Verantwoord gebruik van ChatGPT en andere Generatieve AI-modellen

HOGESCHOOL ROTTERDAM

https://pretalx.surf.nl/spc24/talk/JGWJ7U/

hr.nl/ai

“De publieke beschikbaarheid van ChatGPT heeft de schijnwerpers gericht op generatieve-AI [GEN-AI].

Hogeschool Rotterdam [HR] onderzoekt hoe deze vorm van kunstmatige intelligentie kan worden benut om interactieve spraakmodellen te creëren die zowel Betrouwbaar (waarheidsgetrouw), reproduceerbaar als veilig zijn.

De nadruk ligt hierbij op het waarborgen van privacy, explainability & bias-mitigation van de informatie-uitwisseling tussen mens en machine."

HOGESCHOOL ROTTERDAM

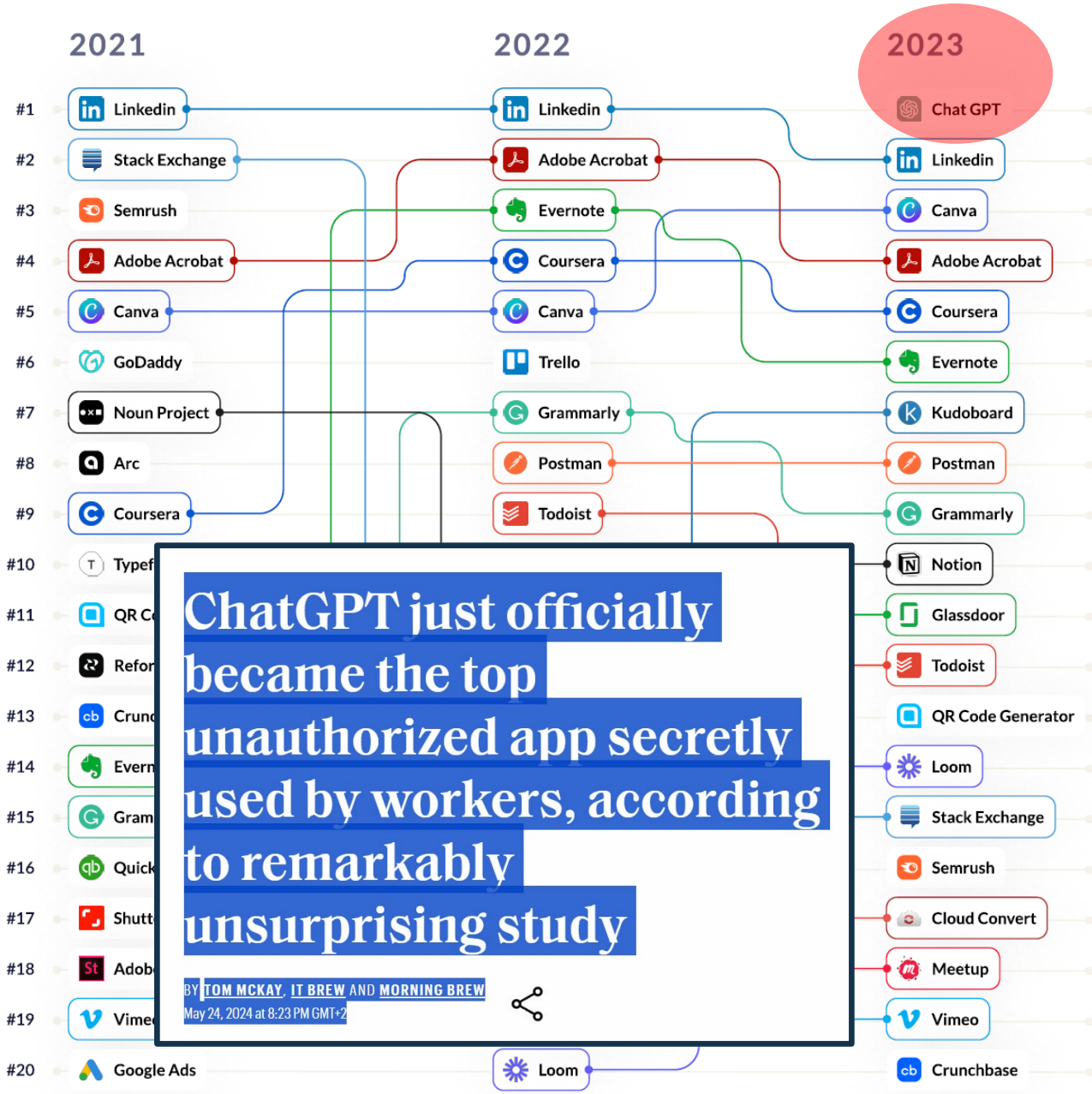
Interest in AI applications and features increased shadow IT. ChatGPT claimed #1 spot

https://www.itbrew.com/stories/2024/05/22/chatgpt-is-the-number-one-offender-in-shadow-it-report-finds

KEY TAKEAWAYS

- ChatGPT has jumped to the top of the shadow IT chart as employees continue to adopt Artificial Intelligence.
- As the innovators and early adopters within a company continue to seek out AI-native applications (like ChatGPT and Grammarly) and AI solutions (like those offered by Canva and Evernote) for unmet needs, organizations should be developing a cohesive AI strategy.
- Nearly every application here offers, or will likely offer, some type of AI functionality; #5, Coursera, saw signups for AI courses every minute in 2023, on average. They also demonstrate the continued strength of the PLG go-to-market motion, with most offering free signup variants.
- Outside of AI trends, use of LinkedIn stayed consistent during a time of increased revenue pressure and insecurity in the job market. Trello fell off the shadow IT chart in 2022, as more companies purchase Atlassian's suite of products.

MOST POPULAR SHADOW IT



HOGESCHOOL ROTTERDAM

Wat is schaduw-ICT [shadow IT / gray IT]

Schaduw-ICT omvat ongeautoriseerde *hardware, software of diensten die voor "zakelijke" doeleinde (onderwijs en/of onderzoek) worden ingeregeld, ingevoerd en/of gebruikt zonder uitdrukkelijke goedkeuring of medeweten van de organisatie / systeembeheerders en/of technische staf*.

Omdat schaduw-IT niet wordt meegenomen in assetmanagement en evenmin aansluiten bij het AVG-compliance beleid, vormen ze een veiligheid en/of compliance risico.

Dit kan leiden tot het lekken van gevoelige gegevens (datalekken)of de verspreiding van malware binnen de organisatie.

https://www.ncsc.gov.uk/guidance/shadow-it

SURF: Cybersecurity

Schaduw-ict in onderwijs en onderzoek: wat moet je er als instelling mee? Latest update: 15 februari 2023

https://communities.surf.nl/cybersecurity/artikel/schaduw-ict-in-onderwijs-en-onderzoek-wat-moet-je-er-als-instelling-mee

In het onderwijs en onderzoek krijgen studenten, wetenschappers en docenten veel vrijheid in de manier waarop ze ict gebruiken.

Faculteiten en instituten werken autonoom, zodat ze snel kunnen inspringen op nieuwe ontwikkelingen. Mede daardoor ontstaat schaduw-ict.

Dit kan voor ict-afdelingen leiden tot een verlies van controle.

In het bijzonder voor de netwerk- en informatiebeveiliging zijn de risico's moeilijk te overzien en veel instellingen vragen zich af wat ze er mee aan moeten.

Beveiligingsincidenten en datalekken melden (bijgewerkt op 31 augustus 2023)

Als school probeer je de persoonsgegevens die je bewaart zo goed mogelijk te beveiligen. Maar soms gaat er is iets mis. Je hebt bijvoorbeeld zelf geen toegang meer tot je gegevens of buitenstaanders krijgen onbedoeld toegang. Dan is er sprake van een beveiligingsincident. Heeft het incident gevolgen voor de privacy van leerlingen of medewerkers? Dan spreken we van een datalek. Voor het melden van datalekken gelden andere regels dan voor andere beveiligingsincidenten. Volgens de Algemene Verordening Gegevensbescherming (AVG) ben je als school verplicht datalekken direct te melden bij de Autoriteit Persoonsgegevens.

https://aanpakibp.kennisnet.nl/beveiligingsincidenten-en-datalekken-melden/

Shadow IT creates the possibility that organizations may run afoul of regulations such as PCI-DSS, GDPR, HIPAA, SOX and others, exposing them to severe penalties and fines. It can also lead to an increase in the likelihood of data breaches when IT and security operations lose control over the software and applications used in an environment.

https://www.forbes.com/sites/forbestechcouncil/2022/07/19/how-shadow-it-can-keep-compliance-efforts-in-the-dark/

Cyber
Resilience & Hygiene

Vereist een
Zero Trust-aanpak

Beveiligingshygiëne op basisniveau vereist:

Phishing-bestendige meervoudige verificatie (MFA)

Zero Trust-aanpak

Gebruik moderne anti-malware

Software/Patch Management (mitigatie van schaduw-IT)

Data Encryptie

https://www.microsoft.com/nl-nl/security/security-insider/practical-cyber-defense/cyber-resilience-hygiene-guide

HOGESCHOOL ROTTERDAM

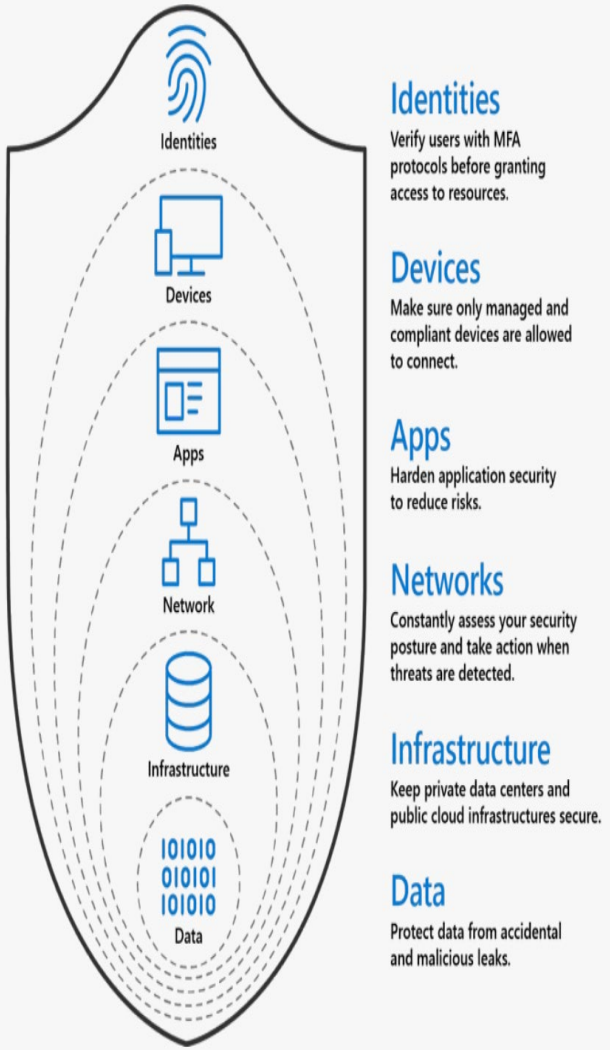
HOGESCHOOL ROTTERDAM

De Zero Trust-principes zijn:

- **Uitgaan van inbreuk** Ga ervan uit dat aanvallers alles kunnen en zullen aanvallen (identiteit, netwerk, apparaat, app, infrastructuur, etc.) en pas je planning daarop aan. Dit betekent dat de omgeving continu moet worden gecontroleerd op mogelijke aanvallen.

- **Uitdrukkelijk verifiëren** Zorg ervoor dat gebruikers en apparaten correct werken voordat ze toegang krijgen tot resources. Bescherm assets tegen controle van aanvallers door expliciet te valideren dat alle beslissingen over vertrouwen en beveiliging gebruik maken van relevante beschikbare informatie en telemetrie.

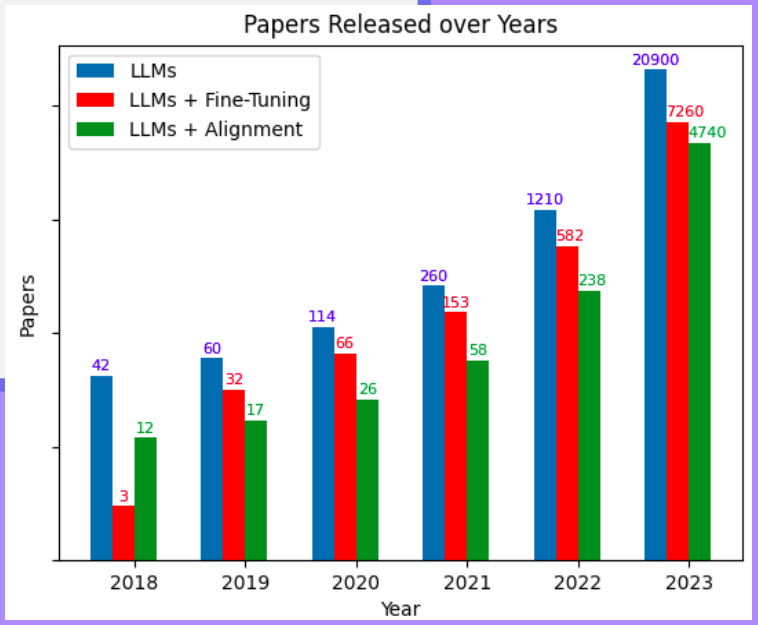
- **Toegang met minimale machtigingen gebruiken** Beperk de toegang tot een mogelijk gecompromitteerde asset met Just-In-Time en Just-Enough-Access (JIT/JEA) en een op risico's gebaseerd beleid, zoals adaptief toegangsbeheer. Je moet alleen het recht toestaan dat nodig is voor een bepaalde resource, en niets meer.



https://www.microsoft.com/nl-nl/security/security-insider/practical-cyber-defense/cyber-resilience-hygiene-guide

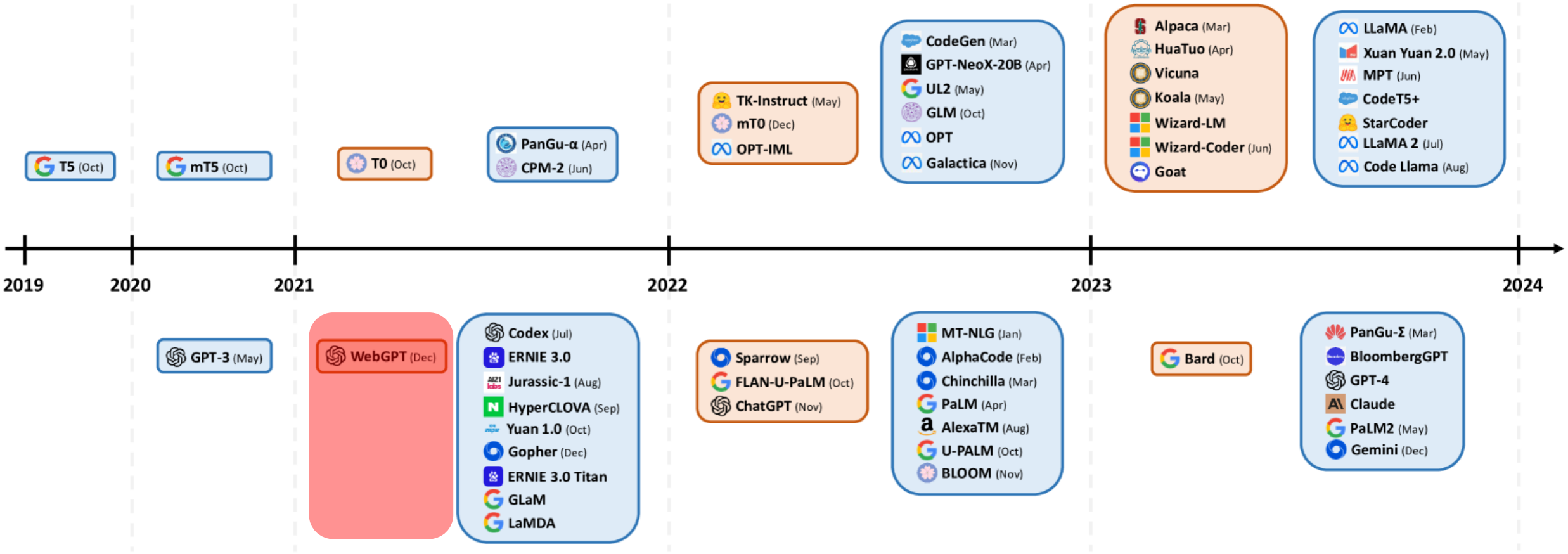
CONTEXT

Waarom heeft Generatieve AI zo'n enorme impact op onze samenleving?



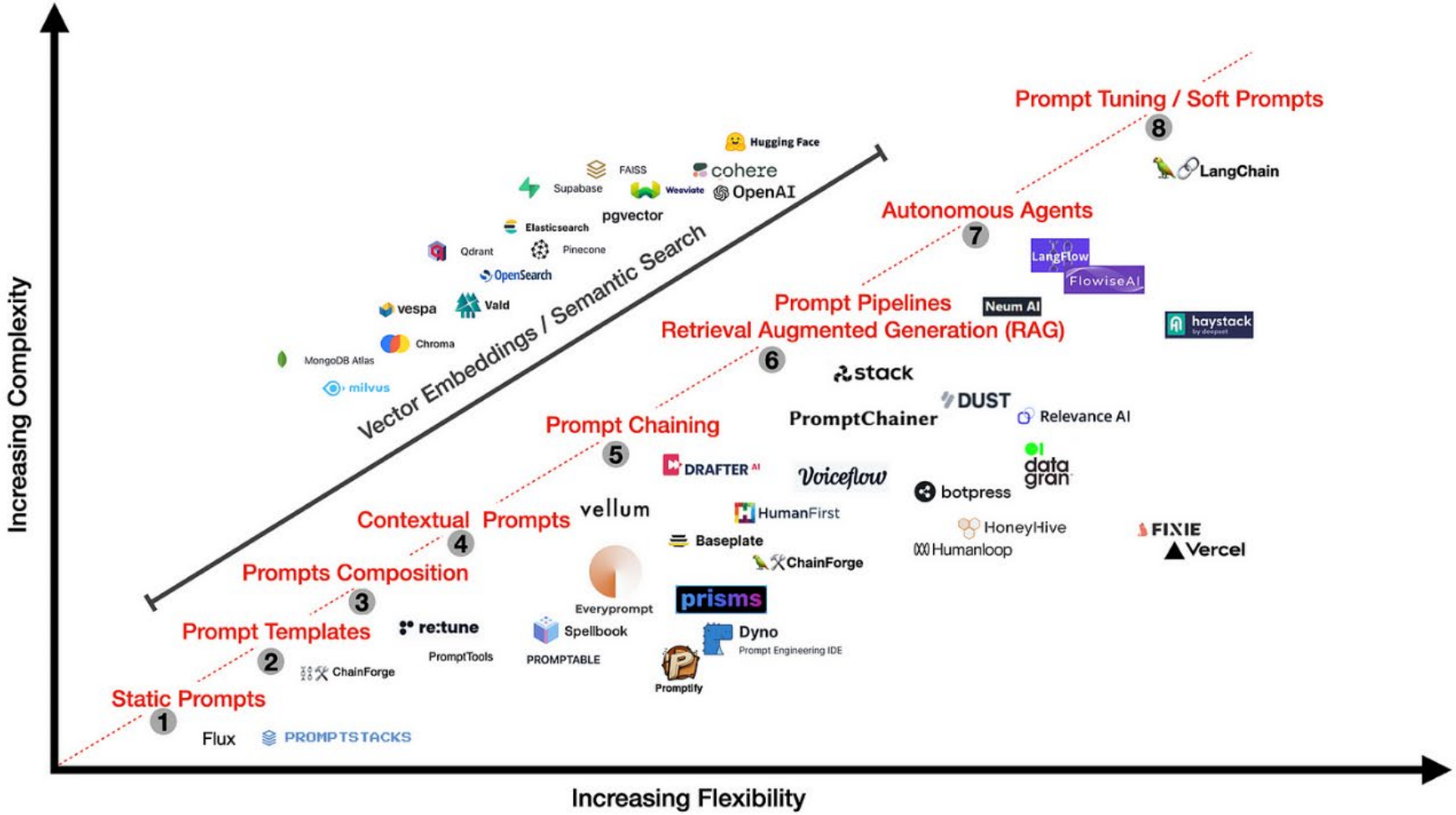
Ontstaansgeschiedenis + evolutie van grote taal modellen {LLM}

HOGESCHOOL ROTTERDAM



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

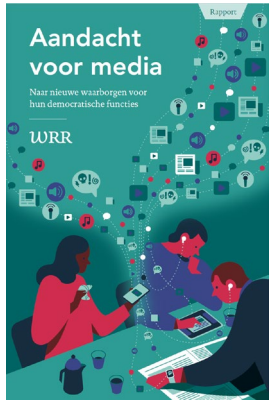
LLM implementaties



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

Kernpunten 3.2 De nieuwe infrastructuur van de publieke ruimte

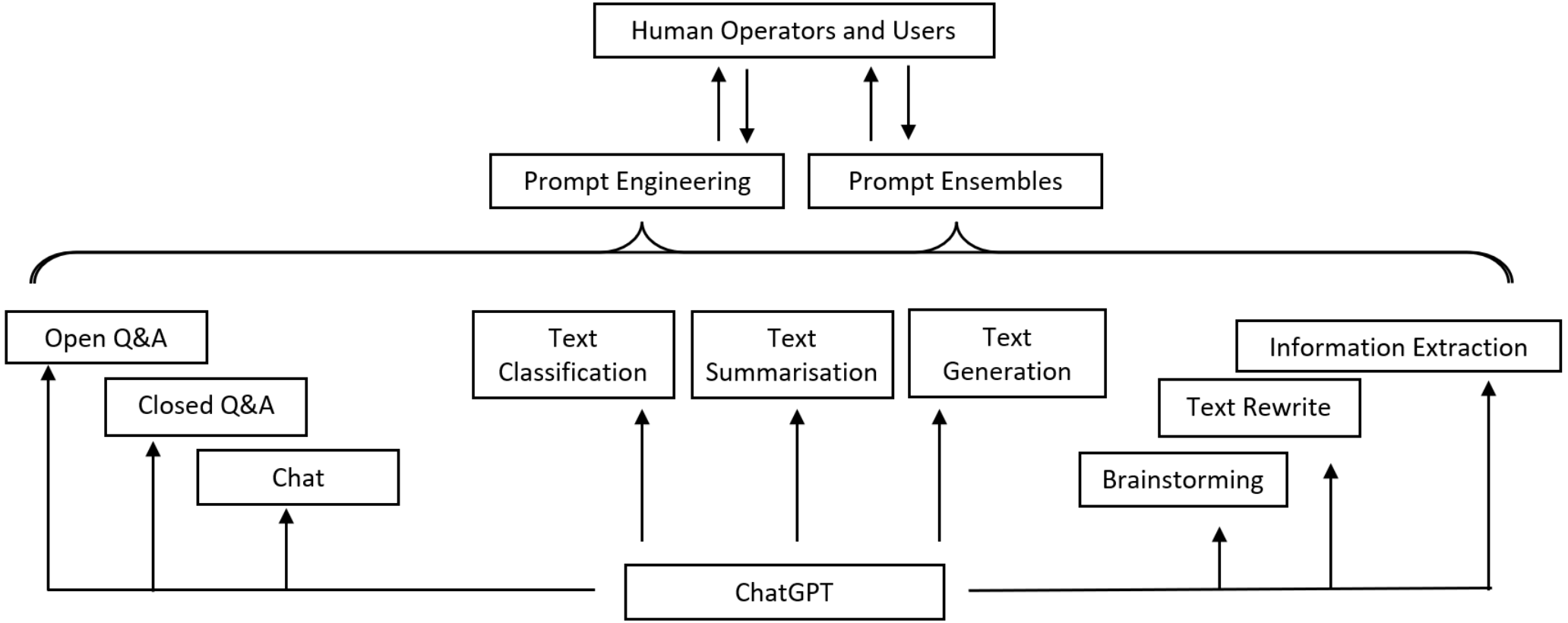
- Technologiebedrijven hebben een infrastructuur ontworpen voor de online ordening van informatie en de interacties tussen gebruikers en content- en dienstenleveraars.
- Deze infrastructuur is dusdanig efficiënt, aantrekkelijk en gebruiksvriendelijk ontworpen dat hij onmisbaar is geworden voor het functioneren van de publieke ruimte. Op platforms vindt het grootste gedeelte van het publieke debat plaats, inclusief persoonlijke en zakelijke communicatie.
- De bedrijven achter deze platforms hebben hierdoor een nieuwe rol verkregen. Zij beheren de toegang tot de publieke ruimte en ordenen de informatie die daarbinnen circuleert.
- Platformisering speelt vooral bij uitgevers; radio (ether) en televisie zijn nog altijd belangrijke en veelgebruikte infrastructuren. Toch ondervinden ook zij in toenemende mate concurrentie van technologiebedrijven als het gaat om de distributie van content.



