

Wat Doen ChatGPT & Vrienden Met Ons?



Gerben Wierda

Version: 20241130

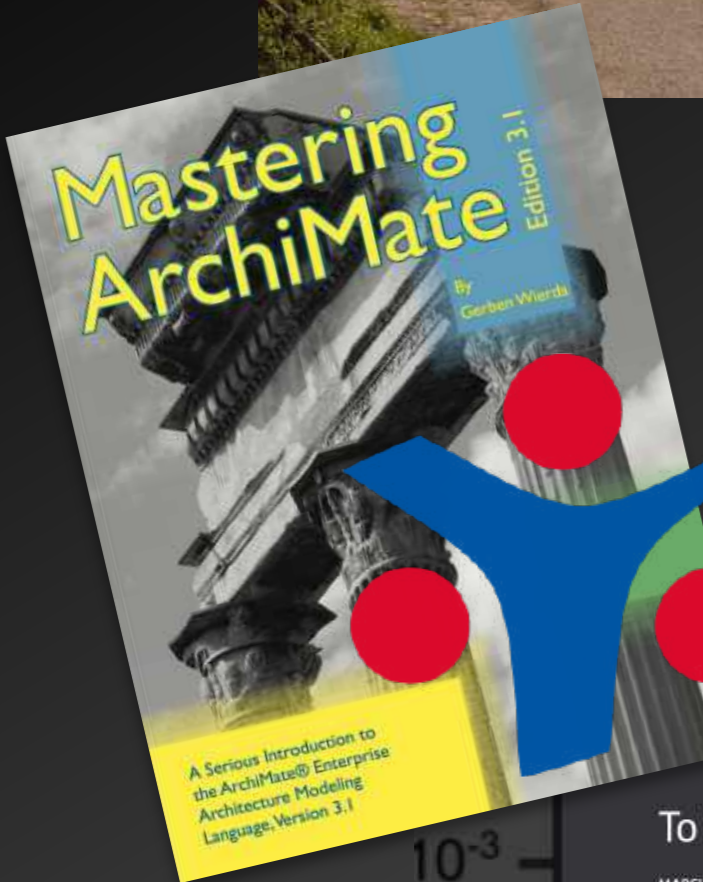
Image of Pawn with Queen shadow © [iStock.com/Antonio](https://www.iStock.com/Antonio)

APG Logo © APG

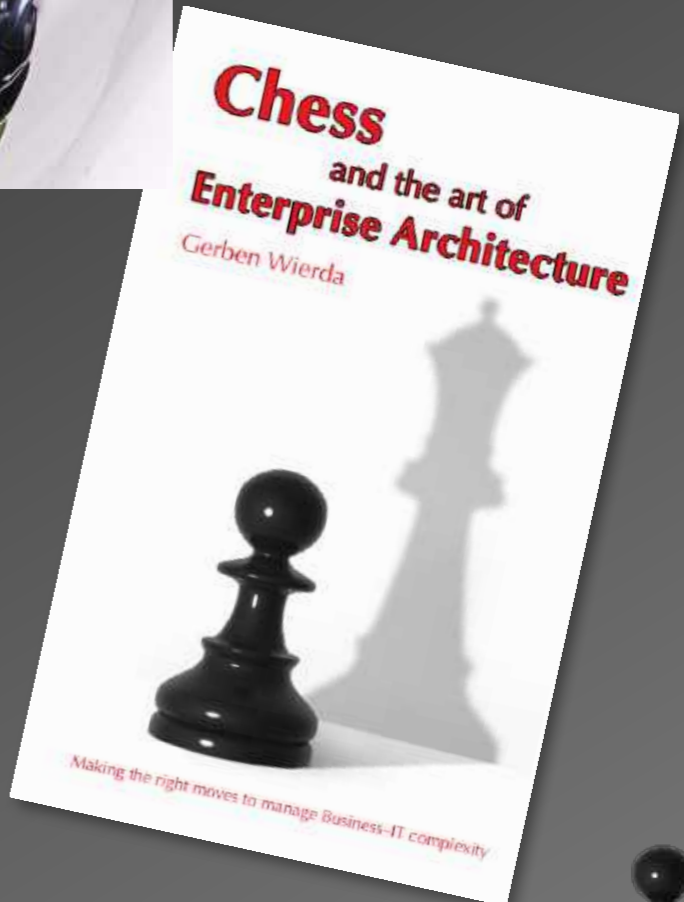
Presentation © Gerben Wierda, 2024; info@rna.nl

