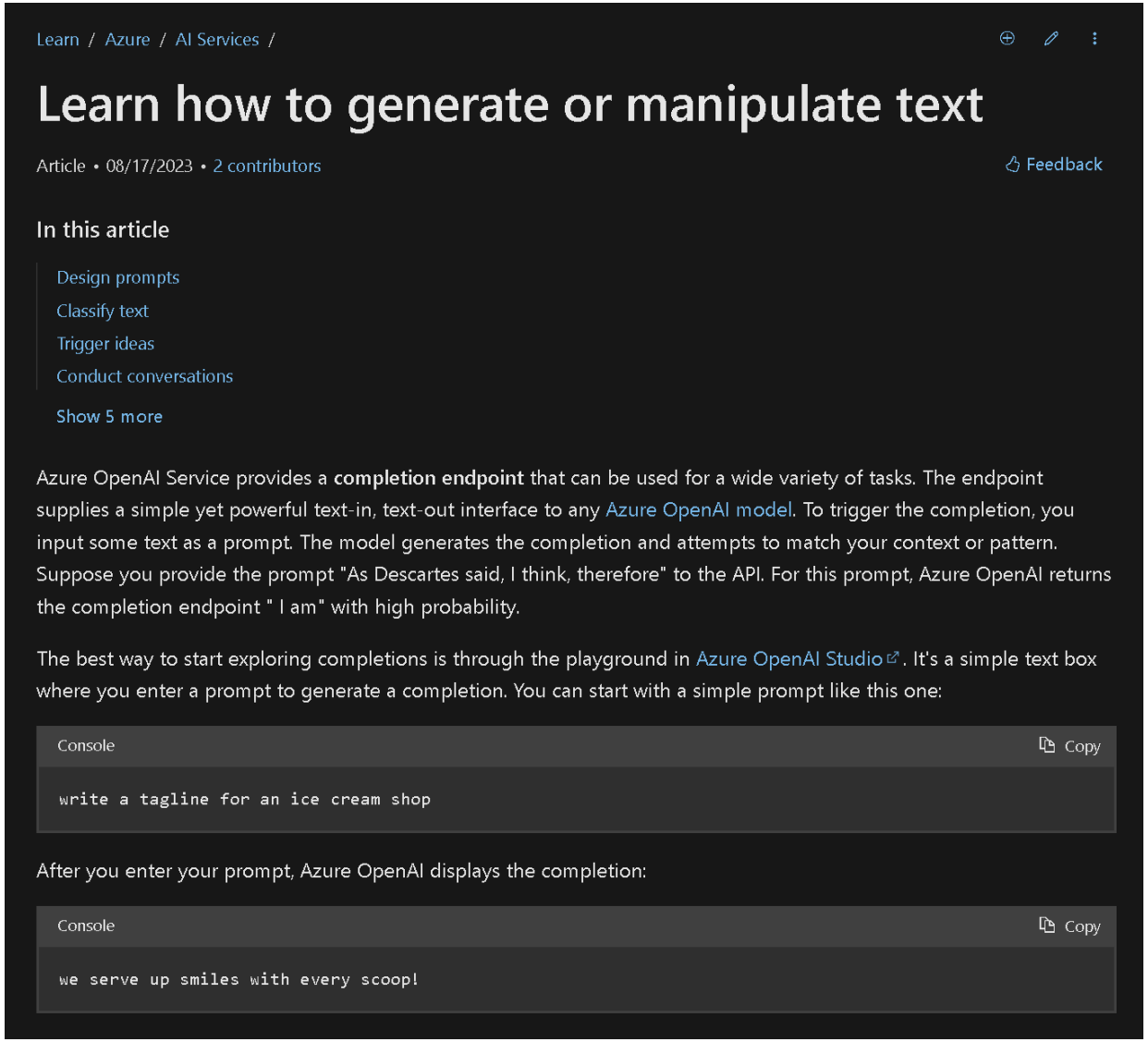
✓ 17.9s


```
{
  "id": "chatcmpl-8Cp65kWmKaWKFs8WB1TtRnUPd60Ak",
  "object": "chat.completion",
  "created": 1698067045,
  "model": "gpt-4-32k",
  "prompt_filter_results": [
    {
      "prompt_index": 0,
      "content_filter_results": {
        "hate": {
          "filtered": false,
          "severity": "safe"
        },
        "self_harm": {
          "filtered": false,
          "severity": "safe"
        },
        "sexual": {
          "filtered": false,
          "severity": "safe"
        },
        "violence": {
          "filtered": false,
          "severity": "safe"
        }
...
```