Ontwikkeling internet	Instantie	Wijze van aggregatie	Voorbeelden
Fase 1	Portals/startpagina's	Statisch overzicht websites met hyperlinks, gegenereerd door gebruikers	*Yahoo!* *Startpagina.nl*
Fase 2	Zoekmachines	Algoritmische selectie van internetbronnen op basis van zoekopdrachten van gebruiker	*Google* *Bing* *Ilse*
Fase 3	Sociale media	Permanente stroom (grotendeels) algoritmisch gecureerde en gepersonaliseerde contentbundels	*TikTok, Facebook, Instagram, YouTube, X*
Fase 4	AI-chatbots	Integrale, op maat gegenereerde content op basis van internetinformatie	*ChatGPT, Llama*

ChatBot Use-Cases

HOGESCHOOL ROTTERDAM



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

HOGESCHOOL ROTTERDAM

Prompt Taxonomie

openeinde spectrum

Uitkomst vertoont weinig variatie

PROMPT is COMPLEX

Uitkomst is onzeker vertoont zeer veel variatie

PROMPT is SIMPEL

Definiëring
karakteristieke kenmerken
GEN-AI *geeft inzicht*

AI definities volgens HAI

Artificial Intelligence Definitions

https://hai.stanford.edu/sites/default/files/2020-09/AI-Definitions-HAI.pdf

Intelligence might be defined as the ability to learn and perform suitable techniques to solve problems and achieve goals, appropriate to the context in an uncertain, ever-varying world. A fully pre-programmed factory robot is flexible, accurate, and consistent but not intelligent.

Artificial Intelligence (AI), a term coined by emeritus Stanford Professor John McCarthy in 1955, was defined by him as "the science and engineering of making intelligent machines". Much research has humans program machines to behave in a clever way, like playing chess, but, today, we emphasize machines that can learn, at least somewhat like human beings do.

Autonomous systems can independently plan and decide sequences of steps to achieve a specified goal without micro-management. A hospital delivery robot must autonomously navigate busy corridors to succeed in its task. In AI, autonomy doesn't have the sense of being self-governing common in politics or biology.

Machine Learning (ML) is the part of AI studying how computer agents can improve their perception, knowledge, thinking, or actions based on experience or data. For this, ML draws from computer science, statistics, psychology, neuroscience, economics and control theory.

In **supervised learning**, a computer learns to predict human-given labels, such as dog breed based on labeled dog pictures; **unsupervised learning** does not require labels, sometimes making its own prediction tasks such as trying to predict each successive word in a sentence; **reinforcement learning** lets an agent learn action sequences that optimize its total rewards, such as winning games, without explicit examples of good techniques, enabling autonomy.

Deep Learning is the use of large multi-layer **(artificial) neural networks** that compute with continuous (real number) representations, a little like the hierarchically organized neurons in human brains. It is currently the most successful ML approach, usable for all types of ML, with better generalization from small data and better scaling to big data and compute budgets.

An **algorithm** lists the precise steps to take, such as a person writes in a computer program. AI systems contain algorithms, but often just for a few parts like a learning or reward calculation method. Much of their behavior emerges via learning from data or experience, a sea change in system design that Stanford alumnus Andrej Karpathy dubbed **Software 2.0**.

Narrow AI is intelligent systems for one particular thing, e.g., **speech** or **facial recognition**. **Human-level AI**, or **Artificial General Intelligence (AGI)**, seeks broadly intelligent, context-aware machines. It is needed for effective **social chatbots** or **human-robot interaction**.

Human-Centered Artificial Intelligence is AI that seeks to augment the abilities of, address the societal needs of, and draw inspiration from human beings. It researches and builds effective partners and tools for people, such as a robot helper and companion for the elderly.

Text by Professor Christopher Manning, September 2020

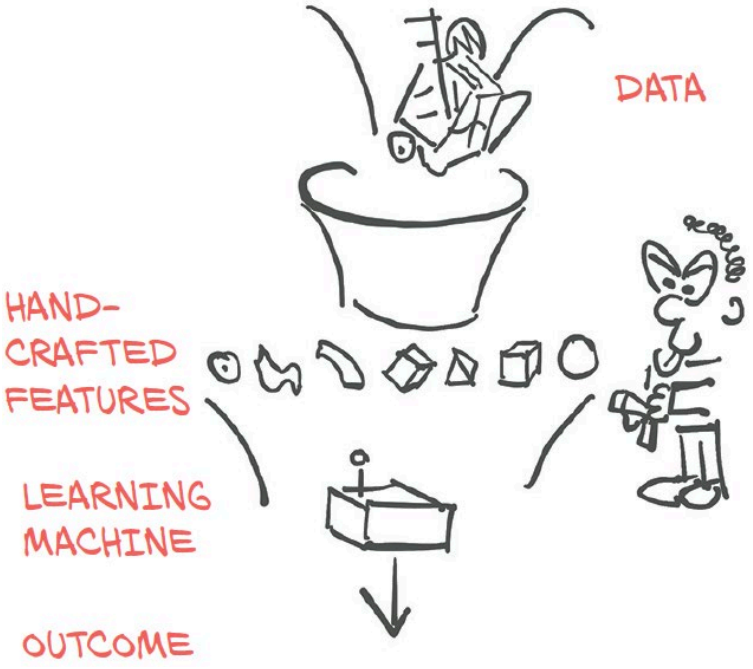
HOGESCHOOL ROTTERDAM

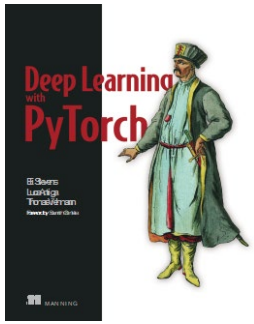
{AI Paradigm-shift ➔ NO Human input needed}

HOGESCHOOL ROTTERDAM

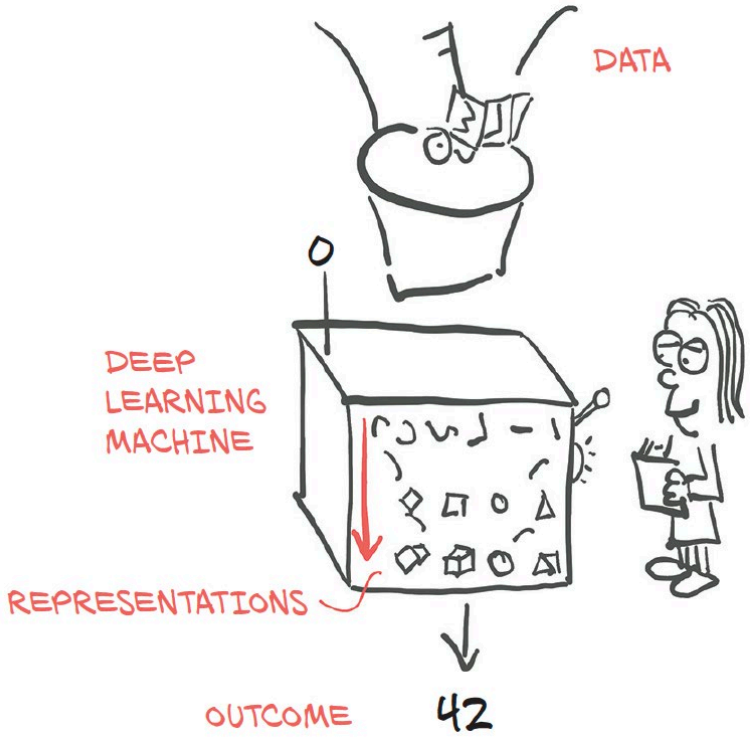
More data, parameters & computing power | Less human-in-the-loop

Machine Learning Paradigm {ML}





Deep Learning Paradigm {DL}



AI-taxonomie

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

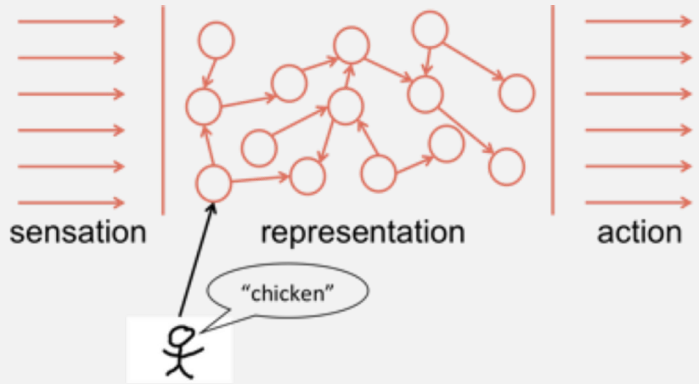
Modern Computational Intelligence

Representations are distributed not readable for humans

SYMBOLIC

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.

SUBSYMBOLIC

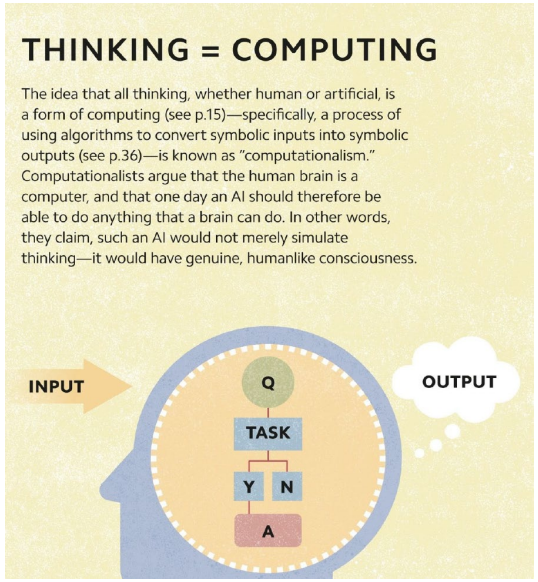


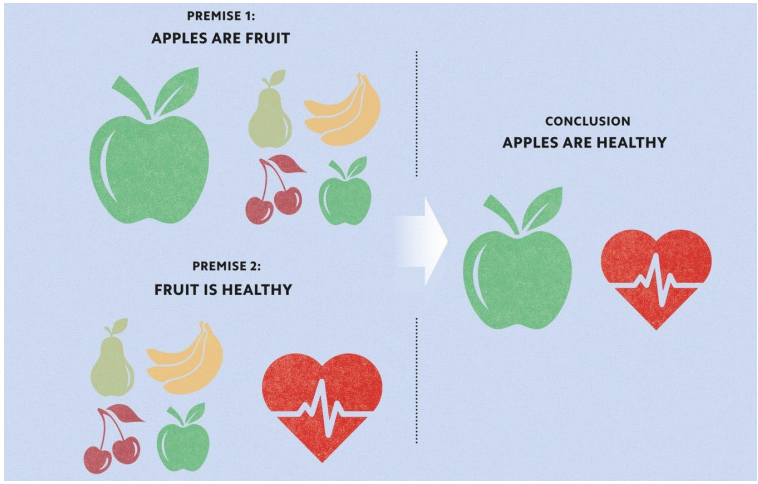
SYMBOLIC AI

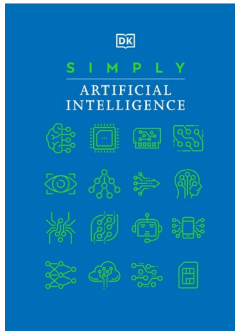
TOP-DOWN / open-loop

Intelligentie (denken) is
een vorm van Logica “Rekenkracht”
(computationalism)

Hersenen zijn een
“following the rules” Computer



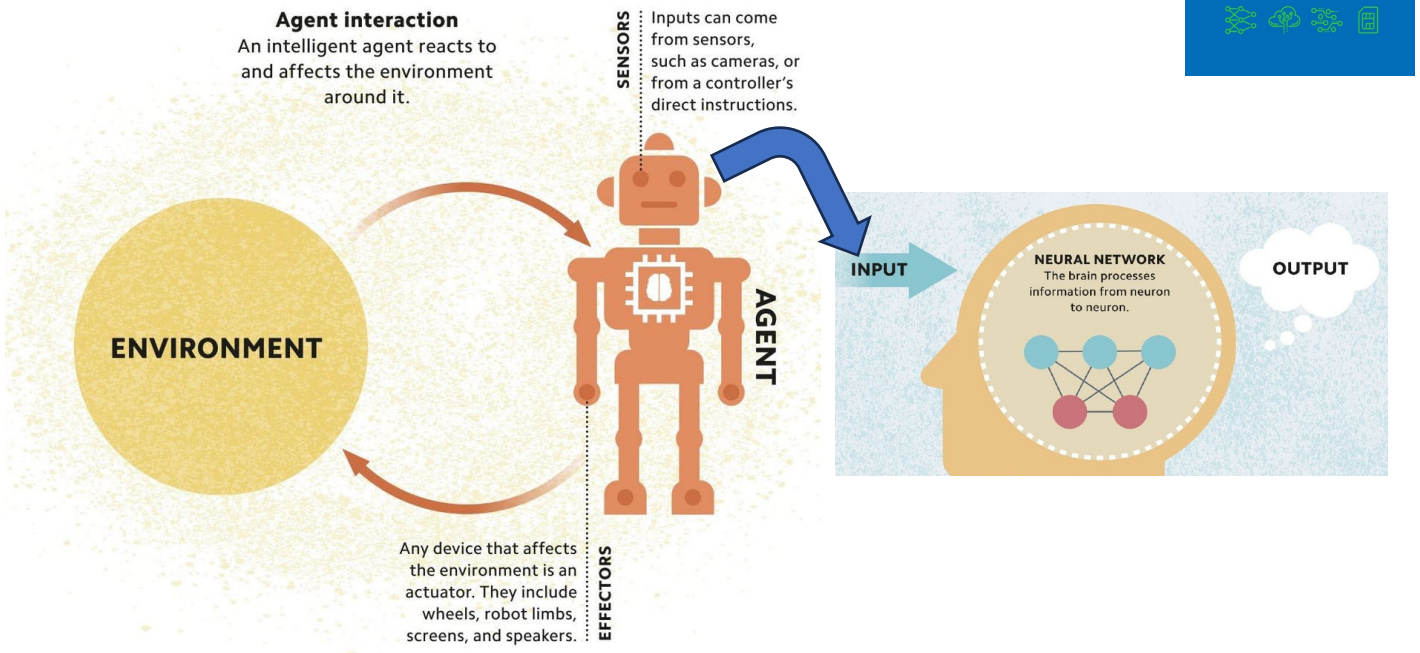




SUBSYMBOLIC AI

BOTTOM-UP / closed feedback-loop

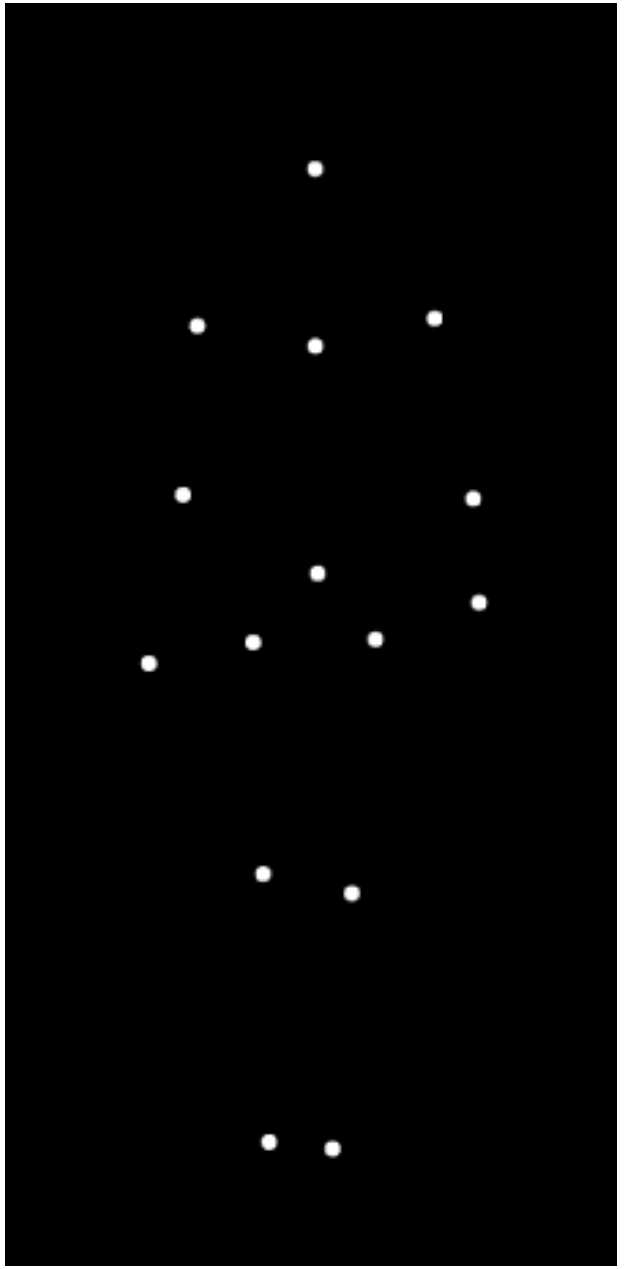
Intelligentie representeert adaptatief-leervermogen ***(trial & error)*** dat wordt mogelijk gemaakt door netwerken bestaande uit simpele rekeneenheden ***(connectionism)*** aangestuurd door een algoritme + feedback loop ***(Cybernetica)***



{Top Down}

Biological motion is a prime example of compuationalism as performed by the human brain.

Only 12 moving-dots are needed to determine direction of motion.



HTML5 BML Walker – BioMotionLab

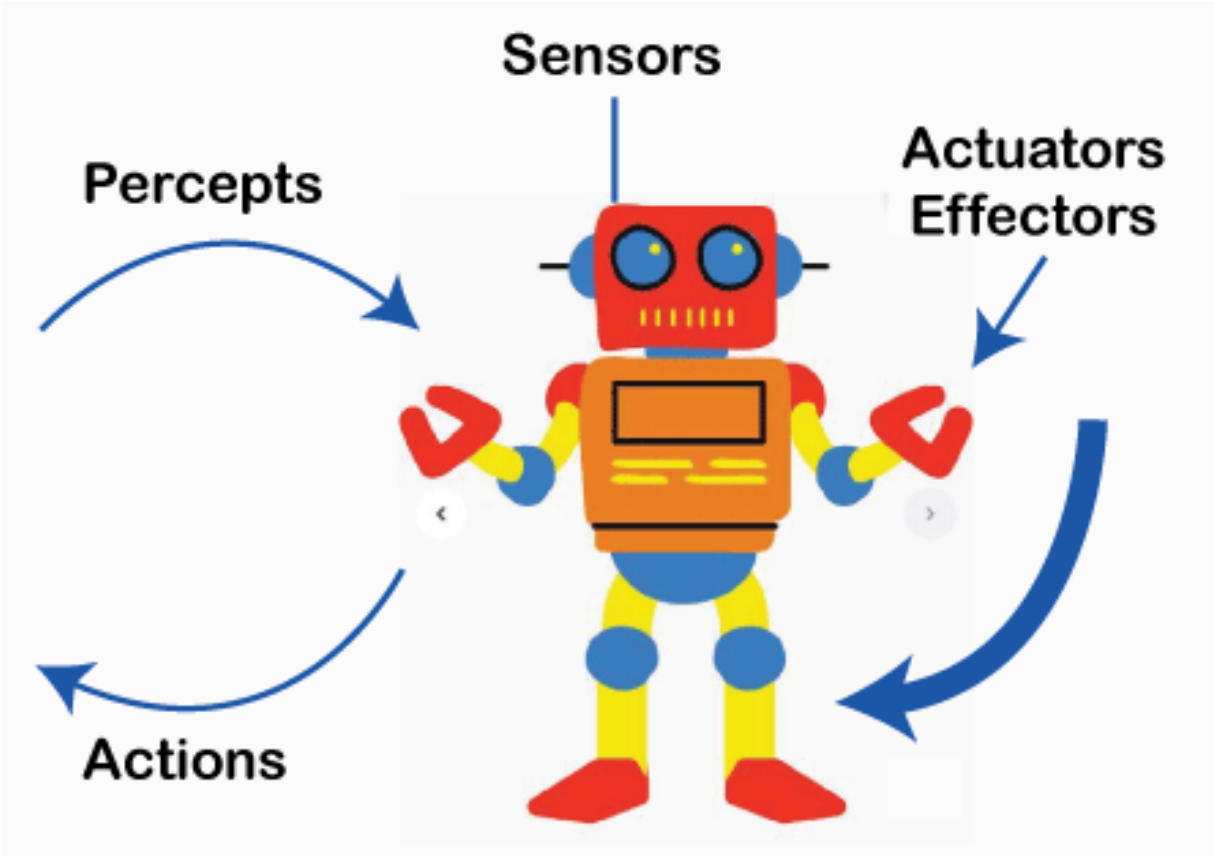
Bradshaw, M. & van der Willigen, R. F., (1999).
The walker's direction affects the perception of biological motion.
In M. A. Grealy & J. A. Thomson (Eds.),
Studies in perception and action V (pp. 3–6). London: Academic Press.

{MULTIMODALE AGENT}

HOGESCHOOL ROTTERDAM

```
AGENT representeert input / output model
```

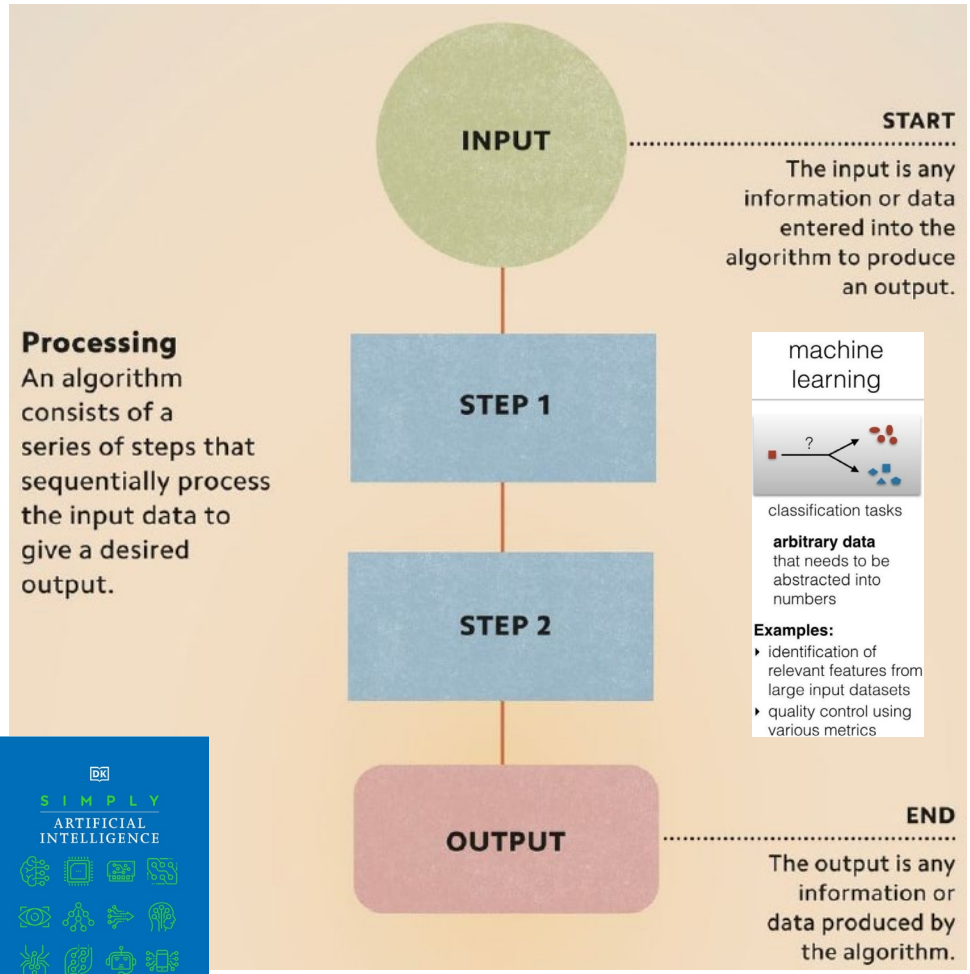
EXTERNAL INPUT

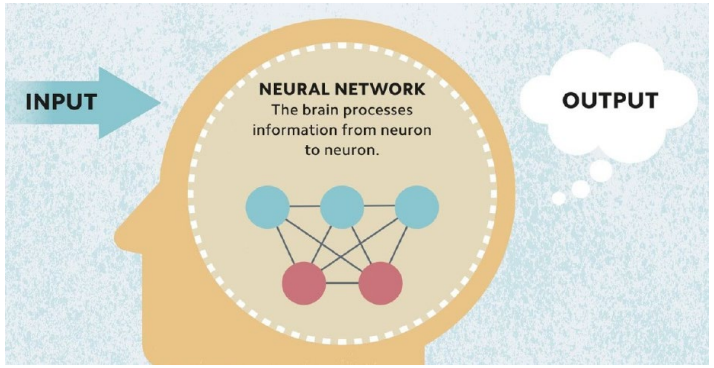


{ALGORITME}

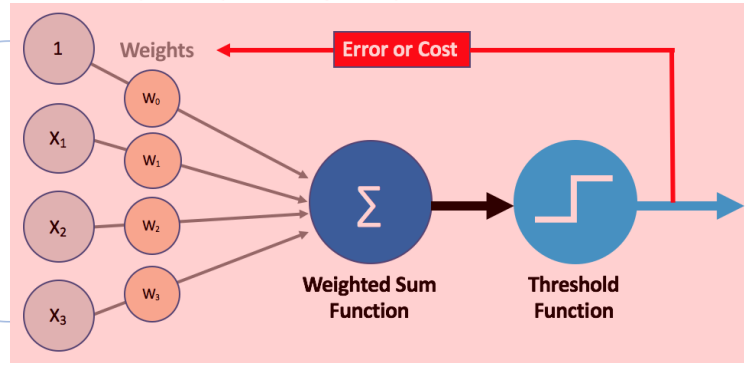
{AI}

https://learn.microsoft.com/nl-nl/dotnet/machine-learning/deep-learning-overview



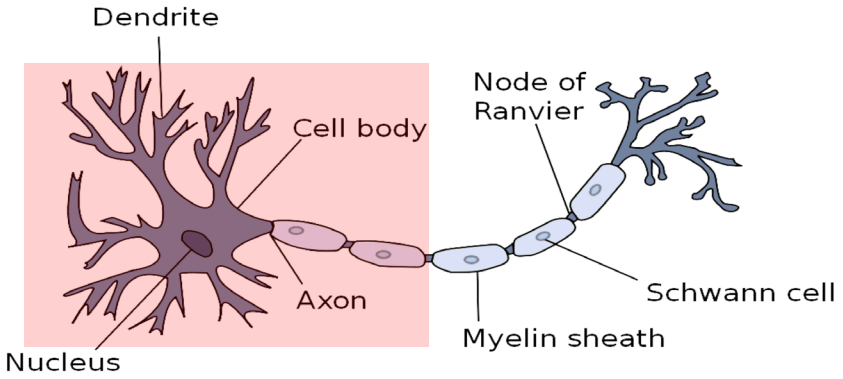


The weights are updated through a process of minimising a cost function



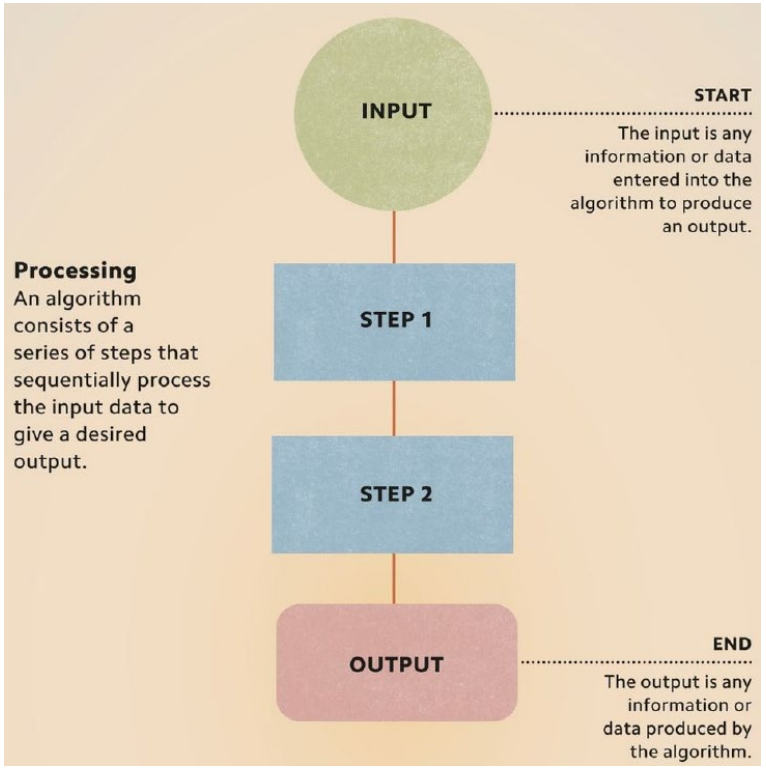
Names for Artificial Neurons

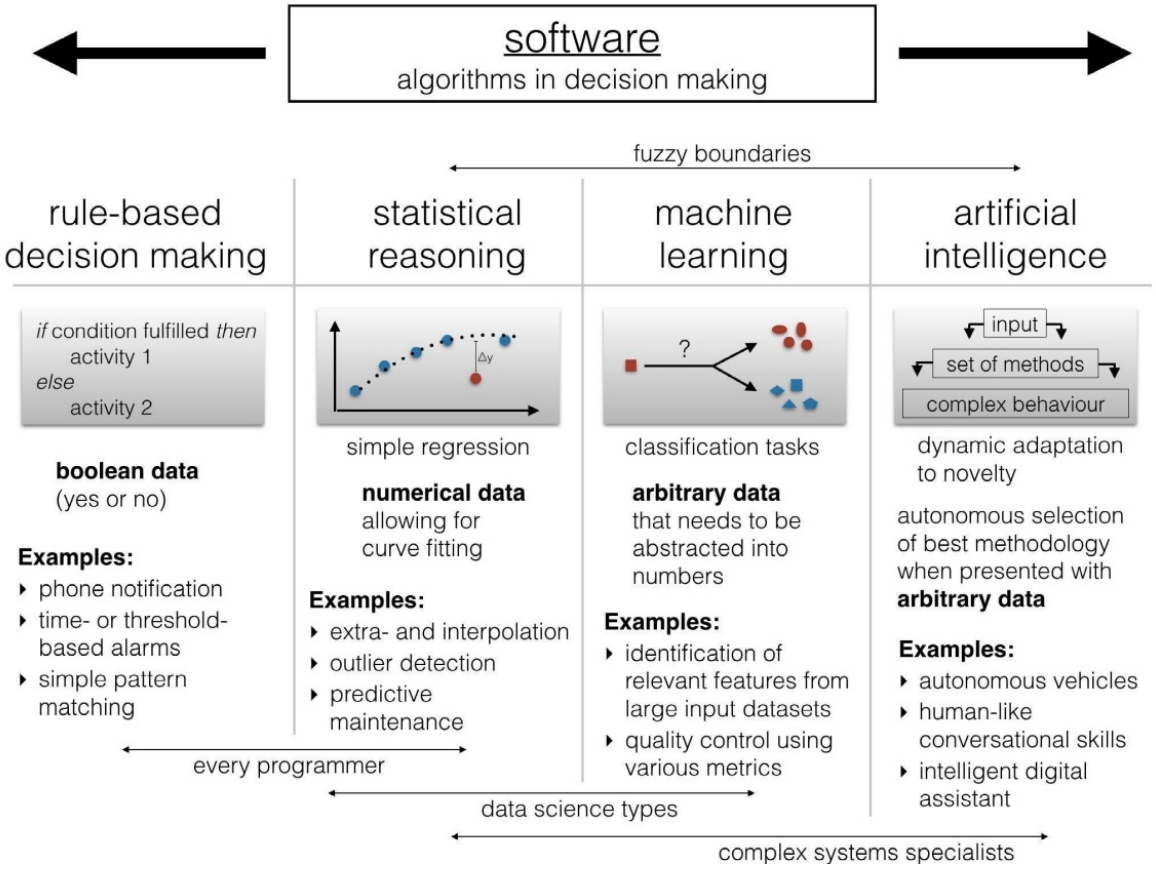
{unit}
{cell}
{node}
{perceptron}



{COMPUTER ALGORITME}

HOGESCHOOL ROTTERDAM



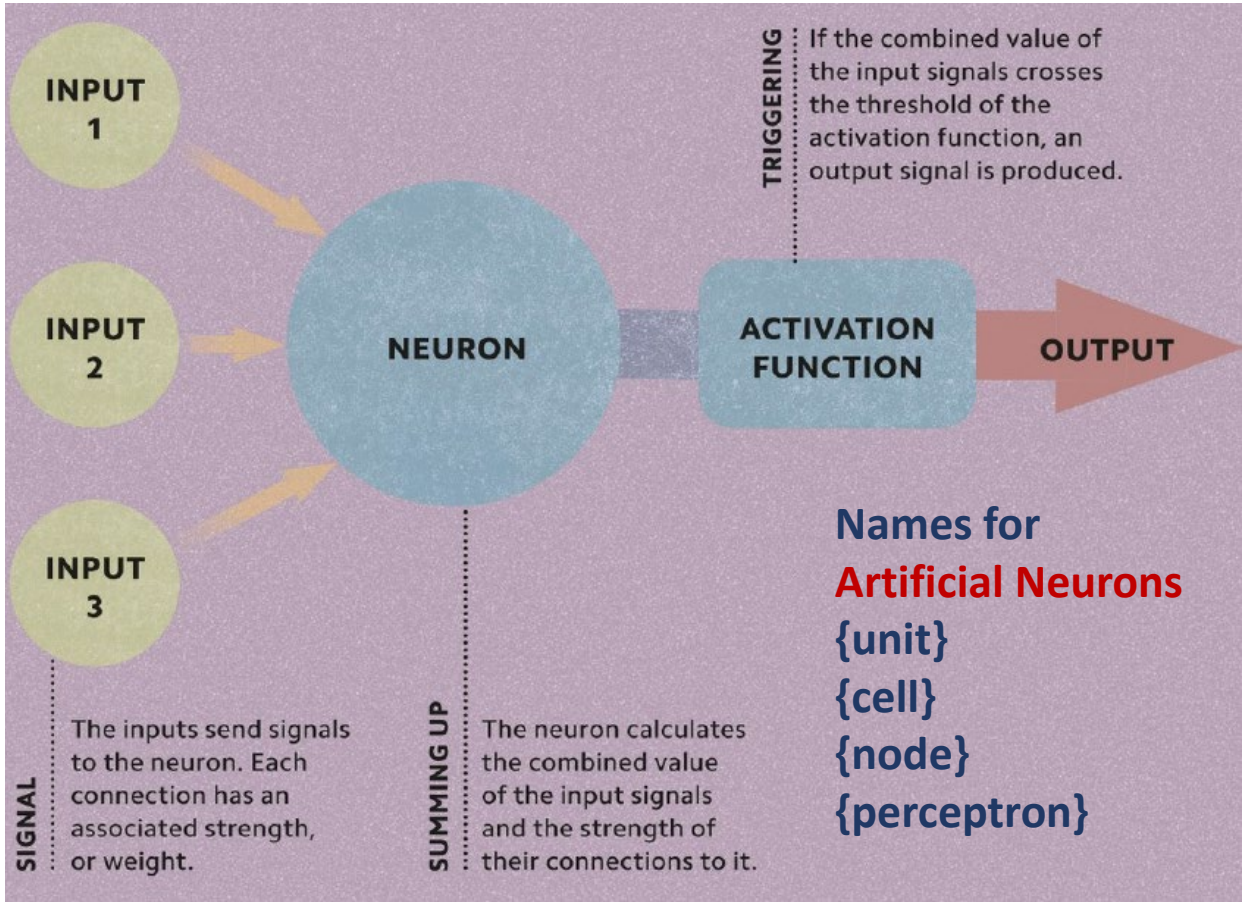


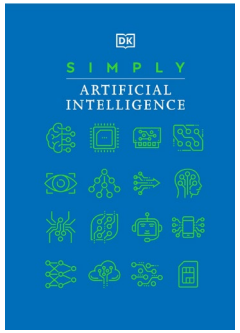
{KUNSTMATIGE NEURON}

HOGESCHOOL ROTTERDAM

computation

Input → Activation((Input * Weight) + Bias) → Output

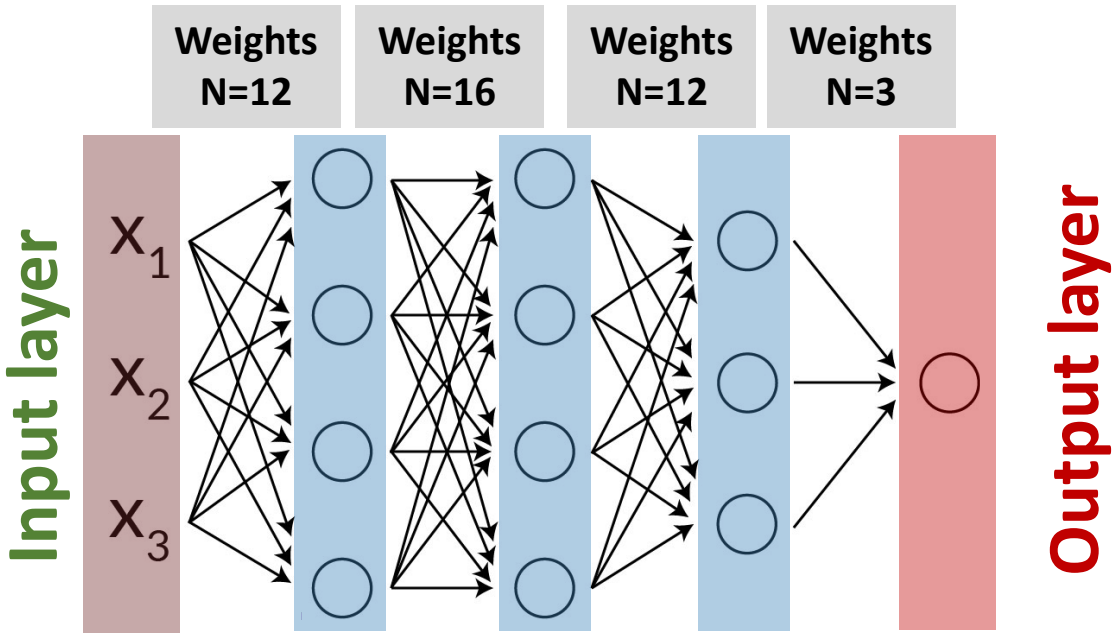




{Neuraal Netwerk [NN] Typen}

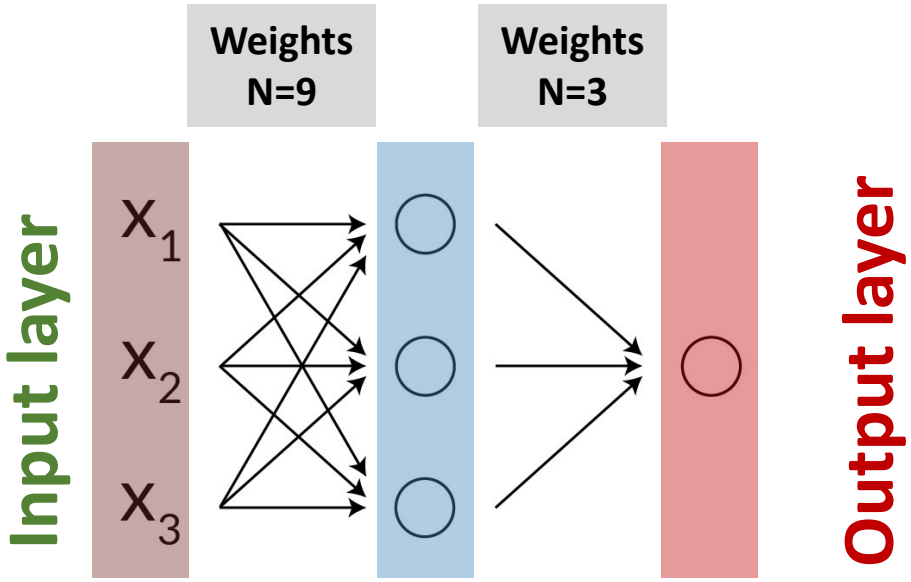
HOGESCHOOL ROTTERDAM

Neural Network {NN} Lagen Achitectuur



hidden layers N>2

Deep network



Hidden layer N=1

Shallow network

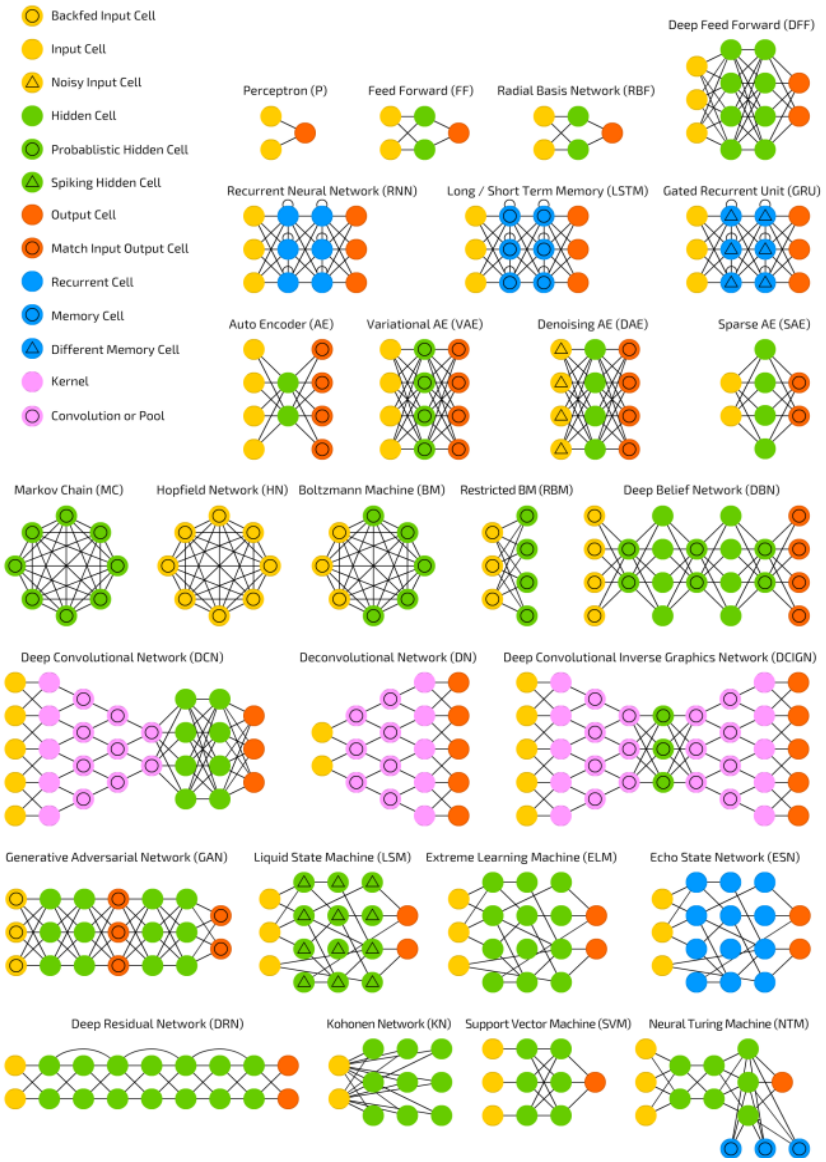
{Topology}

Topology of a neural network refers to the way artificial neurons are connected to form a network.

Form follows function!

The topology of a network determines the degree of perplexity of the tasks it can

https://pub.towardsai.net/main-types-of-neural-networks-and-its-applications-tutorial-734480d7ec8e

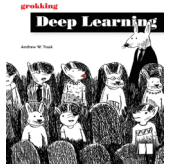


https://www.mdpi.com/2504-3900/47/1/9

{Bottum-UP: Machinaal Leren [ML]}

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

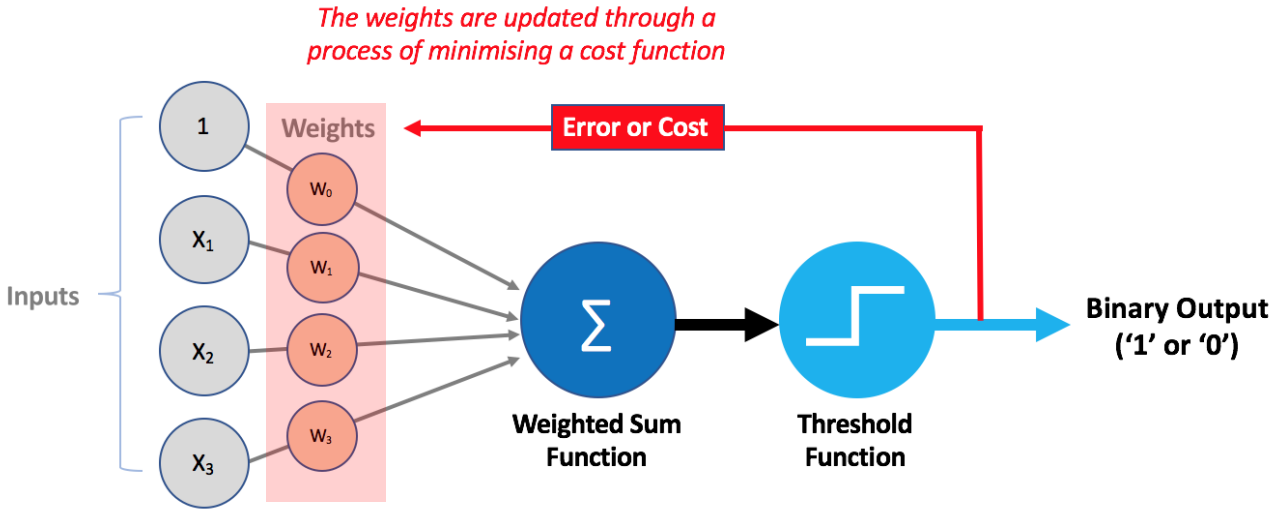
Given that deep learning is a subset of machine learning, what is machine learning? Most generally, it is what its name implies. Machine learning is a subfield of computer science wherein *machines learn* to perform tasks for which they were *not explicitly programmed*. In short, machines observe a pattern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

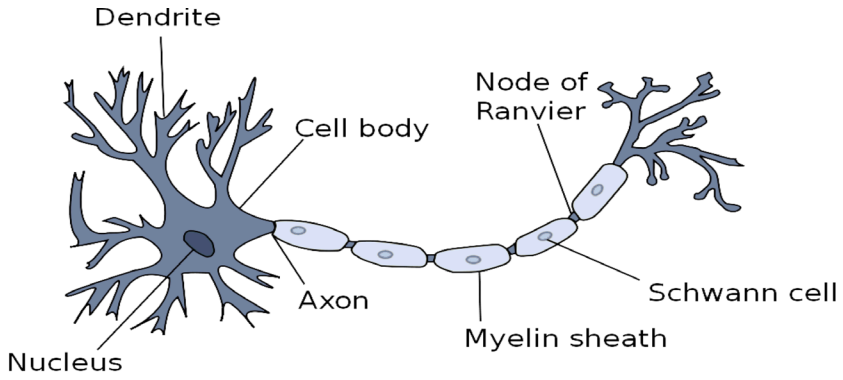
Machine learning ~= Monkey see, monkey do

{Backprogation ML-Leer-regel nodig om een NN te laten leren}

Machinaal Leren [ML]