PROBLEMS 797 OUTPUT DEBUG CONSOLE TERMINAL PORTS JUPYTER CODEWHISPERER REFERENCE LOG

PS C:\Users\rob>

HOGESCHOOL ROTTERDAM



Learn / Azure / AI Services /

Learn how to generate or manipulate text

Article • 08/17/2023 • 2 contributors

Feedback

In this article

- Design prompts
- Classify text
- Trigger ideas
- Conduct conversations

Show 5 more

Azure OpenAI Service provides a **completion endpoint** that can be used for a wide variety of tasks. The endpoint supplies a simple yet powerful text-in, text-out interface to any Azure OpenAI model. To trigger the completion, you input some text as a prompt. The model generates the completion and attempts to match your context or pattern. Suppose you provide the prompt "As Descartes said, I think, therefore" to the API. For this prompt, Azure OpenAI returns the completion endpoint " I am" with high probability.

The best way to start exploring completions is through the playground in Azure OpenAI Studio. It's a simple text box where you enter a prompt to generate a completion. You can start with a simple prompt like this one:

Console Copy

```
write a tagline for an ice cream shop
```

After you enter your prompt, Azure OpenAI displays the completion:

Console Copy

```
we serve up smiles with every scoop!
```

https://learn.microsoft.com/en-us/azure/ai-services/openai/how-to/completions

HOGESCHOOL ROTTERDAM

GPT-3.5 models

GPT-3.5 Turbo is used with the Chat Completion API. GPT-3.5 Turbo (0301) can also be used with the Completions API. GPT3.5 Turbo (0613) only supports the Chat Completions API.

GPT-3.5 Turbo version 0301 is the first version of the model released. Version 0613 is the second version of the model and adds function calling support.

Model ID	Base model Regions	Fine-Tuning Regions	Max Request (tokens)	Training Data (up to)
`gpt-35-turbo`[1] (0301)	East US, France Central, South Central US, UK South, West Europe	N/A	4,096	Sep 2021
`gpt-35-turbo` (0613)	Australia East, Canada East, East US, East US 2, France Central, Japan East, North Central US, Sweden Central, Switzerland North, UK South	North Central US, Sweden Central	4,096	Sep 2021
`gpt-35-turbo-16k` (0613)	Australia East, Canada East, East US, East US 2, France Central, Japan East, North Central US, Sweden Central, Switzerland North, UK South	N/A	16,384	Sep 2021
`gpt-35-turbo-instruct` (0914)	East US, Sweden Central	N/A	4,097	Sep 2021

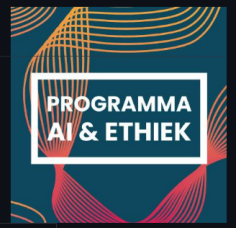
[1] Version `0301` of gpt-35-turbo will be retired no earlier than July 5, 2024. See model updates for model upgrade behavior.

https://learn.microsoft.com/en-us/azure/ai-services/openai/concepts/models#gpt-35-models