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

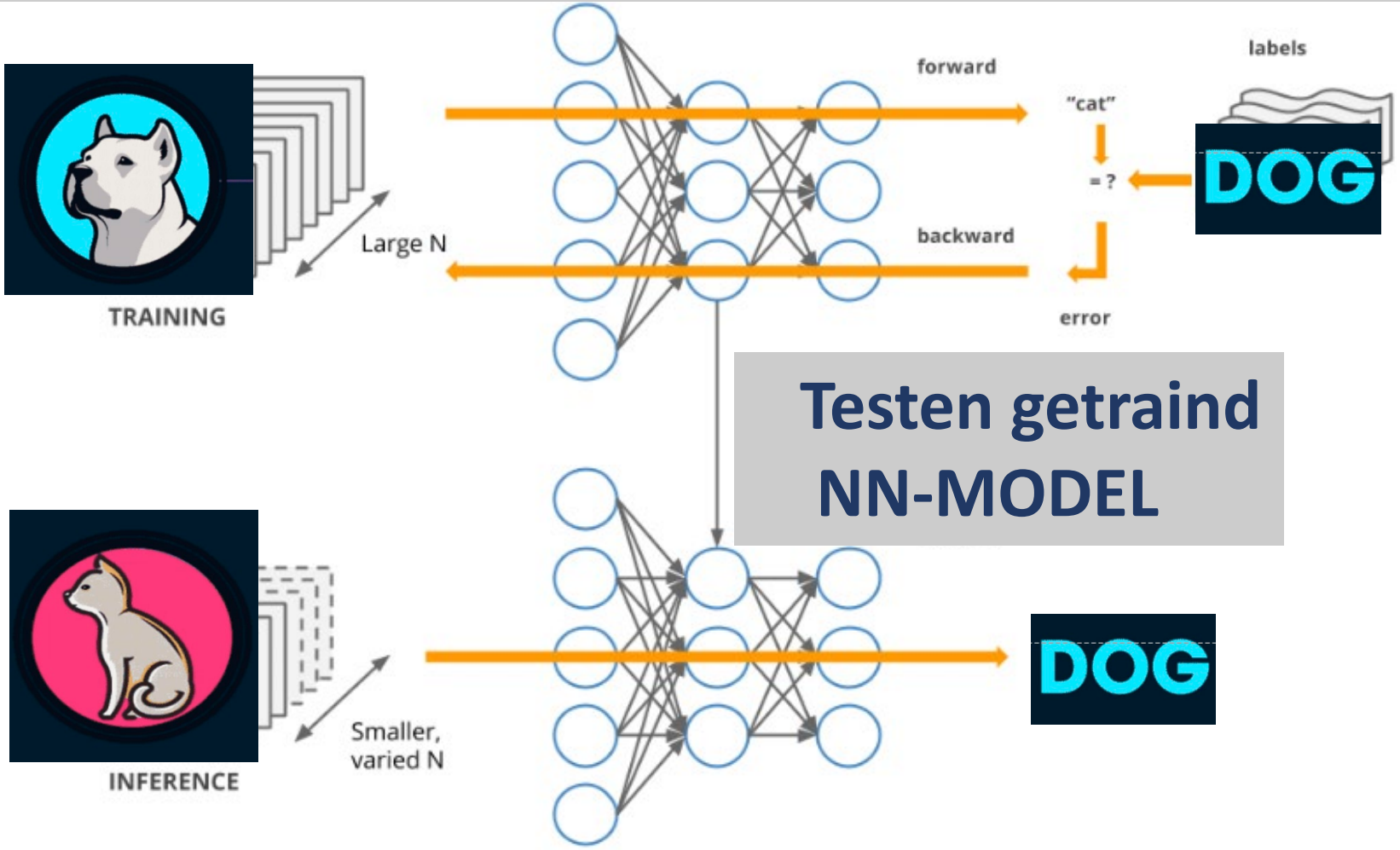




{NN Trainen (leren) versus Testen}

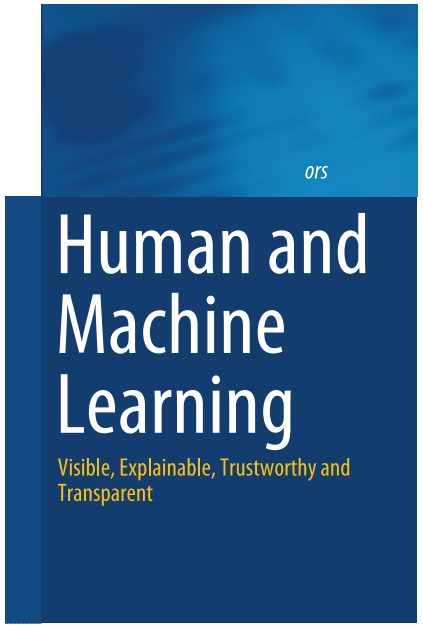
HOGESCHOOL ROTTERDAM

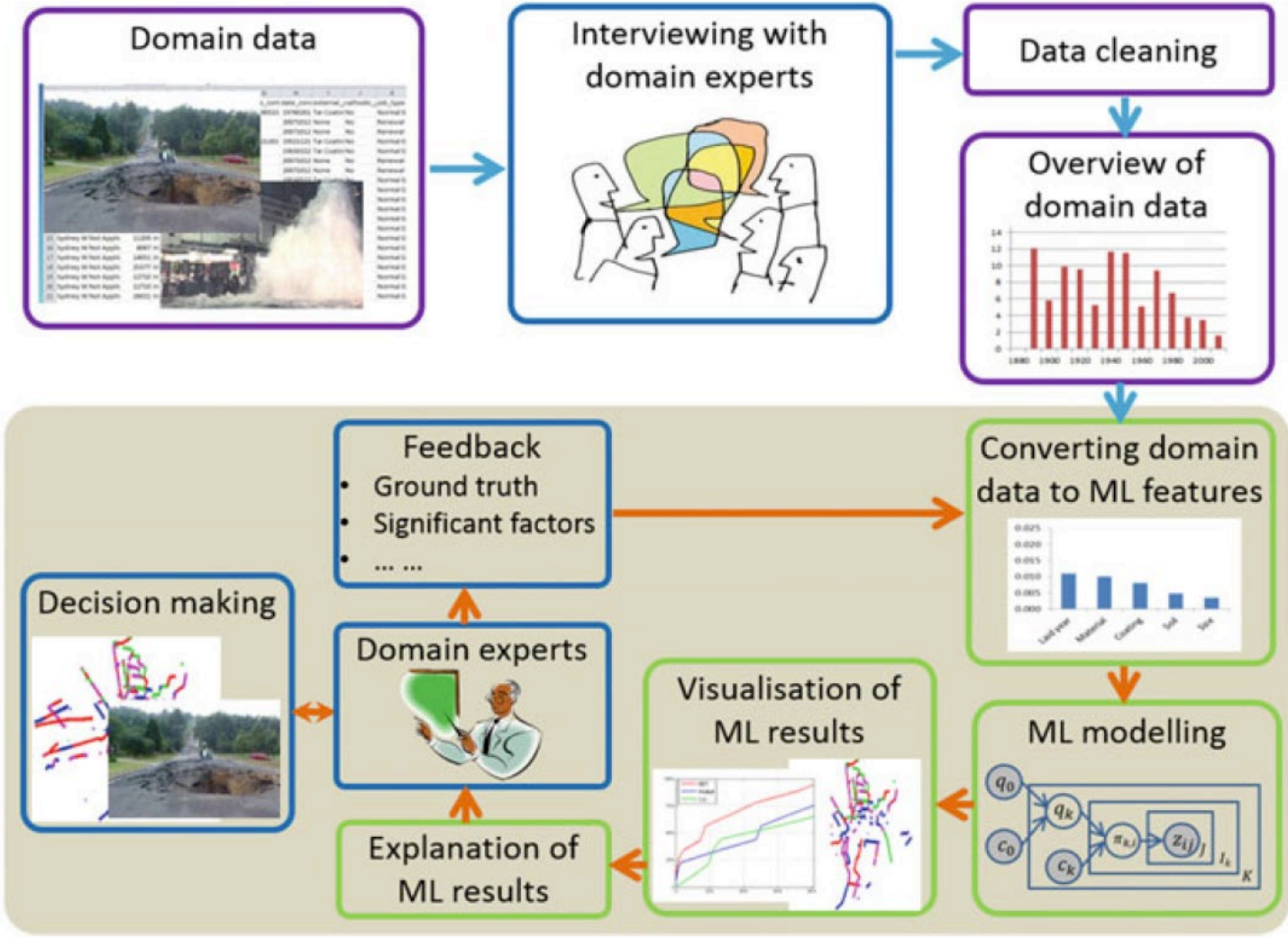
Machinaal Leren [ML] == trainen van het NN



{Modeling: human-in-the-loop}

HOGESCHOOL ROTTERDAM





{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*

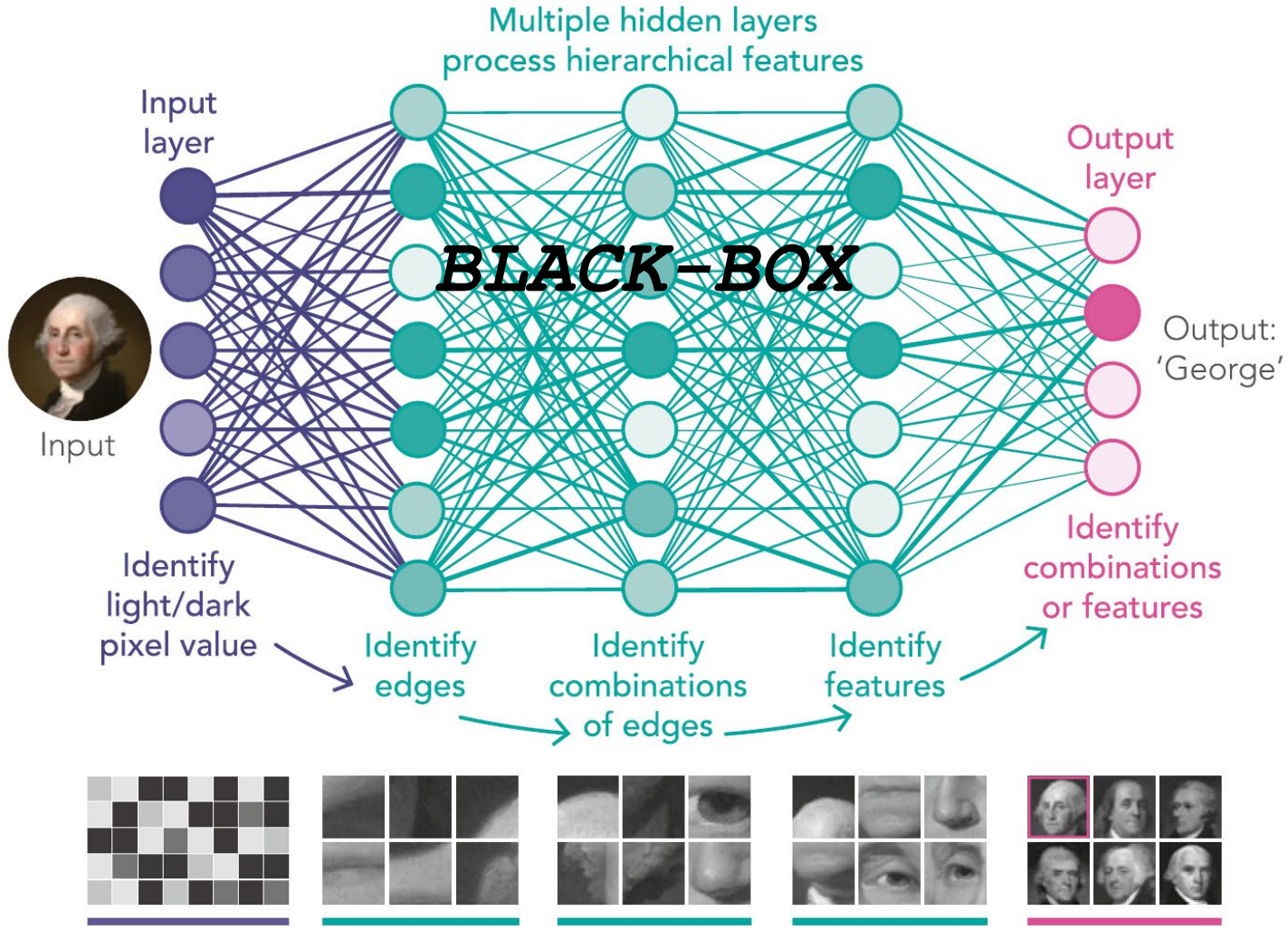
There's no mistaking the image: It's a banana—a big, ripe, bright-yellow banana. Yet the artificial intelligence (AI) identifies it as a toaster, even though it was trained with the same powerful and oft-publicized deep-learning techniques that have produced a white-hot revolution in driverless cars, 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



{Big-data}

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

HOGESCHOOL ROTTERDAM
CREATING

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:

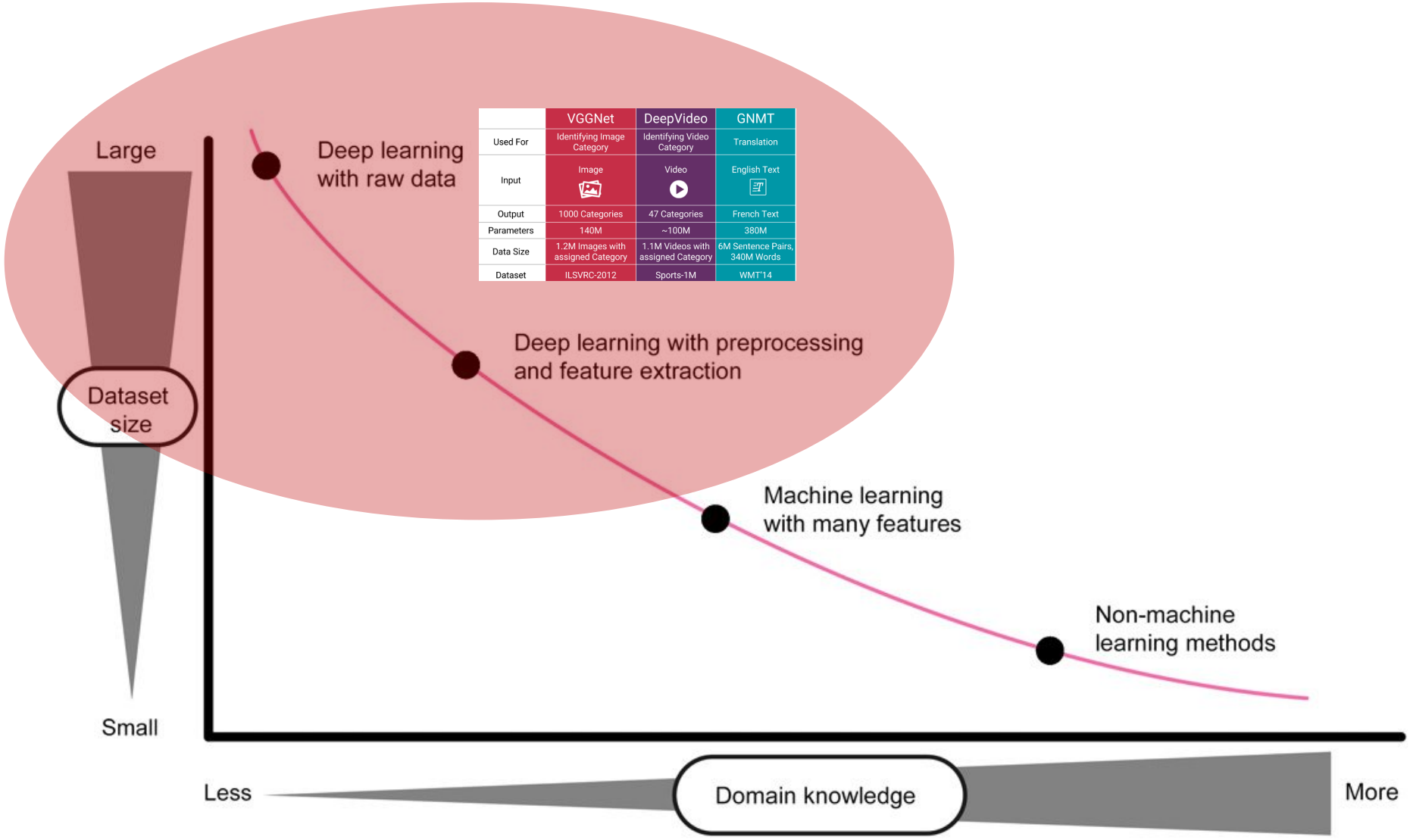
1. We introduce the perspective of viewing *any* explanation of a model's prediction as a model itself, which we term the *explanation model*. This lets us define the class of *additive feature attribution methods* (Section 2), which unifies six current methods.

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

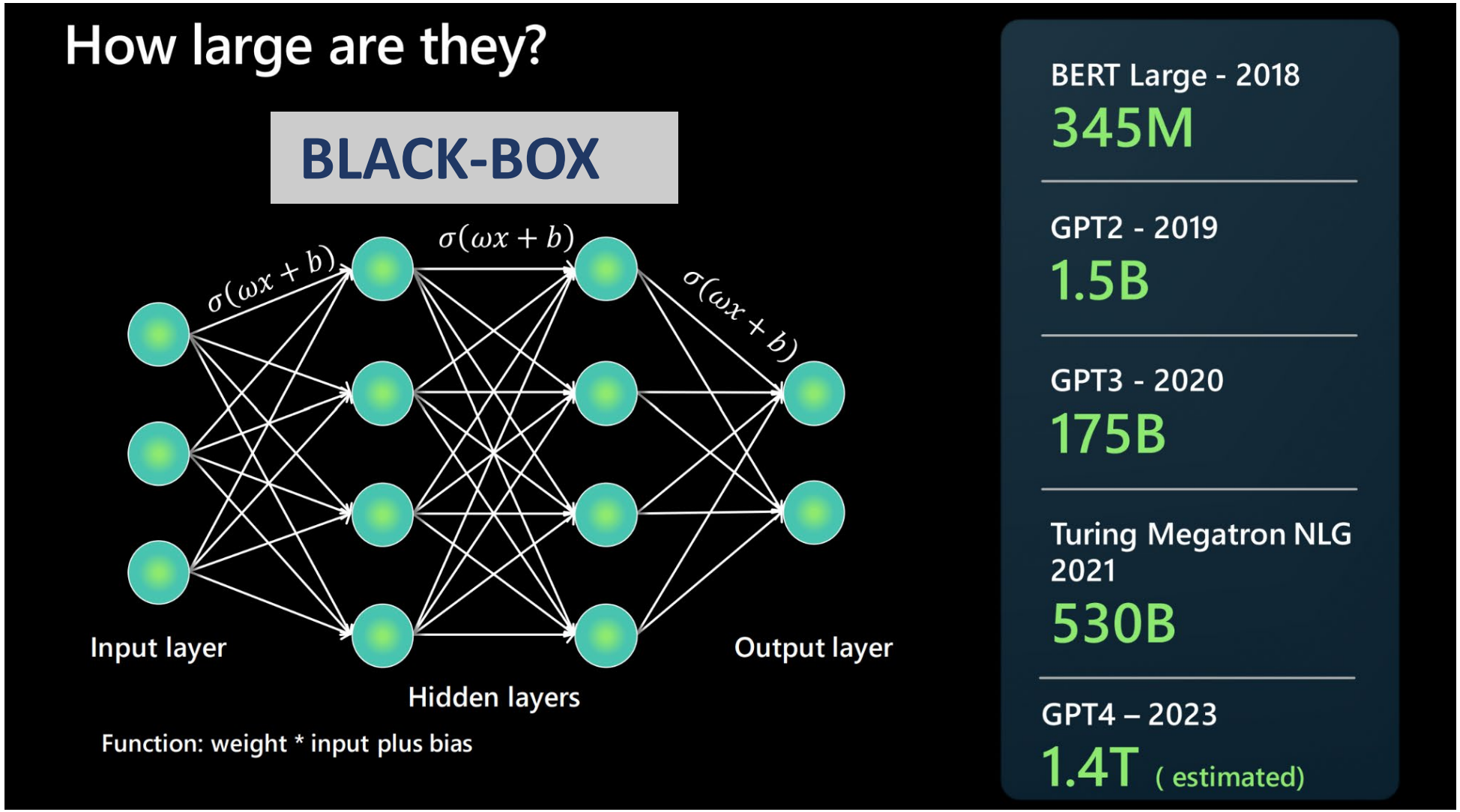
31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

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

	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



{Diep Neuraal Netwerk [NN] ontstaat “vanzelf”}



Het gebruik van publieke Grote Taal Modellen zoals ChatGPT is problematisch

Generatieve AI [GEN-AI]

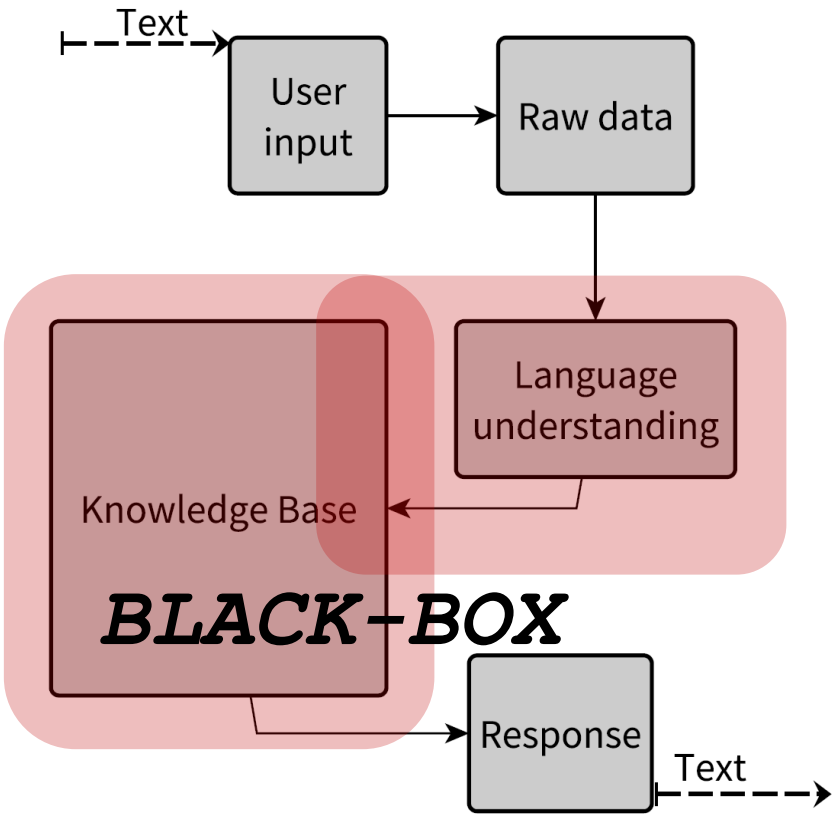
Agenten gebaseerd op Neurale netwerk modellen ---door machinaal leren op basis van bestaande multimodale content
=========>(trainen van het model)
die instaat zijn nieuwe inhoud te creëren,
zoals tekst, afbeeldingen, muziek en/of code.

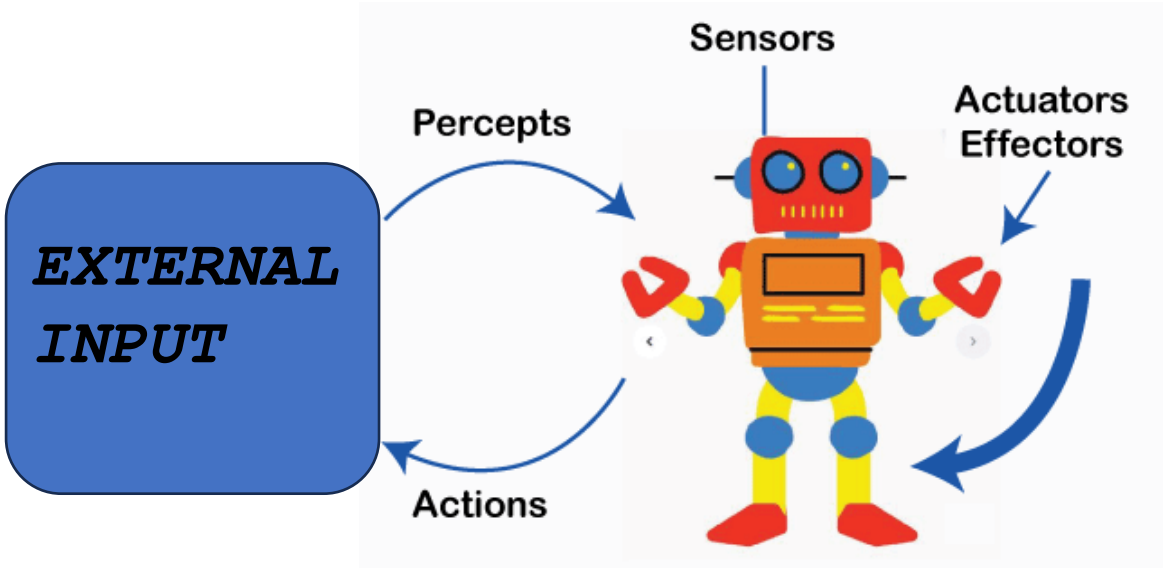
GEN-AI agenten zijn geen informatiedatabases of deterministische informatiezoeksystemen, maar voorspellingssystemen.

GEN-AI maakt dus geen onderscheid tussen goed/fout of waar/nietwaar maar produceert een uitkomst die met grote waarschijnlijkheid kan worden gerelateerd aan de geven input (prompt).