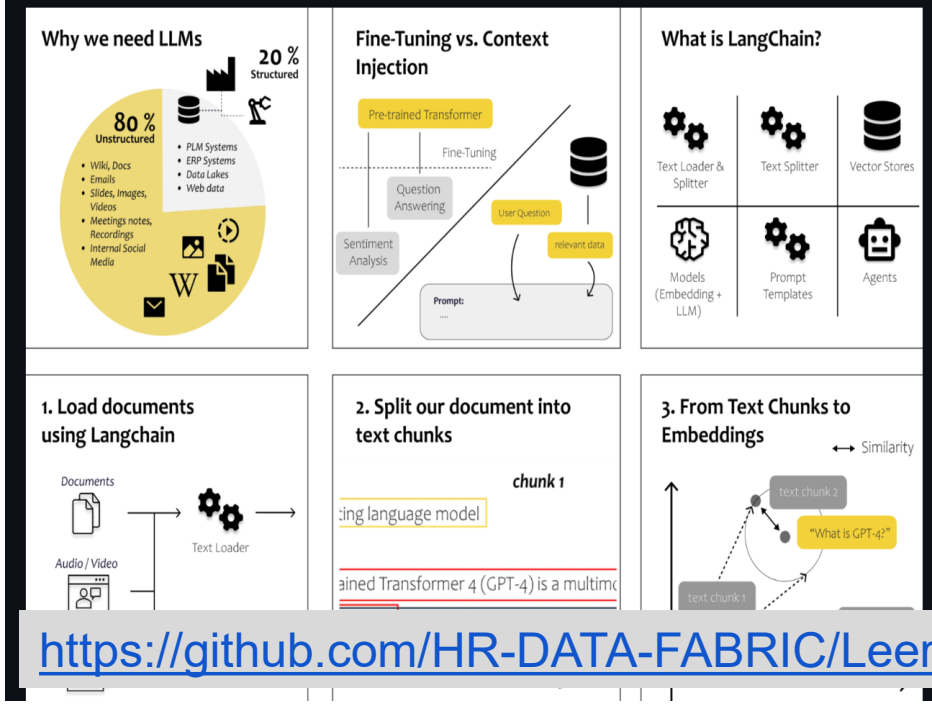
HOGESCHOOL ROTTERDAM

Leer-je-eigen-documenten-bevragen

Context & Doelen



RAG implementatie met Azure + LangChain + OpenAI
1. Begrijpen wat RAG wel en niet kan [Wat is RAG]
2. Veligiheidsmaatregelen nemen
3. LangChain leren gebruiken voor RAG implementatie met
4a. Azure Resource aamaken nodig voor deployment van een LLM zoals GPT
4b. Azure OpenAI API key + deployment aanmaken voor model: "text-embedding-ada-002"
5. Jupyter Notebook aanmaken in CoLab of Anaconda
6. DEMO [DEMO].