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

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



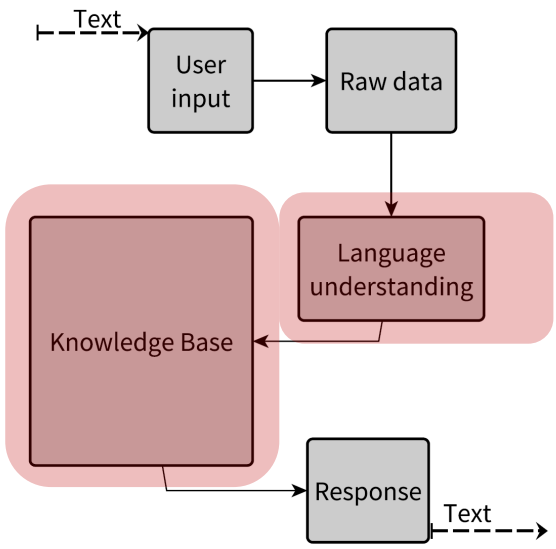


ROBOTIC AI-AGENT

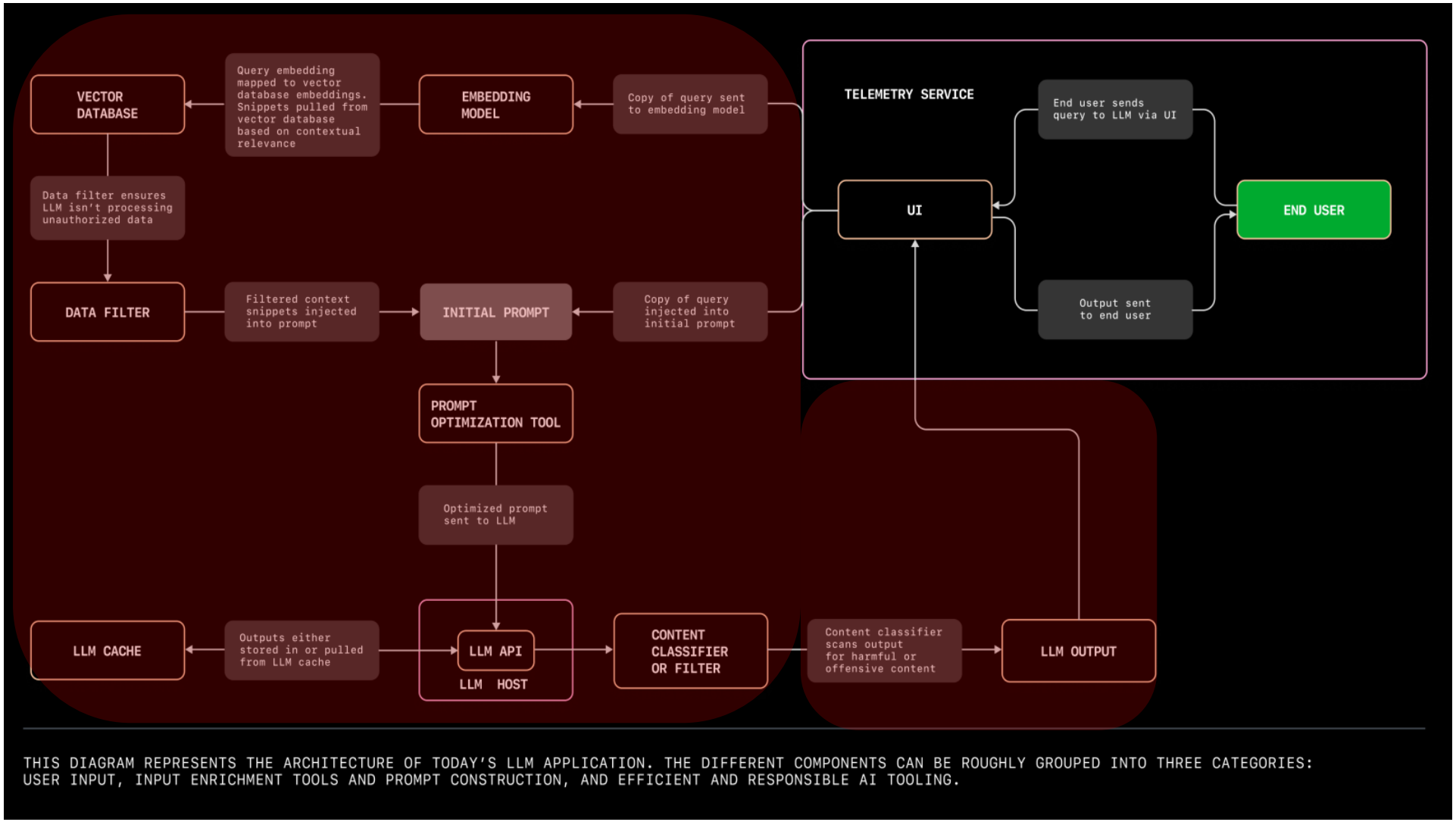
Sánchez-Díaz, X., Ayala-Bastidas, G., Fonseca-Ortiz, P., & Garrido, L. (2018). A knowledge-based methodology for building a conversational chatbot as an intelligent tutor. https://doi.org/10.1007/978-3-030-04497-8_14

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

HOGESCHOOL ROTTERDAM



BLACK-BOX

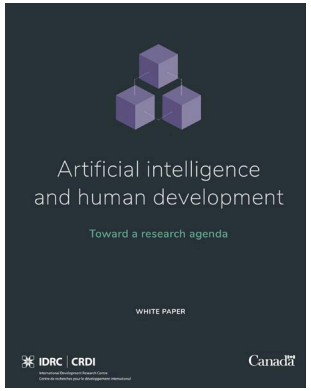


https://blogs.novita.ai/exploring-architectural-structures-and-functional-capacities-of-llms/

{Skewed Data-SETS}

HOGESCHOOL ROTTERDAM

Big-Data is Inherently Skewed



When artificial intelligence learns from humans, it's bad

James Crowder*

Learned bias

ML algorithms are not biased in and of themselves; they *learn* to be biased. This *algorithmic bias* (Danks and London 2017) has received a great deal of attention. It occurs when the learning algorithm is trained on biased datasets and subsequently "accurately" learns the patterns of bias inherent in the data (see, e.g., Caliskan et al. 2017). In some cases, the learned representations within ML algorithms can even exaggerate these biases (Zhao et al. 2017). Algorithmic bias has two sources: incomplete datasets and datasets that represent biased social phenomena.

Incomplete datasets are those that are not representative of the entire range of potential examples. Consequently, an algorithm trained on an incomplete dataset will perform poorly when given an example that falls outside the scope of the available data. For example, a facial recognition algorithm that is not trained on a wide variety of skin colours might not function accurately for faces of all skin tones. Indeed, one AI researcher with dark skin discovered that an otherwise functional facial recognition algorithm failed to recognize her face unless she put on a white mask[45] . Many similar examples have been reported.[46]

When critical decisions are made based on input from an algorithm trained on a database that is not representative of the entire user population, the results can impact health and well-being. Most clinical trials have highly selective criteria that exclude women (especially pregnant women), the elderly, and those with conditions beyond those being studied. Thus, participants tend to be white males.[47] In some cases, the findings of these studies do not generalize well across the broader population, with outcomes potentially compromised for individuals not represented in the research. As a result, AI algorithms trained on this data are

* bit.ly/2HFetvY

Photo by Marius Masalar

{Disparities}

Big Data causes racial & gender disparities

The Race Gap in Speech Recognition Technology

Researchers from Stanford University found that state-of-the-art ASR systems exhibited significant racial and gender disparity—they misunderstand Black speakers twice as often as white speakers. In the paper, titled "Racial Disparities in Automated Speech Recognition," authors ran thousands of audio snippets of white and Black speakers, transcribed from interviews conducted with 42 white speakers and 73 Black speakers, through leading speech-to-text services by Amazon, Apple, Google, IBM, and Microsoft.

The results suggest that, on average, systems made 19 errors every hundred words for white speakers and 35 errors for Black speakers—nearly twice as many. Moreover, the systems performed particularly poorly for Black men, with more than 40 errors for every hundred words (Figure 2.5.3). The breakdown by ASR systems shows that gaps are similar across companies (Figure 2.5.4). This research emphasizes the importance of addressing the bias of AI technologies and ensuring equity as they become mature and deployed.

TESTINGS on LEADING SPEECH-to-TEXT SERVICES: WORD ERROR RATE by RACE and GENDER, 2019

Source: Koenecke et al., 2020 | Chart: 2021 AI Index Report

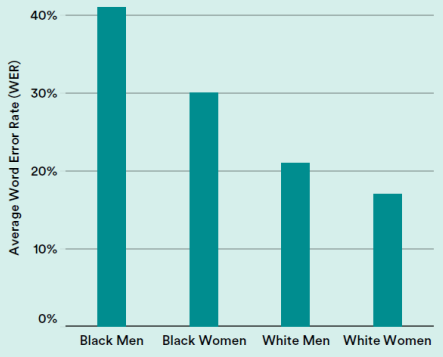


Figure 2.5.3

TESTINGS on LEADING SPEECH-to-TEXT SERVICES: WORD ERROR RATE by SERVICE and RACE, 2019

Source: Koenecke et al., 2020 | Chart: 2021 AI Index Report

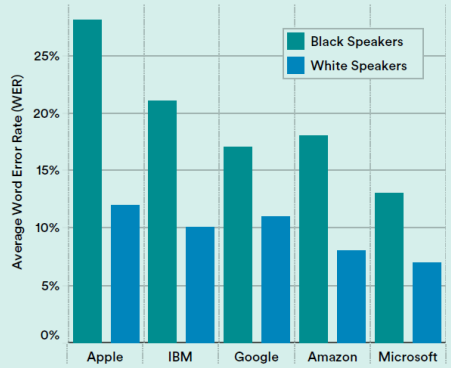


Figure 2.5.4

{Augmentation}

HOGESCHOOL ROTTERDAM

Big Data that is *not* augmented causes Overfitting

SURVEY PAPER Open Access

A survey on Image Data Augmentation for Deep Learning

Connor Shorten* and Taghi M. Khoshgoftaar

*Correspondence:
cshorten2015@fau.edu
Department of Computer and Electrical Engineering and Computer Science, Florida Atlantic University, Boca Raton, USA

Abstract

Deep convolutional neural networks have performed remarkably well on many Computer Vision tasks. However, these networks are heavily reliant on big data to avoid overfitting. Overfitting refers to the phenomenon when a network learns a function with very high variance such as to perfectly model the training data. Unfortunately, many application domains do not have access to big data, such as medical image analysis. This survey focuses on Data Augmentation, a data-space solution to the problem of limited data. Data Augmentation encompasses a suite of techniques that enhance the size and quality of training datasets such that better Deep Learning models can be built using them. The image augmentation algorithms discussed in this survey include geometric transformations, color space augmentations, kernel filters, mixing images, random erasing, feature space augmentation, adversarial training, generative adversarial networks, neural style transfer, and meta-learning. The application of augmentation methods based on GANs are heavily covered in this survey. In addition to augmentation techniques, this paper will briefly discuss other characteristics of Data Augmentation such as test-time augmentation, resolution impact, final dataset size, and curriculum learning. This survey will present existing methods for Data Augmentation, promising developments, and meta-level decisions for implementing Data Augmentation. Readers will understand how Data Augmentation can improve the performance of their models and expand limited datasets to take advantage of the capabilities of big data.

Keywords: Data Augmentation, Big data, Image data, Deep Learning, GANs

Introduction

Deep Learning models have made incredible progress in discriminative tasks. This has been fueled by the advancement of deep network architectures, powerful computation, and access to big data. Deep neural networks have been successfully applied to Computer Vision tasks such as image classification, object detection, and image segmentation thanks to the development of convolutional neural networks (CNNs). These neural networks utilize parameterized, sparsely connected kernels which preserve the spatial characteristics of images. Convolutional layers sequentially downsample the spatial resolution of images while expanding the depth of their feature maps. This series of convolutional transformations can create much lower-dimensional and more useful representations of images than what could possibly be hand-crafted. The success of CNNs has spiked interest and optimism in applying Deep Learning to Computer Vision tasks.

Springer Open

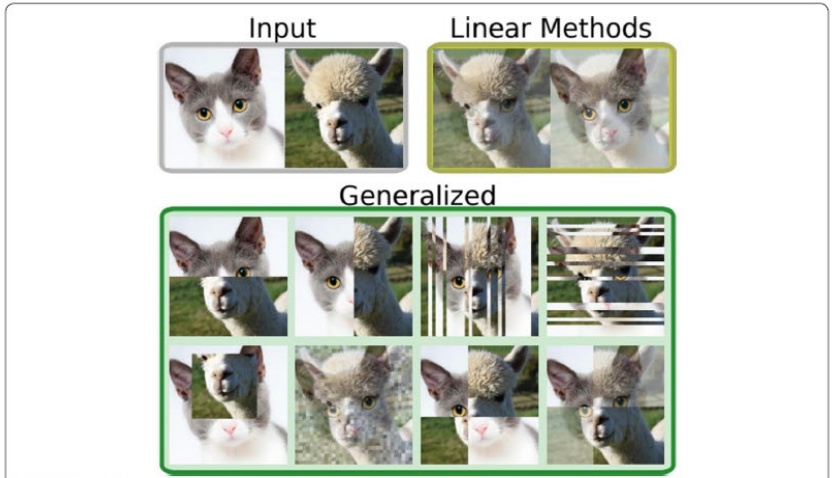
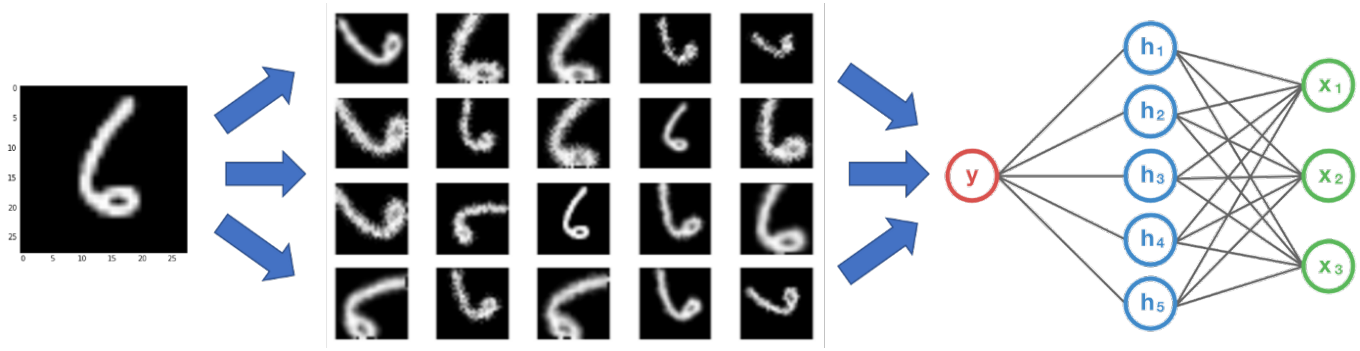
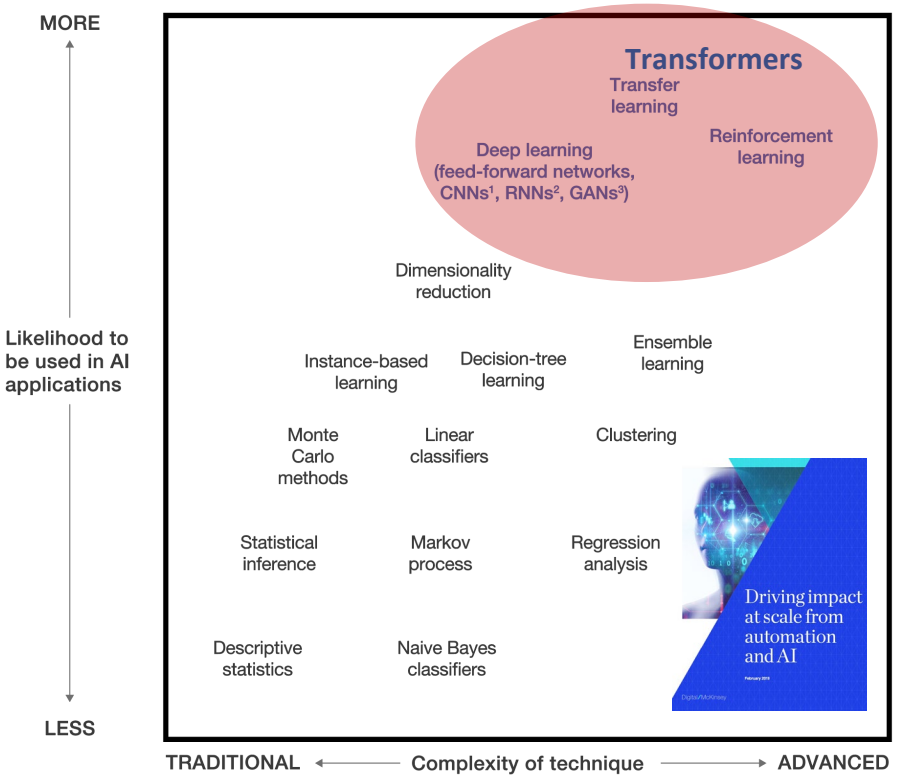


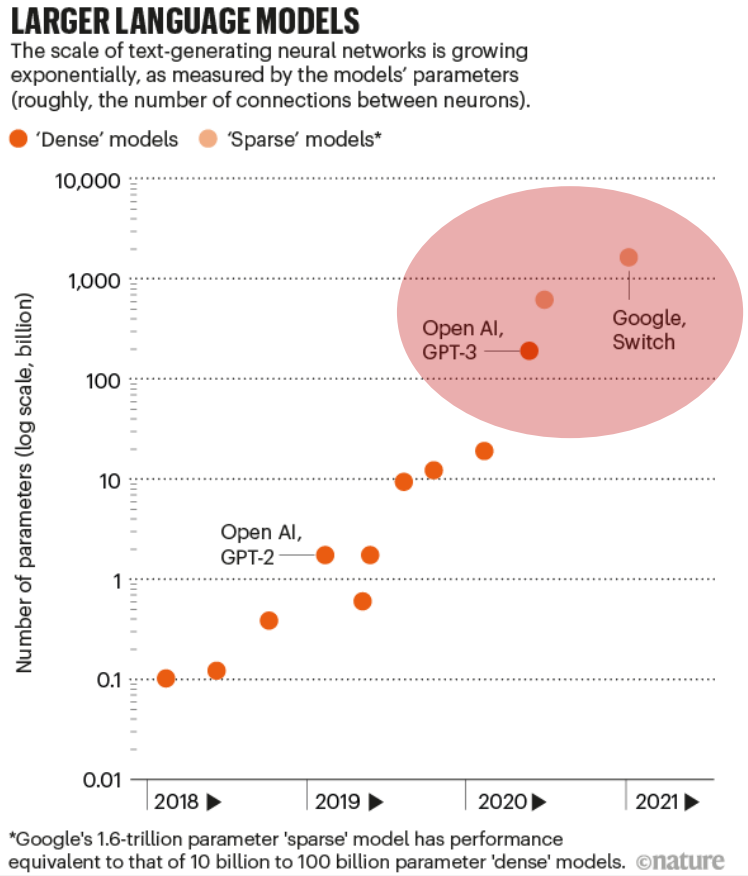
Fig. 9 Non-linearly mixing images [66]



{large scale}

Only very large scale {DNNs} are useful [can compete with human performance]

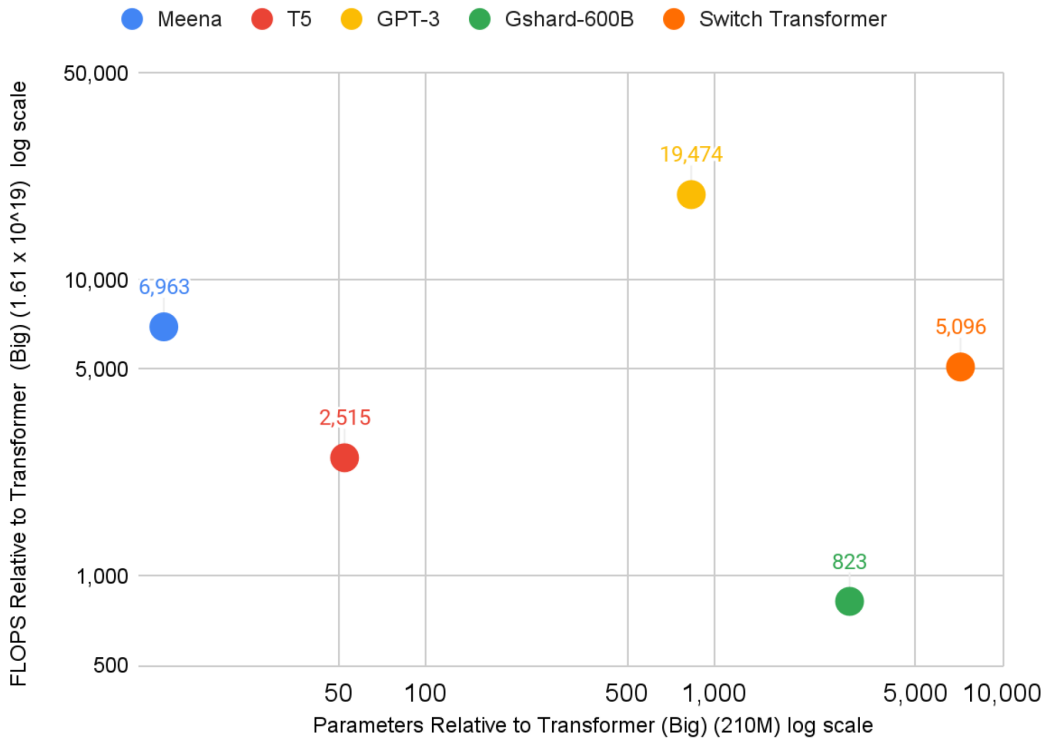


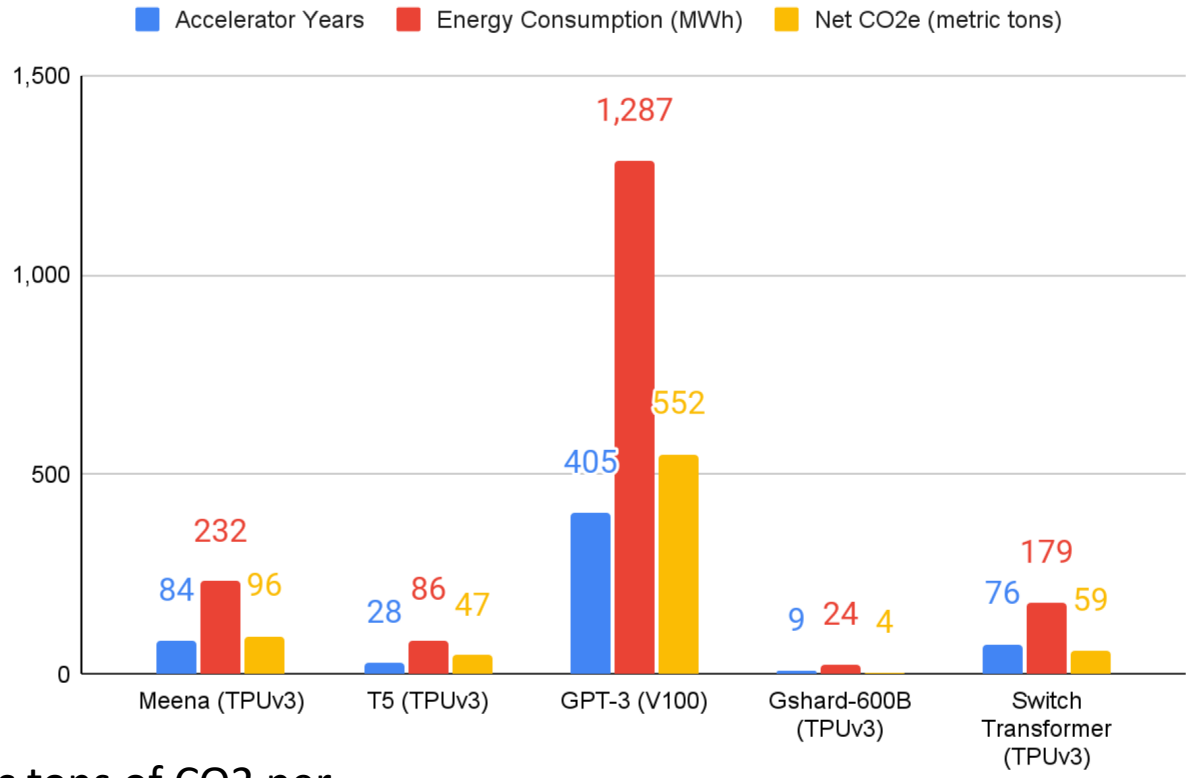


https://www.nature.com/articles/d41586-021-00530-0

{CO2 foot-print}

Training large scale transformer {DNNs} produce massive Carbon Emissions





As of 2007, the average U.S. household emits 20 metric tons of CO2 per year. In comparison to a world average of 4 tons.

Carbon Footprint_CSS09-05_e2021.pdf (umich.edu)

https://arxiv.org/ftp/arxiv/papers/2104/2104.10350.pdf

{computational unsustainability}

HOGESCHOOL ROTTERDAM

The scale of state-of-the-art {SOTA} –near human level– DNNs
– combined with a blind Brute-Force implementation + post-hoc analysis –
is becoming more and more
computationally unsustainable,
even to the point
that hypernetworks are employed
to help humans to make DDNs work.

2110.13100v1.pdf (arxiv.org)

https://paperswithcode.com/sota/

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

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

Generatieve-AI

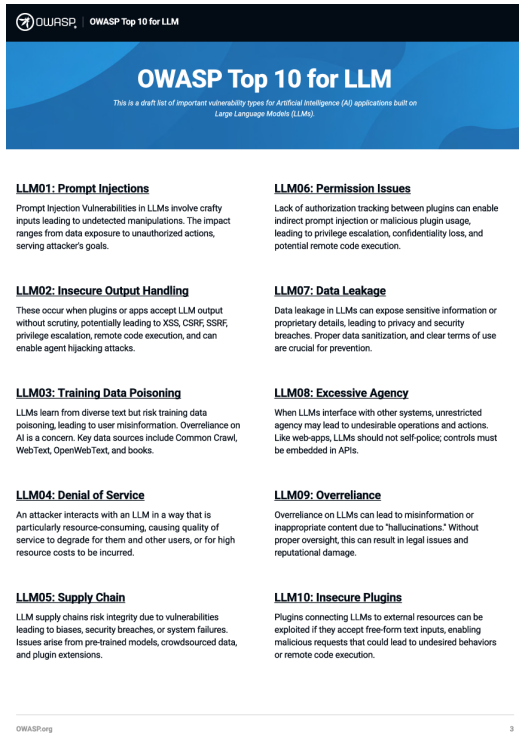
Hoe pas je Zero Trust principes toe op het gebruik van publieke chatbots zoals Chat-GPT?

Human-in-the Loop

“Cyber Security”
beveiligingsrichtlijnen
voor LLMs:

{Open Web Application Security Project (OWASP) }

De **Open Web Application Security Project (OWASP)** is een open-source project waar beveiligingsexperts continu aan werken, om de lijst actueel te houden met de meest voorkomende veiligheidslekken.



https://owasp.org/www-project-top-10-for-large-language-model-applications/assets/PDF/OWASP-Top-10-for-LLMs-2023-v05.pdf

HOGESCHOOL ROTTERDAM

{Bekende LLM “cyber” veiligheidsrisico’s}

Volgen OWASP gebruiken onderwijsinstellingen veelvuldig op LLM’s gebaseerde public Chabots, maar lijken daarbij niet een duidelijk security-protocol te volgen.

Het belang van een **“human-in-the-loop”** is niet te onderschatten, zoals door toestemming voor bepaalde acties afhankelijk te maken van een menselijke goedkeuring.

Het naleven van “**zero-trust principes**” en “**least-privilege**” toegang is eveneens broodnodig. Een LLM moet behandeld worden als een untrusted user om de invloed van een kwaadwillende zoveel mogelijk in te perken.

LLM01: Prompt Injections

Prompt Injection Vulnerabilities in LLMs involve crafty inputs leading to undetected manipulations. The impact ranges from data exposure to unauthorized actions, serving attacker's goals.

LLM02: Insecure Output Handling

These occur when plugins or apps accept LLM output without scrutiny, potentially leading to XSS, CSRF, SSRF, privilege escalation, remote code execution, and can enable agent hijacking attacks.

LLM03: Training Data Poisoning

LLMs learn from diverse text but risk training data poisoning, leading to user misinformation. Overreliance on AI is a concern. Key data sources include Common Crawl, WebText, OpenWebText, and books.

LLM04: Denial of Service

An attacker interacts with an LLM in a way that is particularly resource-consuming, causing quality of service to degrade for them and other users, or for high resource costs to be incurred.

LLM05: Supply Chain

LLM supply chains risk integrity due to vulnerabilities leading to biases, security breaches, or system failures. Issues arise from pre-trained models, crowdsourced data, and plugin extensions.

LLM06: Permission Issues

Lack of authorization tracking between plugins can enable indirect prompt injection or malicious plugin usage, leading to privilege escalation, confidentiality loss, and potential remote code execution.

LLM07: Data Leakage

Data leakage in LLMs can expose sensitive information or proprietary details, leading to privacy and security breaches. Proper data sanitization, and clear terms of use are crucial for prevention.

LLM08: Excessive Agency

When LLMs interface with other systems, unrestricted agency may lead to undesirable operations and actions. Like web-apps, LLMs should not self-police; controls must be embedded in APIs.

LLM09: Overreliance

Overreliance on LLMs can lead to misinformation or inappropriate content due to "hallucinations." Without proper oversight, this can result in legal issues and reputational damage.

LLM10: Insecure Plugins

Plugins connecting LLMs to external resources can be exploited if they accept free-form text inputs, enabling malicious requests that could lead to undesired behaviors or remote code execution.

https://owasp.org/www-project-top-10-for-large-language-model-applications/assets/PDF/OWASP-Top-10-for-LLMs-2023-v05.pdf

HOGESCHOOL ROTTERDAM

{LLM01: ONVEILIGE PROMPT (“prompt injection”)}

De input bepaalt wat een LLM produceert. Vaak zijn deze resultaten op voorhand lastig te voorspellen.
Toch kunnen ontwikkelaars de resultaten verfijnen met betere training en sleutelwerk aan de parameters.
Bij publieke chatbots zoals ChatGPT en Google Bard is bekend dat ze vrij snel voorzichtig worden als je ze gedurfde vragen stelt.
In een professionele context gaat het alleen om meer dan lolligheden, en kan een interne LLM bedrijfsgeheimen in huis hebben.
Hackers kunnen prompts manipuleren, waardoor het systeem de intenties van de aanvaller uitvoert (denk aan “role playing”).
Dit kan leiden tot extractie van gegevens, social engineering en andere problemen.

Kwetsbaarheid (“Jailbeak”):
OWASP spreekt over directe en indirecte prompt injections. In het eerste geval karakteriseert men deze aanvalstactiek als een “jailbreak” voor een LLM: een kwaadwillende heeft in dat geval de onderliggende systemprompt ontdekt of herschreven. Daardoor kan men bij een aanval wellicht bij gevoelige data stores komen waar de LLM op rust.
Het indirecte scenario komt voor als een LLM via een externe bron aanspreekbaar warbij de aanvaller een prompt injection uitvoert om de conversatie-context te kapen. Vervolgens opereert de LLM als een “confused deputy”, in de woorden van OWASP.
Dat houdt in dat het taken kan uitvoeren die het normaal gesproken niet zou mogen doen voor een gebruiker.

Preventieve maatregel:
Privilegebeheer wordt afgedwongen voor LLM-toegang tot achterliggende systemen
Validatie door een mens wordt afgedwongen voor beslisfunctionaliteit (“human-in-the-loop”)
Externe inhoud is gescheiden van gebruikersprompts

HOGESCHOOL ROTTERDAM

{LLM07: Onbedoelde openbaarmaking van gevoelige informatie}

Hoe goed een LLM ook getraind en getest is,
het kan zo zijn dat een generatief AI-model op een onverwachte manier informatie door kan spelen **(“data-leakage”).**
Een toevallige prompt kan opeens gevoelige data of algoritmes prijsgeven,
zelfs als de eindgebruiker geen kwaadwillende gedachtes erop nahoudt.
LLM-toepassingen kunnen onbedoeld gevoelige informatie, gepatenteerde algoritmen of vertrouwelijke gegevens openbaar maken,
wat kan leiden tot onbevoegde toegang, diefstal van intellectueel eigendom en privacyschendingen.

Kwetsbaarheid:
Onbedoelde bekendmaking van vertrouwelijke informatie (onzorgvuldig/ondoordacht handelen)
Gecompliceerde (van het web gekopieerde) prompts die worden gebruikt om invoerfilters te omzeilen en gevoelige gegevens te onthullen

Preventieve maatregel:
Validatie en opschoning (pseudonimiseren + de-identificatie) van invoer op basis van geautoriseerde protocollen/use-cases
OWASP raadt aan om consumenten van LLM-applicaties bewust te maken van een veilige omgang met AI-modellen.

Autoriteit Persoonsgegevens
https://www.autoriteitpersoonsgegevens.nl/themas/beveiliging/beveiliging-van-persoonsgegevens/gegevens-pseudonimiseren

Making Qualitative Data Reusable - A Short Guidebook For Researchers And Data Stewards Working With Qualitative Data
https://zenodo.org/doi/10.5281/zenodo.7777518

HOGESCHOOL ROTTERDAM

{LLM08: Te veel vrijheid “Excessive Agency”}

Een LLM-systeem kan erg effectief zijn als het toegang heeft tot andere systeem.
Zo kan een AI-assistent in opdracht van een gebruiker in een andere applicatie duiken om iets uit te zoeken.
Wie bij het opzetten van deze functionaliteit onoplettend is, kan de LLM toestaan om ongewenste acties uit te voeren.

Deze vorm van ***‘agency’ of autonomie*** kan bijvoorbeeld ruimte geven om shell commands uit te voeren of om een database te updaten, zoals een gewone gebruiker van de andere app zou kunnen.
Met te veel permissies kan dit grote gevolgen hebben.

Kwetsbaarheid:
Onbedoelde bekendmaking van vertrouwelijke informatie (onzorgvuldig/ondoordacht handelen)
Gecompliceerde (van het web gekopieerde) prompts die worden gebruikt om invoerfilters te omzeilen en gevoelige gegevens te onthullen

Preventieve maatregel:
Legt vast waartoe een LLM toegang toe heeft/mag hebben.
Beschrijf verificatiestappen / authorisatie-process om van een ChatBot gebruik te mogen maken.
Wederom is het nuttig om iemand in real-time te laten kijken wat een AI-model precies uitvoert, zoals ook doodgewone gebruikers worden gecontroleerd in een IT-omgeving.

HOGESCHOOL ROTTERDAM

{LLM09: te grote afhankelijkheid/vertrouwen “Overreliance”}

LLM’s kunnen veel, maar zijn zoals gezegd in staat om fouten te maken
en produceren altijd een output ook al is deze onzinnig (**“hallucineren”).**
Ze *‘denken’ niet* en zijn uiteindelijk een zeer complexe voorspelmachines.
Dit houdt in dat er in een professionele context erg voorzichtig mee omgegaan moet worden.

LLM’s kunnen in veel gevallen programmeercode produceren, maar ook daar geldt dat er onjuiste of gevaarlijke inhoud in kan zitten.
Het is een bekend probleem dat deze AI-modellen de mist in kunnen gaan, waardoor het belangrijk is dat organisaties regelmatig controleren of de output naar wens is.

Kwetsbaarheid:
Het opvallende aan deze kwetsbaarheid is dat de aanval niet kwaadwillend is: een nietsvermoedende developer kan programmeercode inzetten die men niet genoeg had bekeken om de fouten op te pikken.
LLM geeft onjuiste informatie
LLM genereert onzinnige tekst
Onvoldoende risicocommunicatie van aanbieders van LLM

Preventieve maatregel:
Communiceer de risico's en beperkingen van LLM duidelijk aan eindgebruikers, laat ze workshops volgen.
De resultaten van LLM monitoren.
Triangulatie / cross-checking van LLM-uitvoer met betrouwbare bronnen om de uitvoer op juistheid te controleren

HOGESCHOOL ROTTERDAM

Hoe bouw en test je veilig & betrouwbaar
GEN-AI technologie?

“human-in-the-loop” +
“private endpoint” or
“virtual networks” +
“embedded vector store”

Data, privacy, and security for Azure OpenAI Service

ⓘ Important

Your prompts (inputs) and completions (outputs), your embeddings, and your training data:

- are NOT available to other customers.
- are NOT available to OpenAI.
- are NOT used to improve OpenAI models.
- are NOT used to train, retrain, or improve Azure OpenAI Service foundation models.
- are NOT used to improve any Microsoft or 3rd party products or services without your permission or instruction.
- Your fine-tuned Azure OpenAI models are available exclusively for your use.

The Azure OpenAI Service is operated by Microsoft as an Azure service; Microsoft hosts the OpenAI models in Microsoft's Azure environment and the Service does NOT interact with any services operated by OpenAI (e.g. ChatGPT, or the OpenAI API).

https://learn.microsoft.com/en-us/legal/cognitive-services/openai/data-privacy?tabs=azure-porta

What data does the Azure OpenAI Service process?

What data does the Azure OpenAI Service process?

Azure OpenAI processes the following types of data:

Text prompts, queries and responses submitted by the user via the completions, search, and embeddings operations.

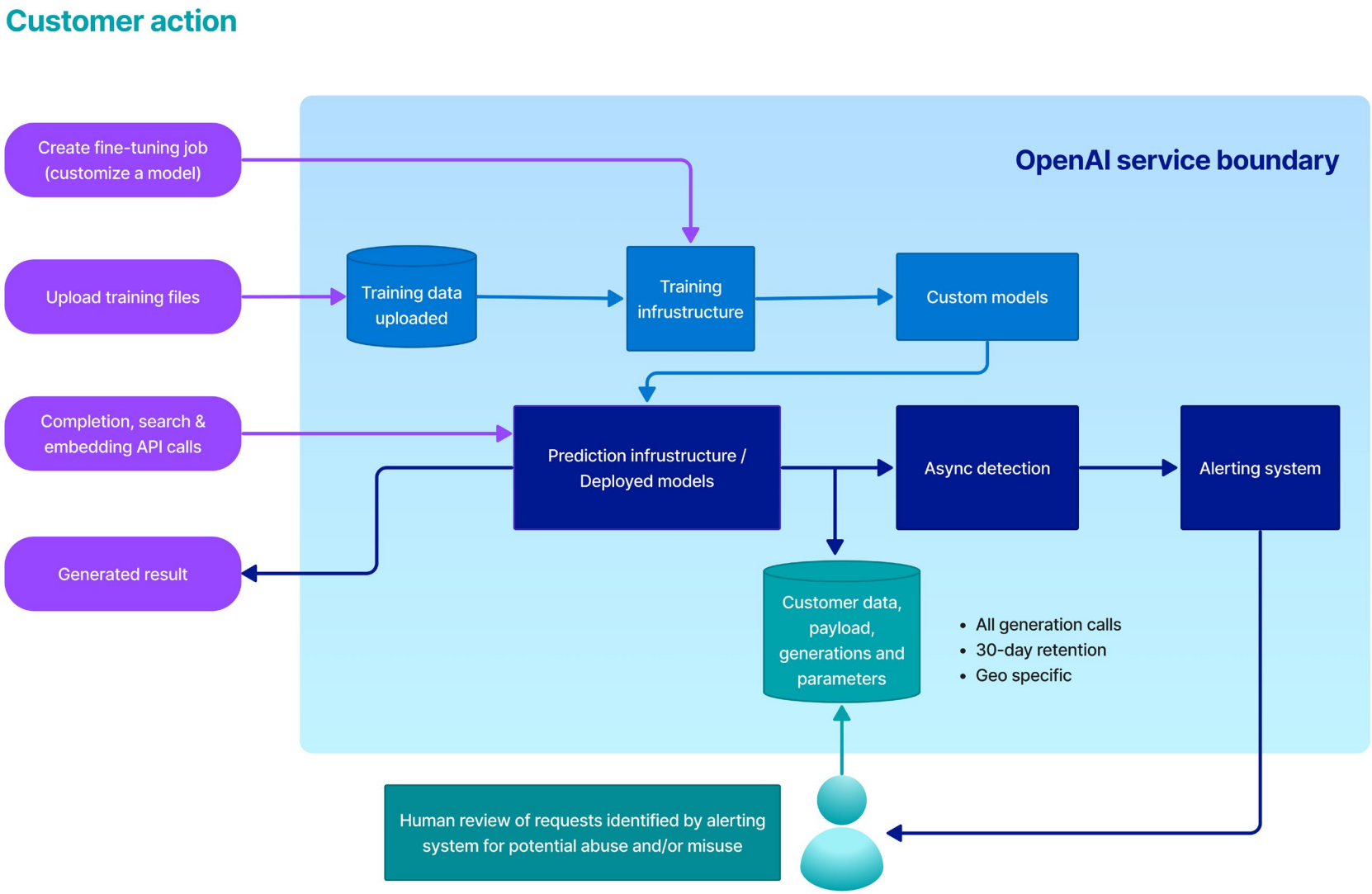
Training & validation data. You can provide your own training data consisting of prompt-completion pairs for the purposes of fine-tuning an OpenAI model.

Results data from training process. After training a fine-tuned model, the service will output meta-data on the job which includes tokens processed and validation scores at each step.

HOGESCHOOL ROTTERDAM

https://learn.microsoft.com/en-us/answers/questions/1183772/how-does-azure-open-ai-handles-user-input-data

What data does the Azure OpenAI Service process?



HOGESCHOOL ROTTERDAM

Azure OpenAI service

OVERVIEW

Build intelligent apps with AI models

Cutting edge models

Quickly develop generative AI experiences with a diverse set of prebuilt and curated models from OpenAI, Meta and beyond.

Try the Azure AI Studio

Data grounding

Trust and transparency

Data, privacy and security

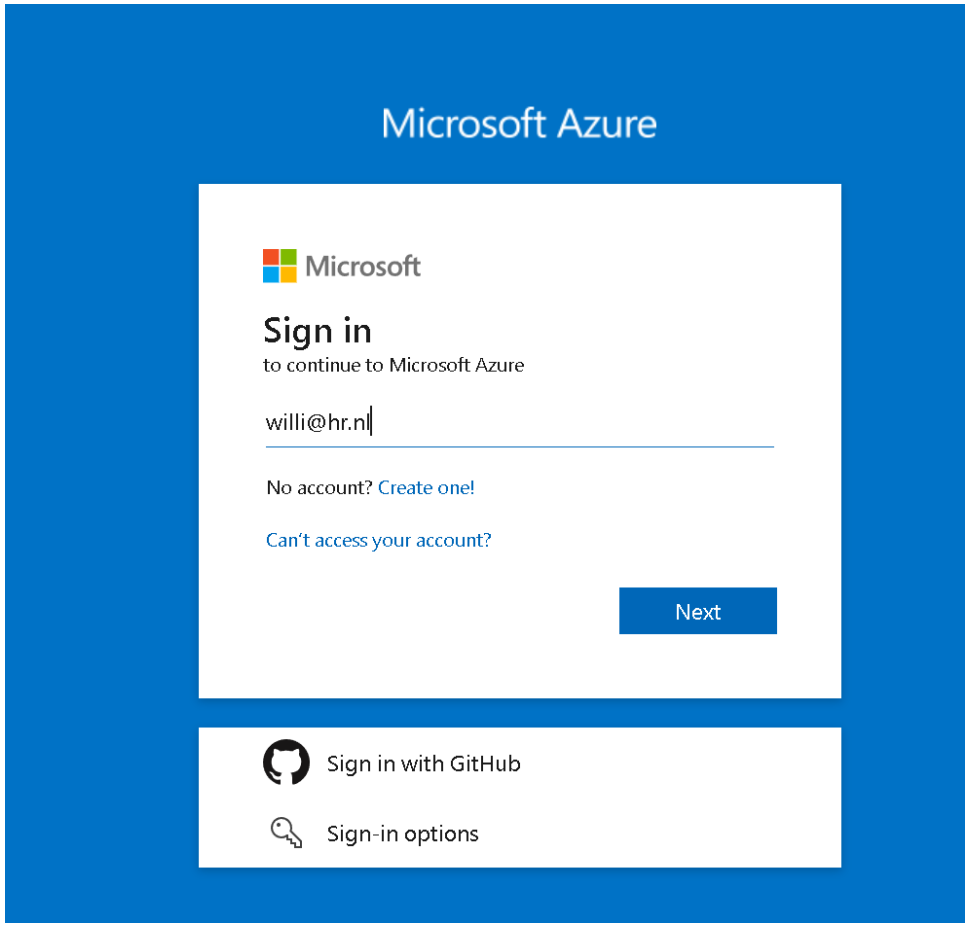
USE CASES

Apply generative AI to a variety of use cases

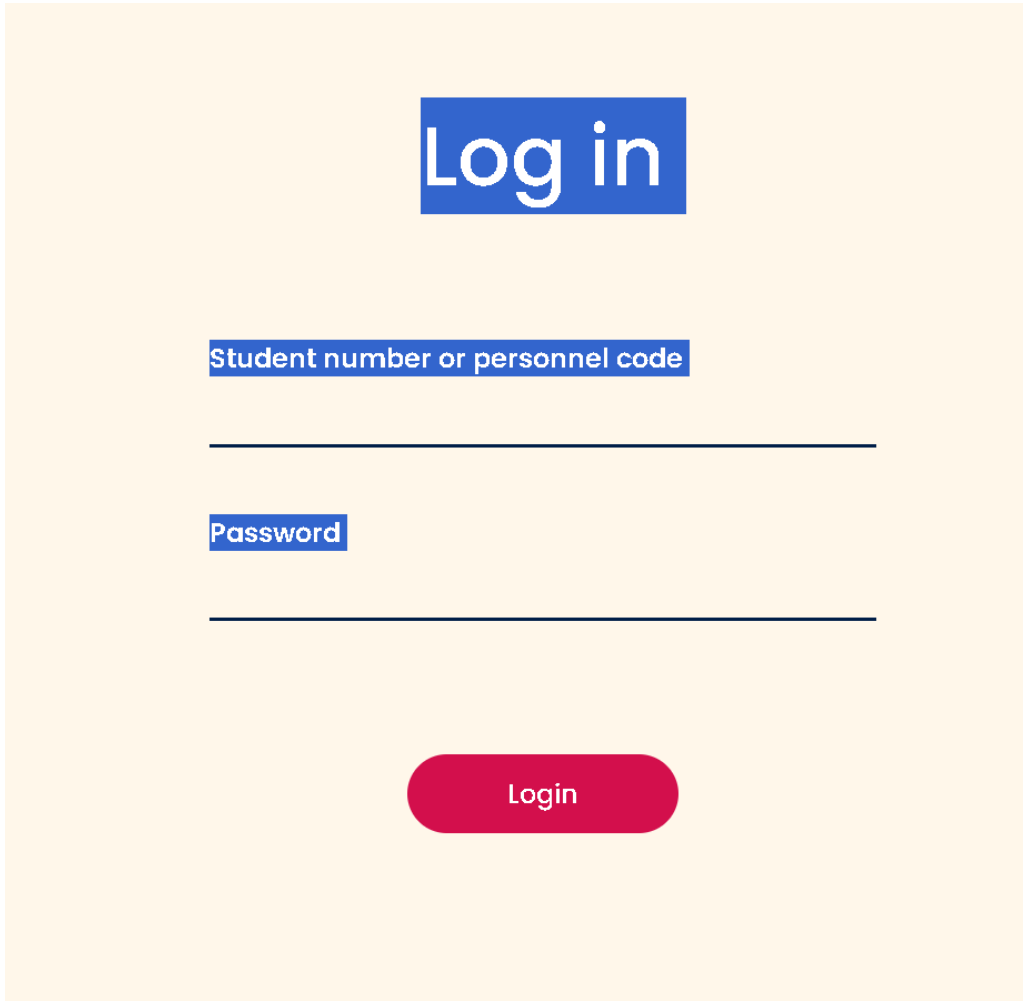
https://azure.microsoft.com/en-us/products/ai-services/openai-service