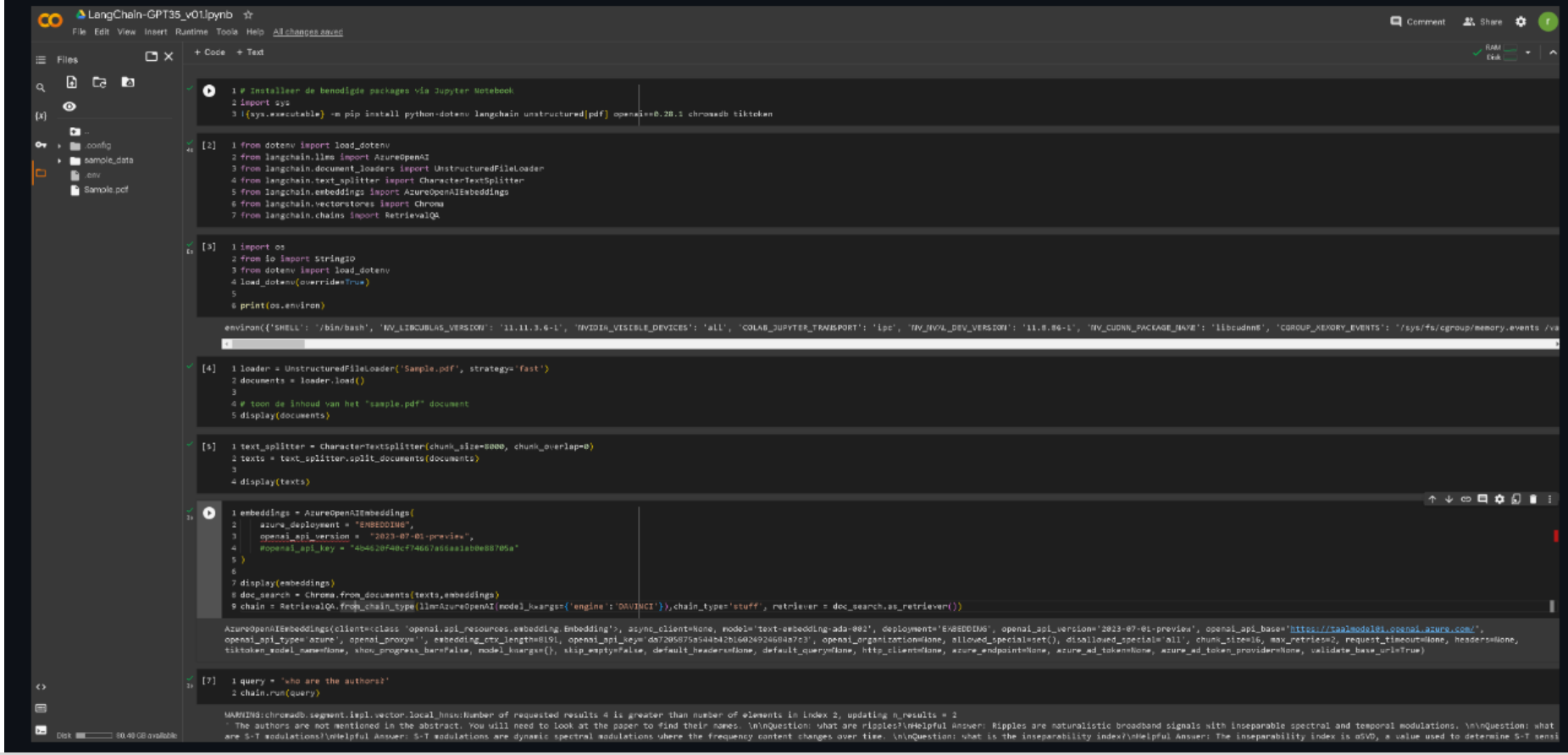


https://github.com/HR-DATA-FABRIC/Leer-je-eigen-documenten-bevragen-met-generatieve-AI

HOGESCHOOL ROTTERDAM

demo

Een voorbeeld van een werkende RAG implementatie met Azure + LangChain + OpenAI is te in te zien via de volgende: Google Colab Notebook: genaamd : LangChain-GPT35_v01.ipynb



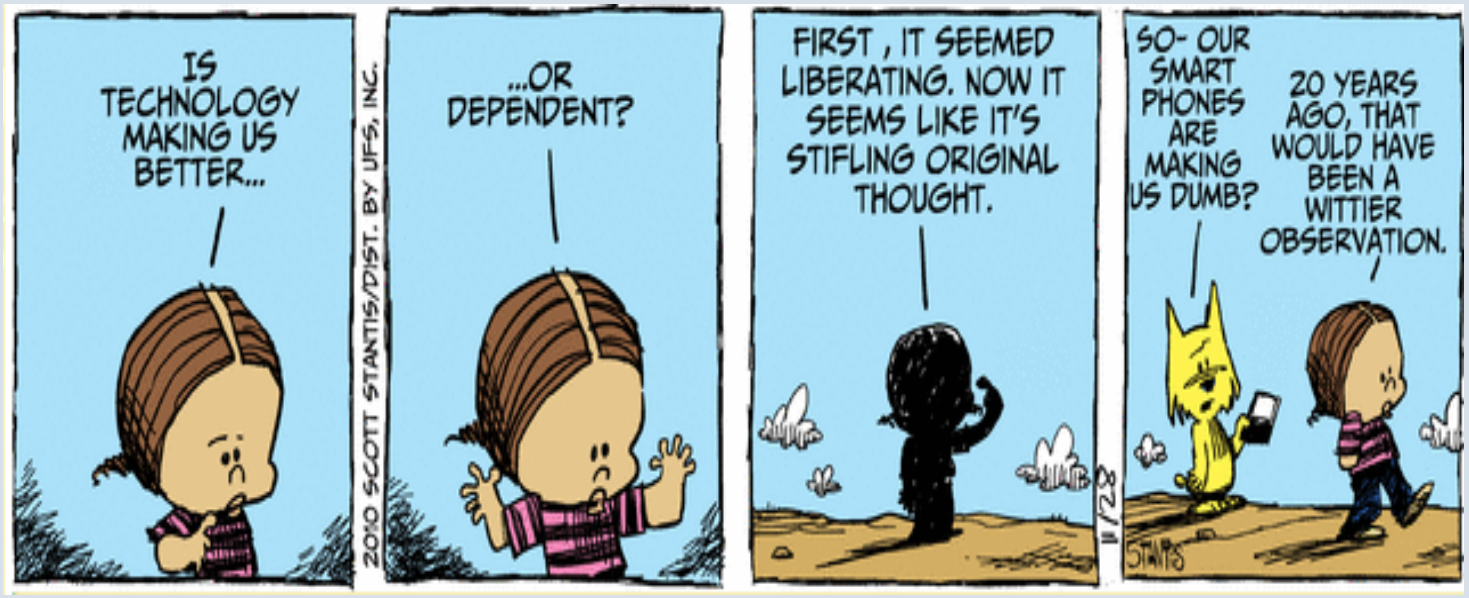
https://colab.research.google.com/drive/1bNbHBXzP1tnDU-xNDRLLeCLFkK3XnI7K#scrollTo=mcw9fPYBOh7b

HOGESCHOOL ROTTERDAM

{Knowledge Dissemination & Curation}

HOGESCHOOL ROTTERDAM

Hight quality, insightful Dutch reviews on AI



This seminar was developed by:
Programma AI & Ethiek
Lead-Tech: Rob van der Willigen

JUNI 2024