HOGESCHOOL ROTTERDAM

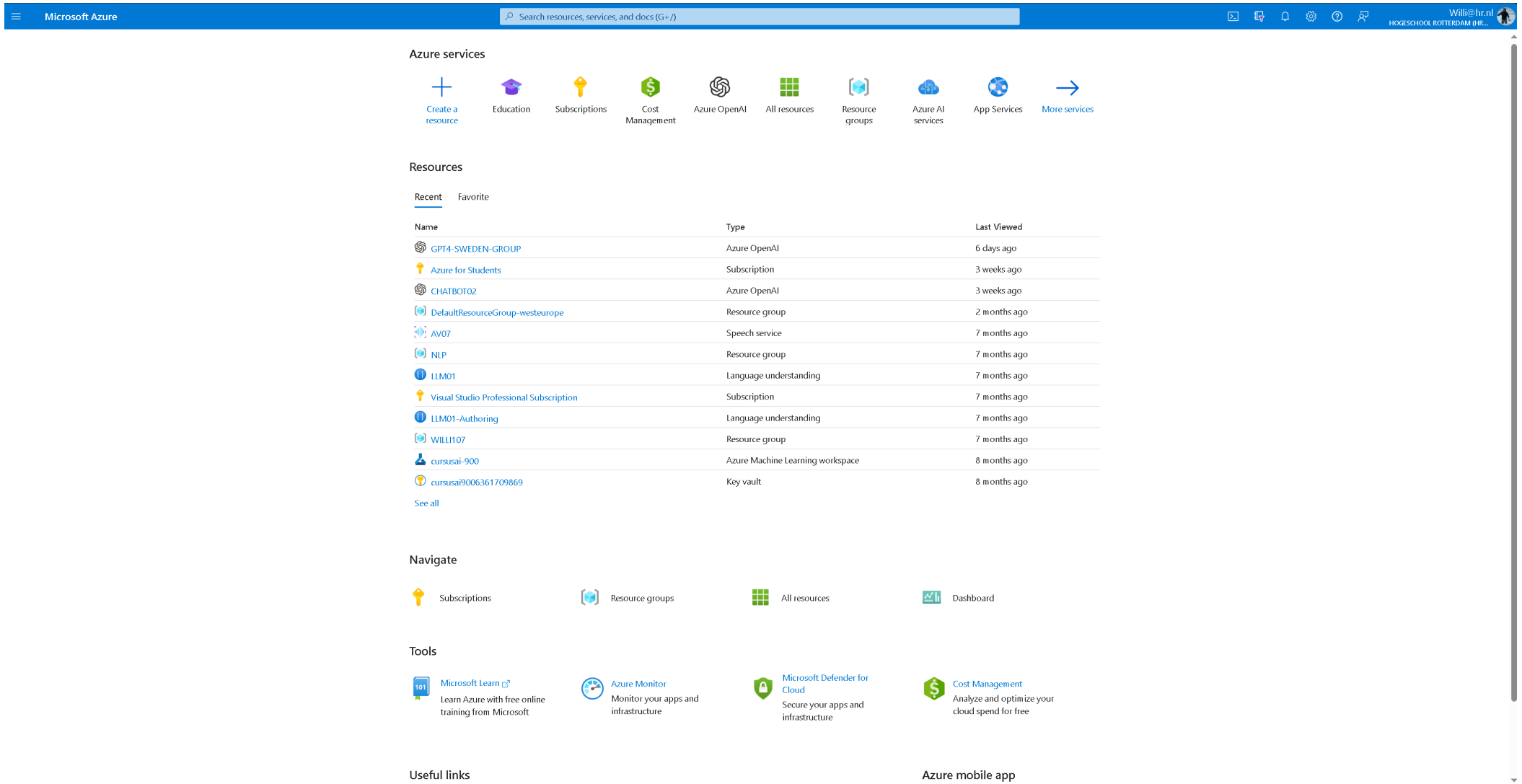
portal.azure.com



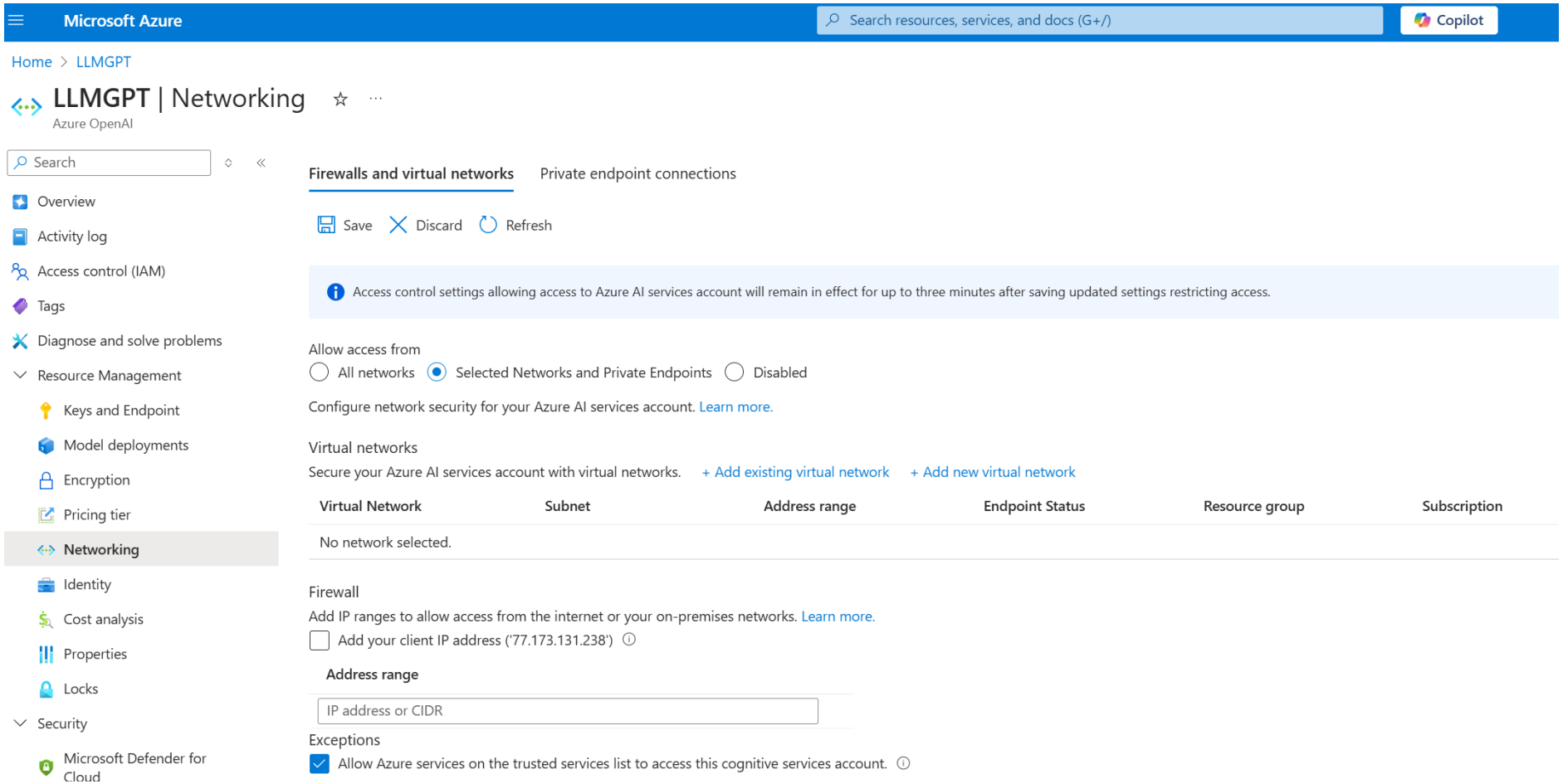
HOGESCHOOL ROTTERDAM



HOGESCHOOL ROTTERDAM



HOGESCHOOL ROTTERDAM



HOGESCHOOL ROTTERDAM

HOGESCHOOL ROTTERDAM

Get started with chat private endpoints for Python

Article • 07/23/2024 • 5 contributors

In this article

This article shows you how to deploy and run the Enterprise chat app sample for Python accessible by private endpoints.

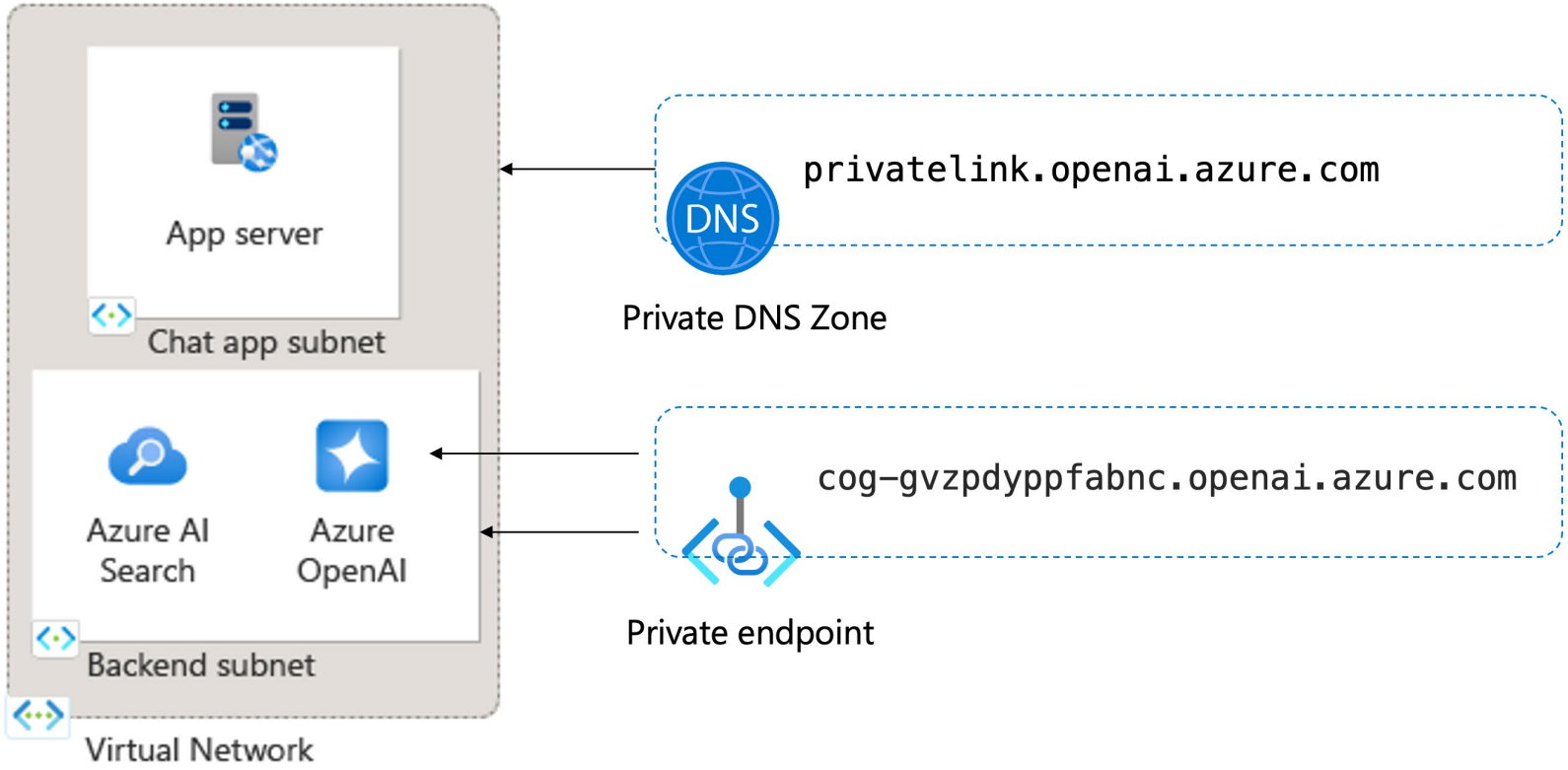
This sample implements a chat app using Python, Azure OpenAI Service, and Retrieval Augmented Generation (RAG) in Azure AI Search to get answers about employee benefits at a fictitious company. The app is seeded with PDF files including the employee handbook, a benefits document and a list of company roles and expectations.

By following the instructions in this article, you will:

- Deploy a chat app to Azure for public access in a web browser.
- Redeploy chat app with private endpoints.

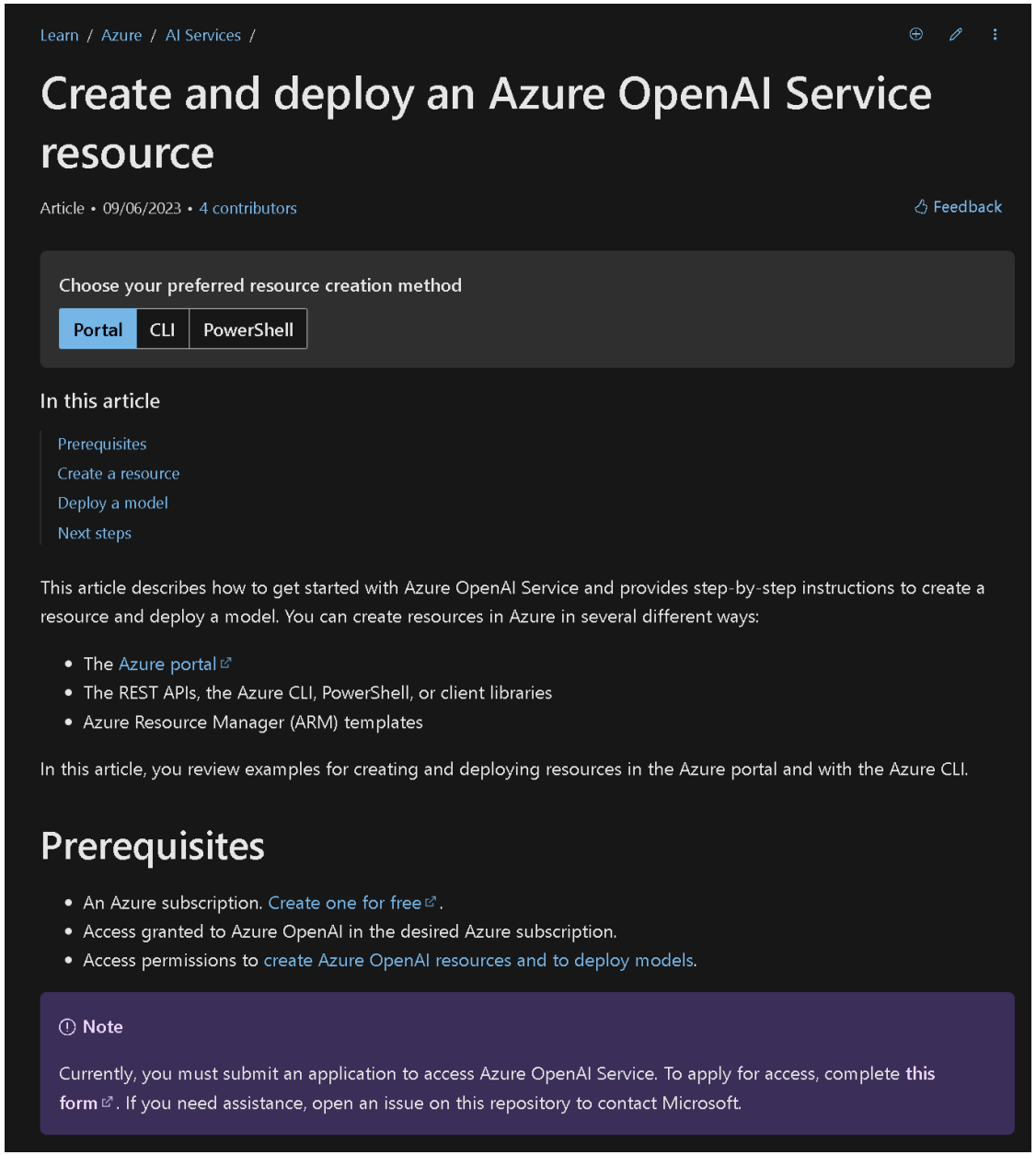
Once you complete this procedure, you can start modifying the new project with your custom code and redeploy, knowing your chat app is accessible only through the private network.

https://learn.microsoft.com/en-us/azure/developer/python/get-started-app-chat-private-endpoint?tabs=visual-studio-code



https://medium.com/@luiz.braz/using-azure-openai-in-a-secure-way-17b9e51cdcaa

https://learn.microsoft.com/en-us/shows/azure-essentials-show/connecting-openai-private-endpoints-across-vnets



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

Learn / Microsoft legal resources /

Limited access for Azure OpenAI Service

Article • 07/23/2024 • 6 contributors

Feedback

In this article

As part of Microsoft's commitment to responsible AI, we have designed and operate Azure OpenAI Service with the intention of protecting the rights of individuals and society and fostering transparent human-computer interaction. For this reason, Azure OpenAI is a Limited Access service, and access and use is subject to eligibility criteria determined by Microsoft. Unless otherwise indicated in the service, all Azure customers are eligible for access to Azure OpenAI models, and all uses consistent with the Product Terms and Code of Conduct are permitted, so customers are not required to submit a registration form unless they are requesting approval to modify content filters and/or abuse monitoring.

Azure OpenAI Service is made available to customers under the terms governing their subscription to Microsoft Azure Services, including Product Terms such as the Universal License Terms applicable to Microsoft Generative AI Services and the product offering terms for Azure OpenAI. Please review these terms carefully as they contain important conditions and obligations governing your use of Azure OpenAI Service.

Registration for modified content filters and/or abuse monitoring

https://learn.microsoft.com/en-us/legal/cognitive-services/openai/limited-access

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

HOGESCHOOL ROTTERDAM

Request Access to Azure OpenAI Service

* Required

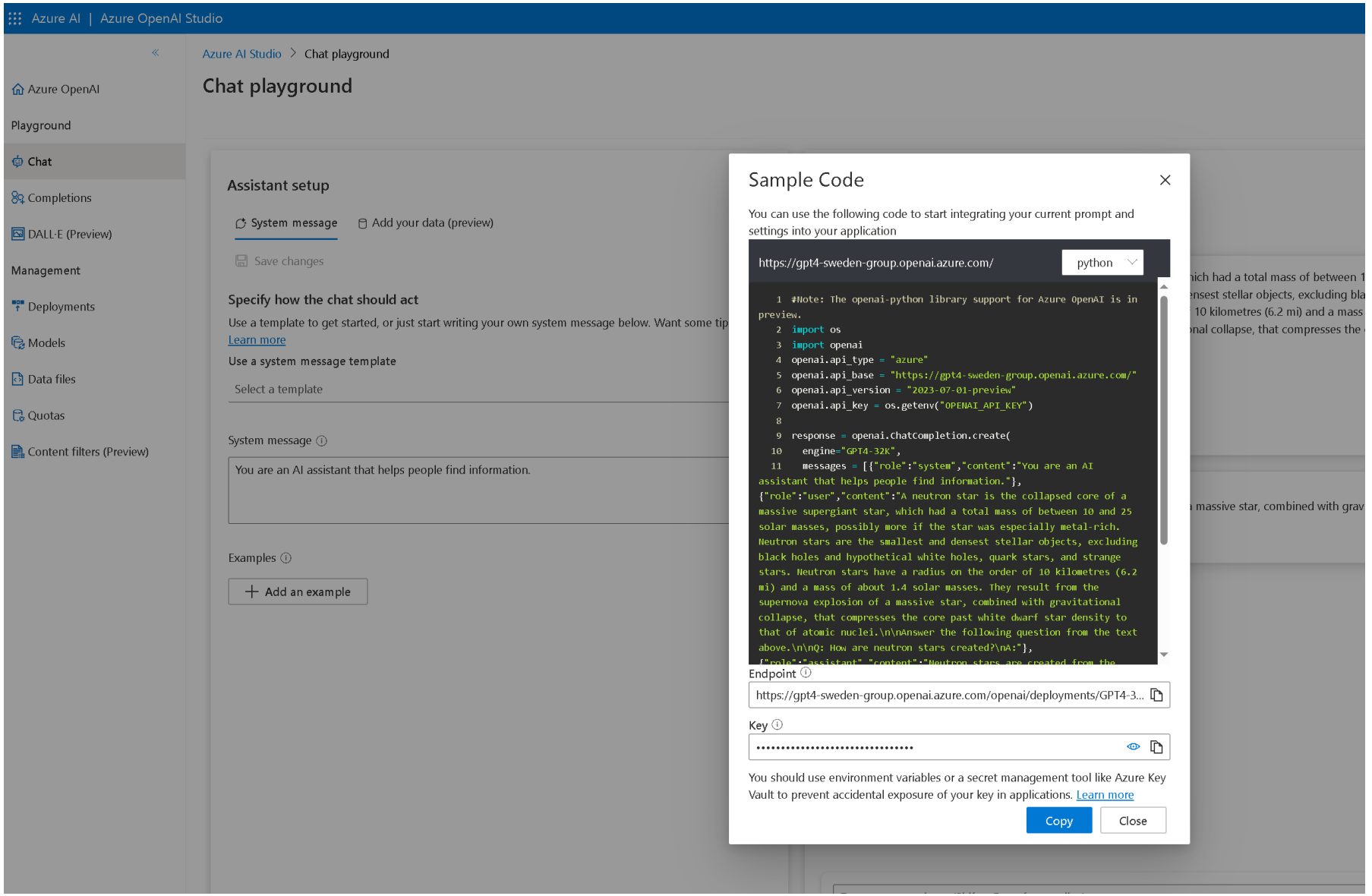
Please read all instructions carefully and complete form as instructed

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



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'])

```

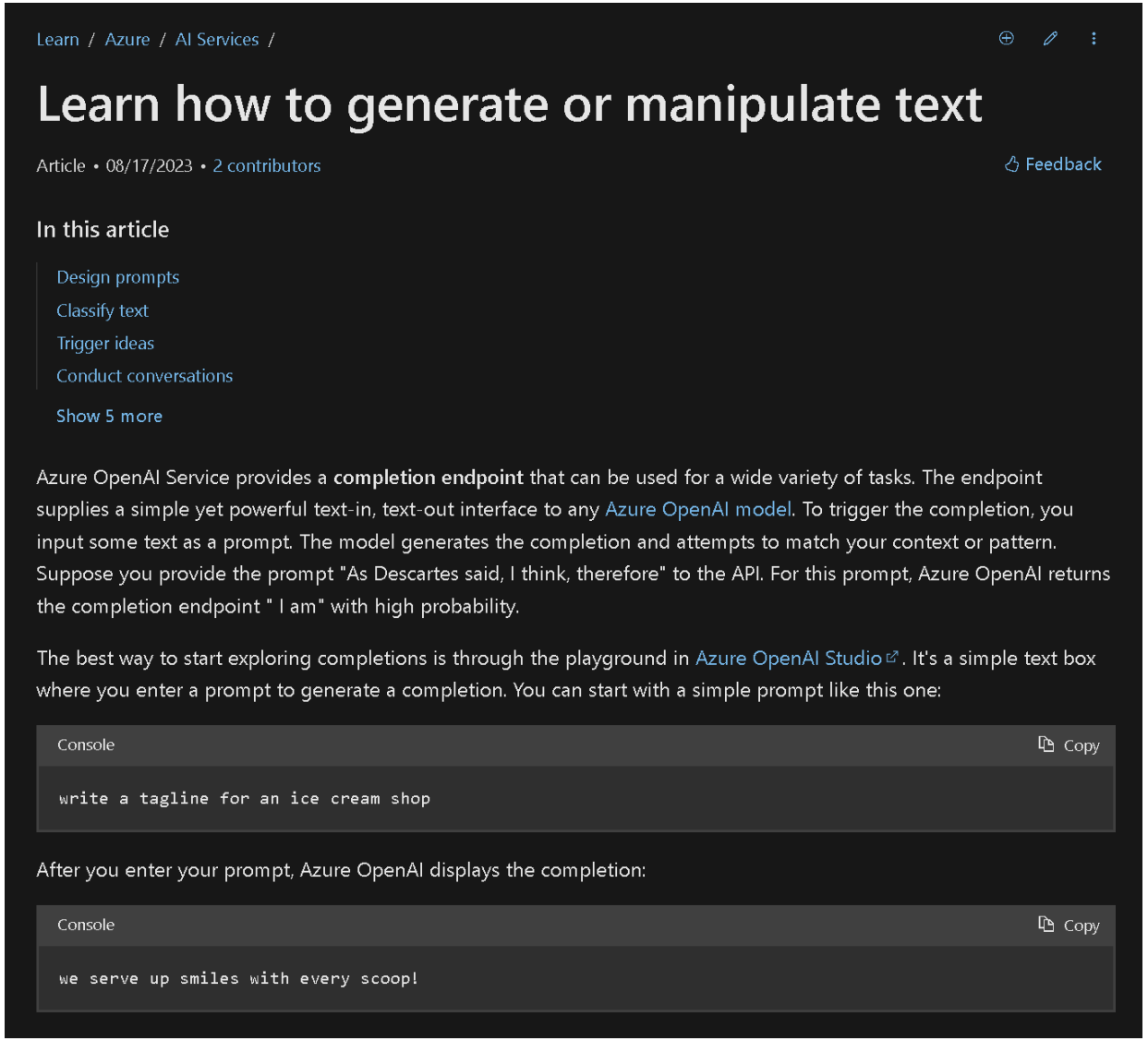
✓ 17.9s


```
{
  "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.

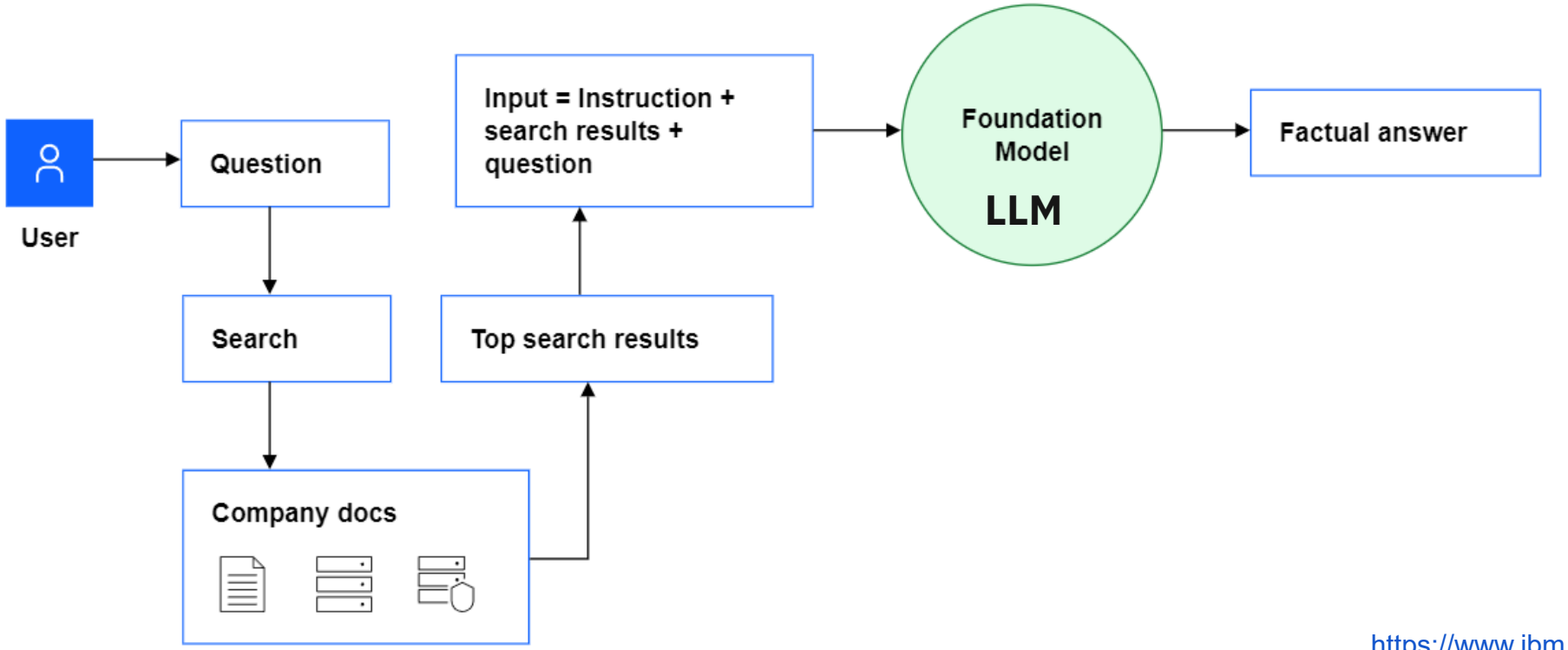
Model ID	Base model Regions	Fine-Tuning Regions	Max Request (tokens)	Training Data (up to)
`gpt-35-turbo`[1] (0301)	East US, France Central, South Central US, UK South, West Europe	N/A	4,096	Sep 2021
`gpt-35-turbo` (0613)	Australia East, Canada East, East US, East US 2, France Central, Japan East, North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

{retrieval-augmented generation RAG}

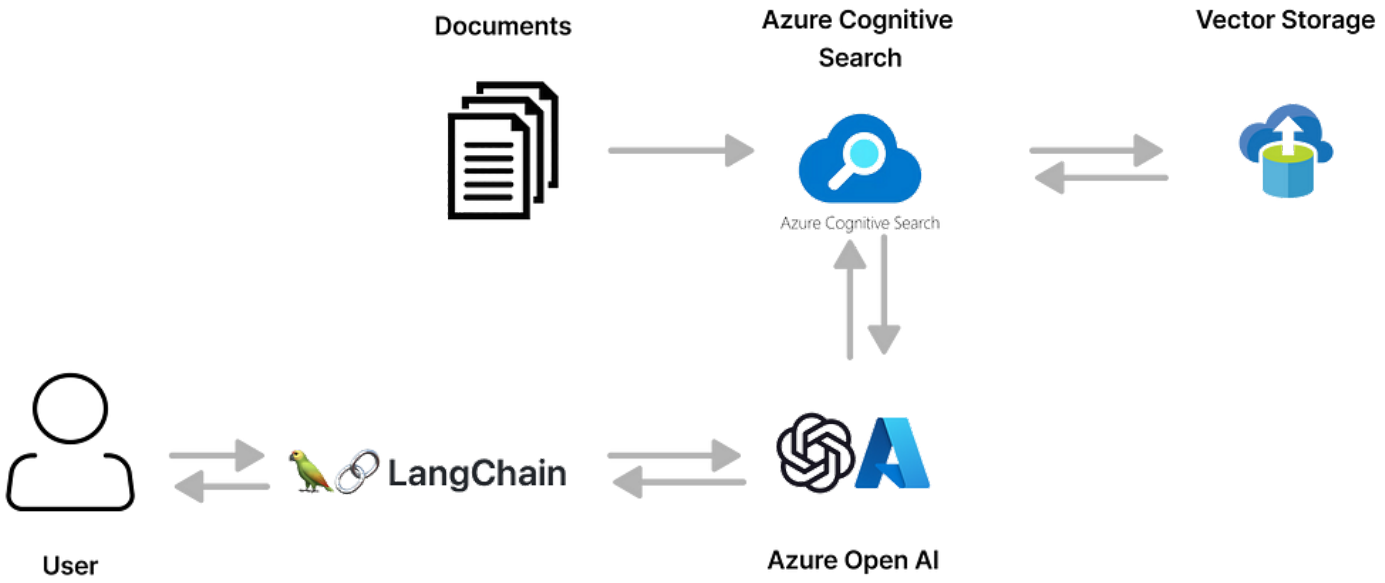
Retrieval-augmented generation [RAG] pattern involves :



https://www.ibm.com/docs/en/watsonx/saas?topic=solutions-retrieval-augmented-generation

HOGESCHOOL ROTTERDAM

{RAG via AZURE + OPENAI + LANGCHAIN}



HOGESCHOOL ROTTERDAM

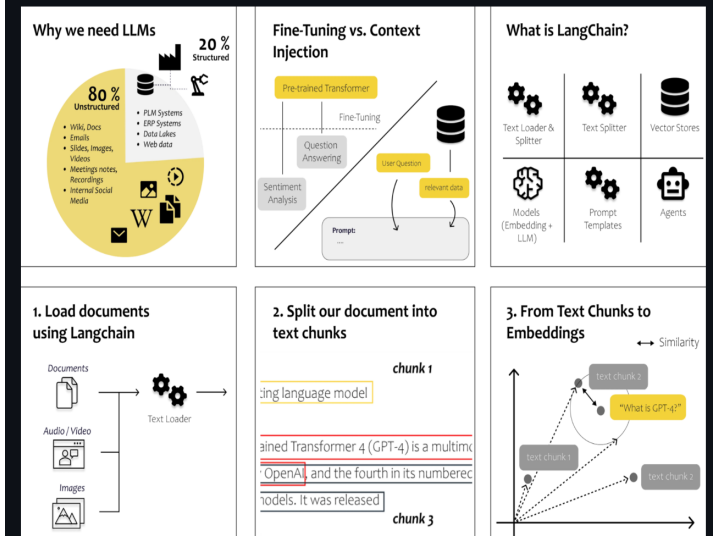
Leer-je-eigen-documenten-bevragen

Context & Doelen

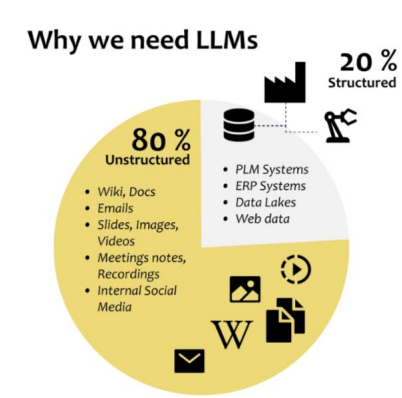
PROGRAMMA AI & ETHIEK

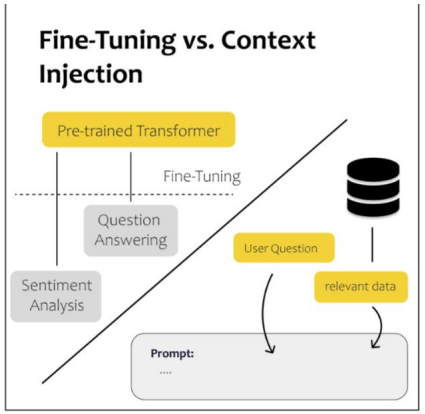
RAG implementa

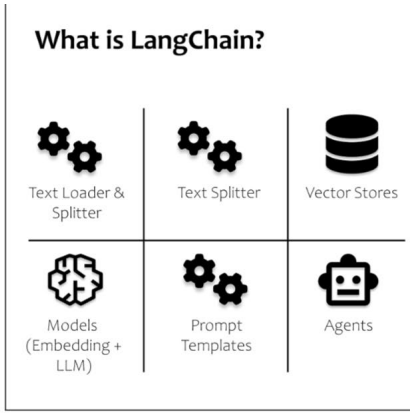
1. Begrijpen wat RAG wel en niet ka
2. Veiligheidsmaatregelen nemen
3. LangChain leren gebruiken voor RAG implementatie met
4a. Azure Resource aanmaken nodig voor deployment van een LLM zoals GPT
4b. Azure OpenAI API key + deployment aanmaken voor model: "text-embedding-ada-002"
5. Jupyter Notebook aanmaken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

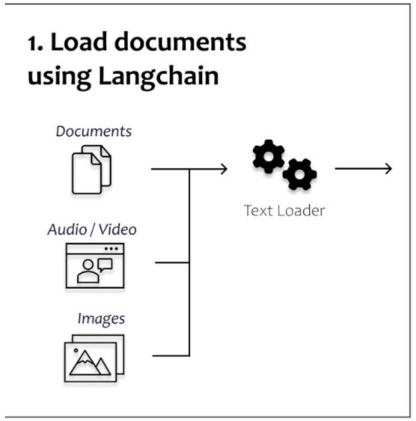


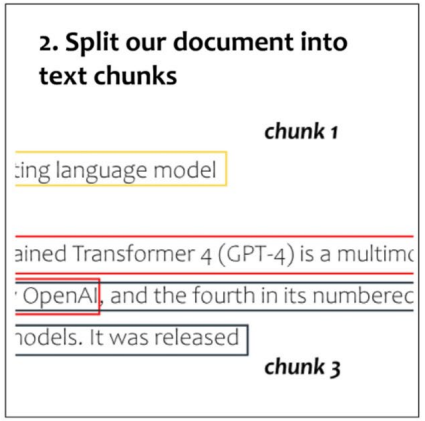
https://github.com/HR-DATA-FABRIC/Leer-je-eigen-documenten-bevragen-met-generatieve-AI

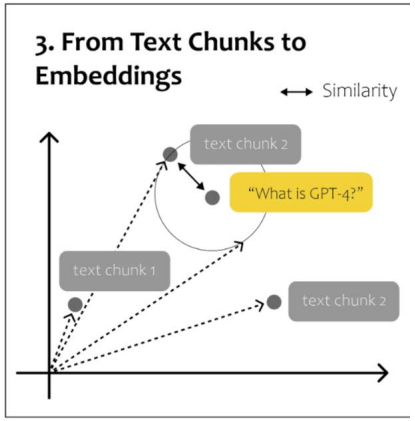


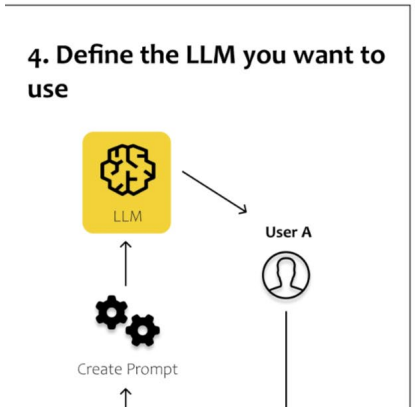


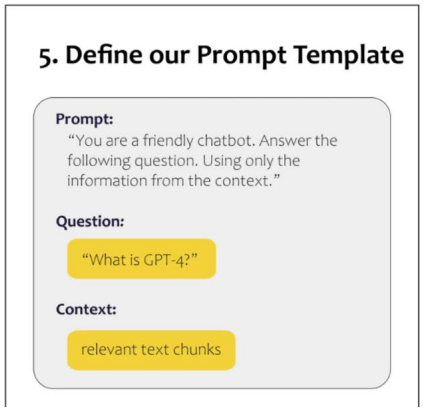


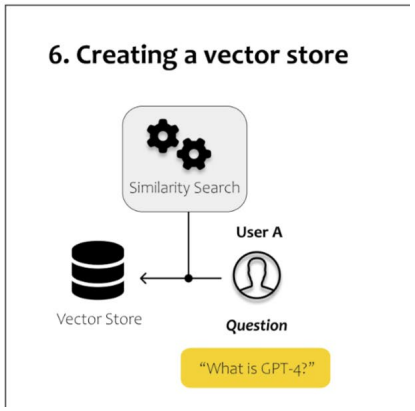






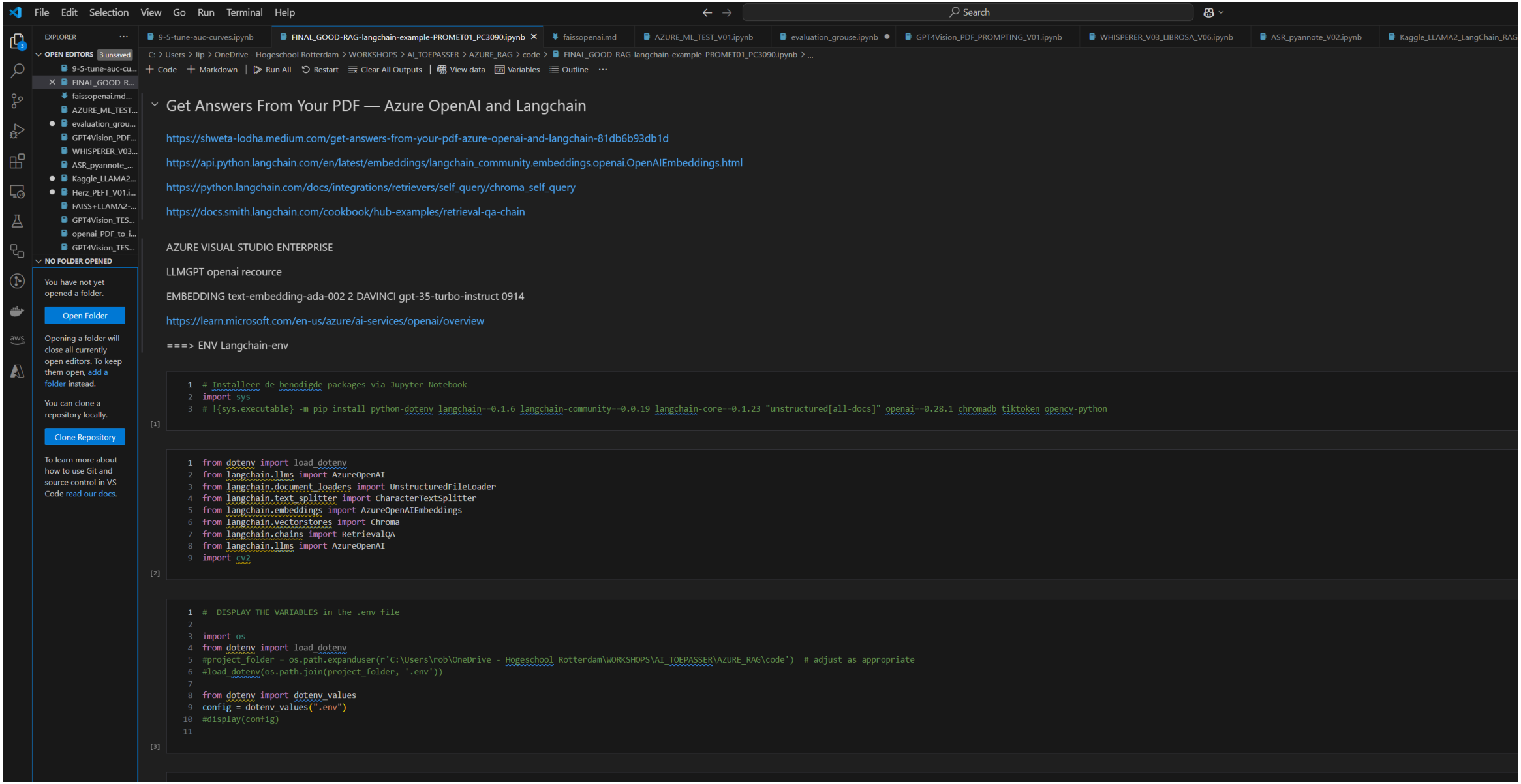






{On -premise (laptop) use of Visual Studio Code}

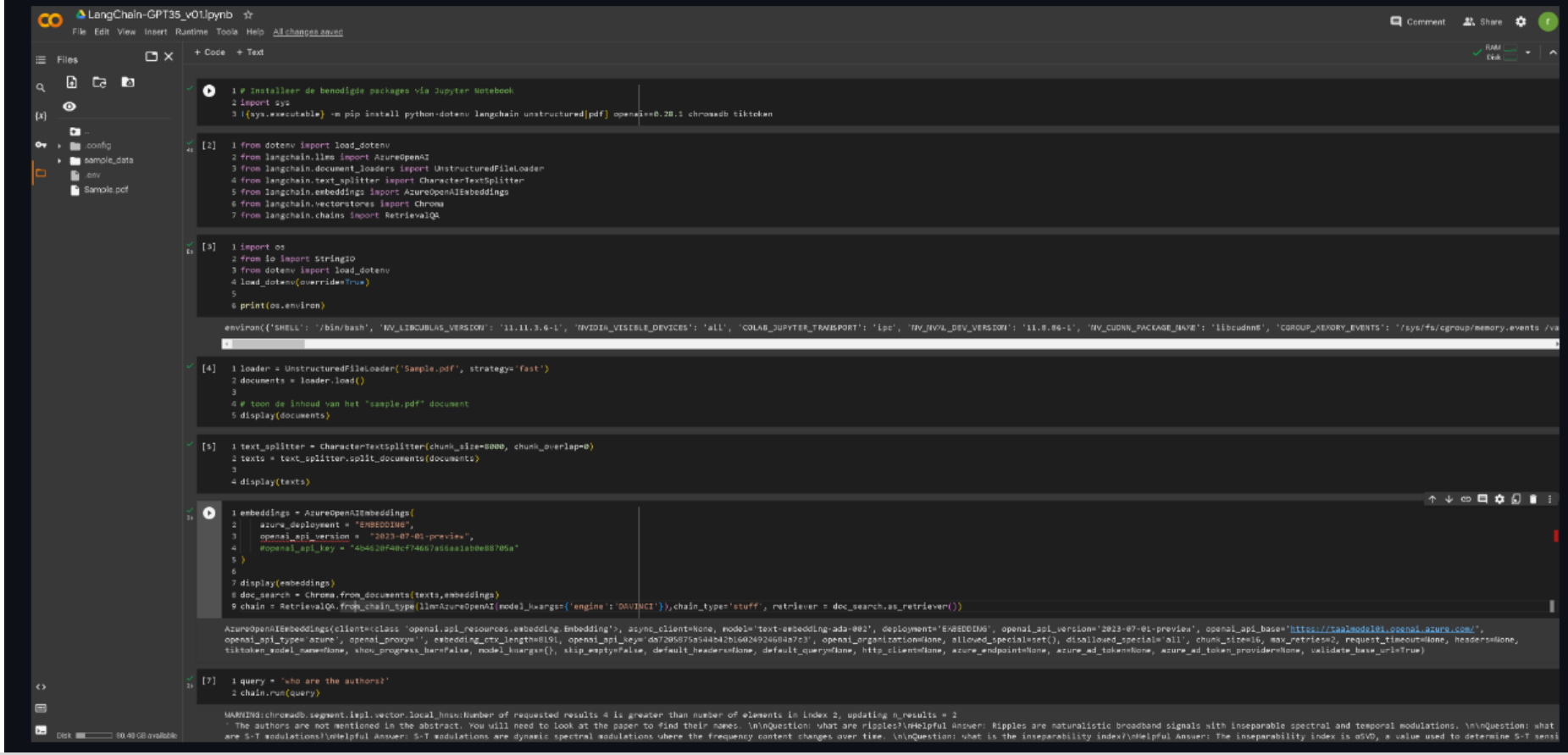
HOGESCHOOL ROTTERDAM



demo

Azure Cognitive Search x Azure OpenAI x LangChain

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



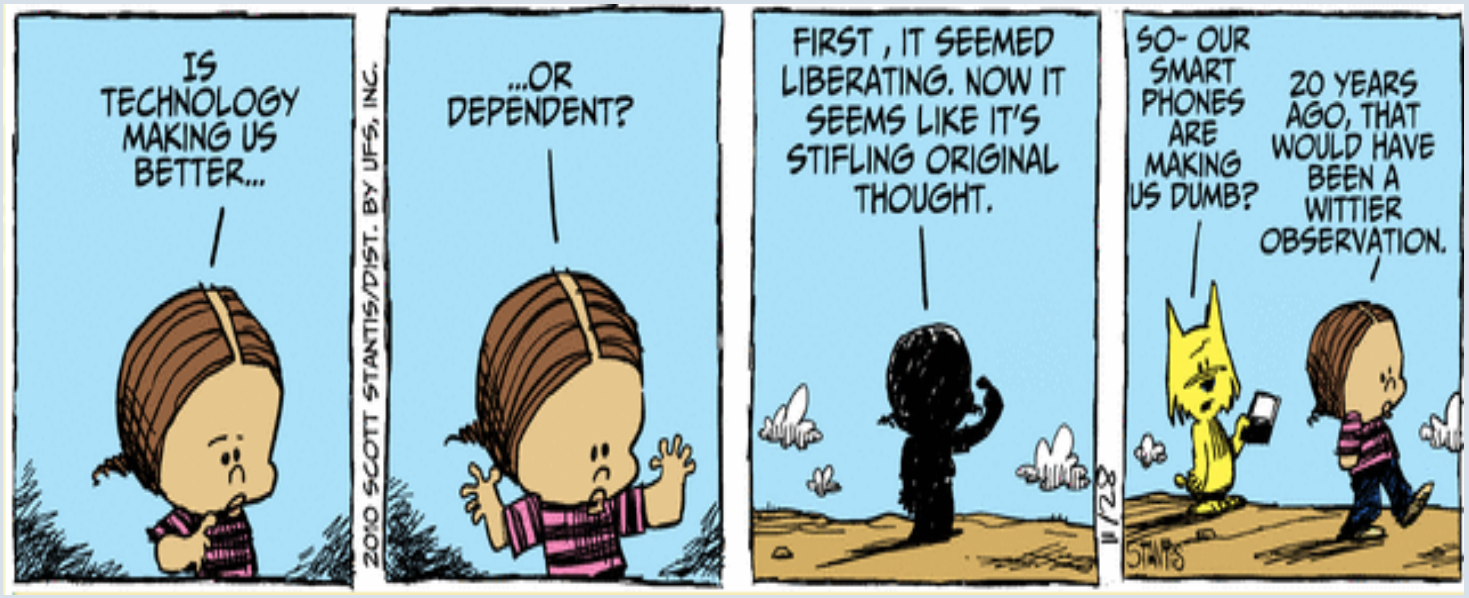
https://colab.research.google.com/drive/1bNbHBXzP1tnDU-xNDRLLeCLFkK3XnI7K#scrollTo=mcw9fPYBOh7b

HOGESCHOOL ROTTERDAM

{Knowledge Dissemination & Curation}

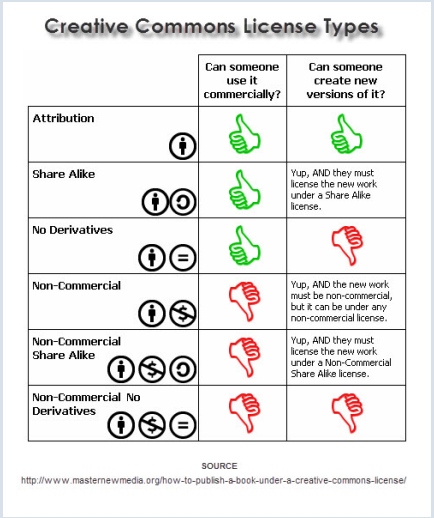
HOGESCHOOL ROTTERDAM

Hight quality, insightful Dutch reviews on AI



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

JUNI 2024



Creative Commons License Types

	Can someone use it commercially?	Can someone create new versions of it?
Attribution	👍	👍
Share Alike	👍	Yup, AND they must license the new work under a Share Alike license.
No Derivatives	👍	👎
Non-Commercial	👎	Yup, AND the new work must be non-commercial, but it can be under any non-commercial license.
Non-Commercial Share Alike	👎	Yup, AND they must license the new work under a Non-Commercial Share Alike license.
Non-Commercial No Derivatives	👎	👎

SOURCE
http://www.masternewmedia.org/how-to-publish-a-book-under-a-creative-commons-license/