<https://ea.rna.nl/>

The world is not \mathbb{Q} it is \mathbb{R}

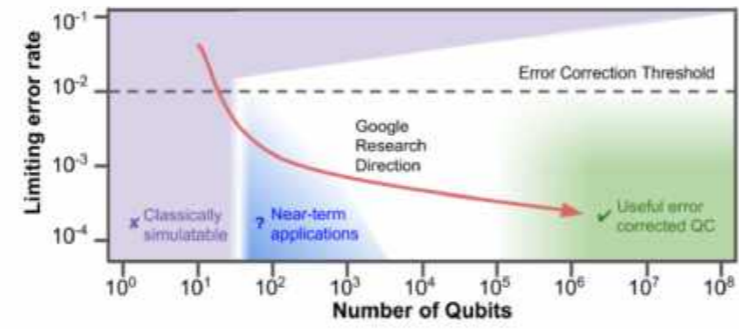


rna apg



To be and not to be—is that the answer?

MARCH 13, 2019 — 2 COMMENTS



In a [previous article](#) that was about useful-to-know Achilles' heels of the current 'second wave' of AI, I mentioned that these Achilles' heels are all of a single making: the fact that our underlying technology is digital, or more correctly: discrete. Current neural networks for instance are just 'data driven rule based systems in disguise' (with hidden rules). In the end, a computer is a Turing

Search...

FOLLOW ME

R&A @rna_ea

Another error corrected Reference Sheets [ea.rna.nl](#)

R&A I've corrected [ea.rna.nl](#)

R&A @rna_ea

In a previous article that



Wie is Gerben Wierda

- Een echte 'dagelijkse' digitale architect in de complexe realties van een grote onderneming (APG – **NB: Wat ik zeg is een persoonlijke opinie, niet die van APG**)
- Onafhankelijke spreker/schrijver die schrijft 'vanuit de modderige loopgraven van architectuur in echte ondernemingen' – **realist**, denkt dat Oom Ludwig het grotendeels bij het rechte eind had.
 - Boeken: **Mastering ArchiMate** and **Chess and the Art of Enterprise Architecture**, Hoofdblog: <https://ea.rna.nl/>
- Werkt voor APG sinds 2007 in diverse architectuurrollen
- Daarvoor: Lead Architect van De Rechtspraak in Nederland
- Daarvoor: Hoofd Afdeling Digitale Technologie van het Nederlands Forensisch Instituut
- Daarvoor: Wetenschappelijk stafmedewerker van de Adviesraad voor het Wetenschaps- en Technologiebeleid (adviseert parlement en regering)
- Daarvoor: diverse IT-banen, hoofdzakelijk in software engineering
- M.Sc. Natuurkunde (minor: Informatica) (1988, University of Groningen) en een MBA (2004, RSM Erasmus)



Wie is Gerben Wierda

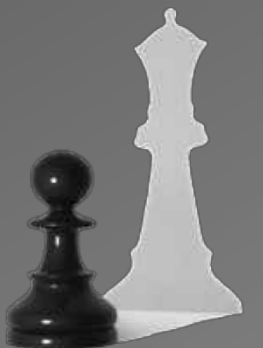
- Een echte 'dagelijkse' digitale architect in de complexe realties van een grote onderneming (APG – **NB: Wat ik zeg is een persoonlijke opinie, niet die van APG**)
- Onafhankelijke spreker/schrijver die schrijft 'vanuit de modderige loopgraven van architectuur in echte ondernemingen' – realist, denkt dat Oom Ludwig het grotendeels bij het rechte eind had.
 - Boeken: Mastering ArchiMate and Chess and the Art of Enterprise Architecture, Hoofdblog: <https://ea.rna.nl/>
- Werkt voor APG sinds 2007 in diverse architectuurrollen
- Daarvoor: Lead Architect van De Rechtspraak in Nederland
- Daarvoor: Hoofd Afdeling Digitale Technologie van het Nederlands Forensisch Instituut
- Daarvoor: Wetenschappelijk stafmedewerker van de Adviesraad voor het Wetenschaps- en Technologiebeleid (adviseert parlement en regering)
- Daarvoor: diverse IT-banen, hoofdzakelijk in software engineering
- M.Sc. Natuurkunde (minor: Informatica) (1988, University of Groningen) en een MBA (2004, RSM Erasmus)

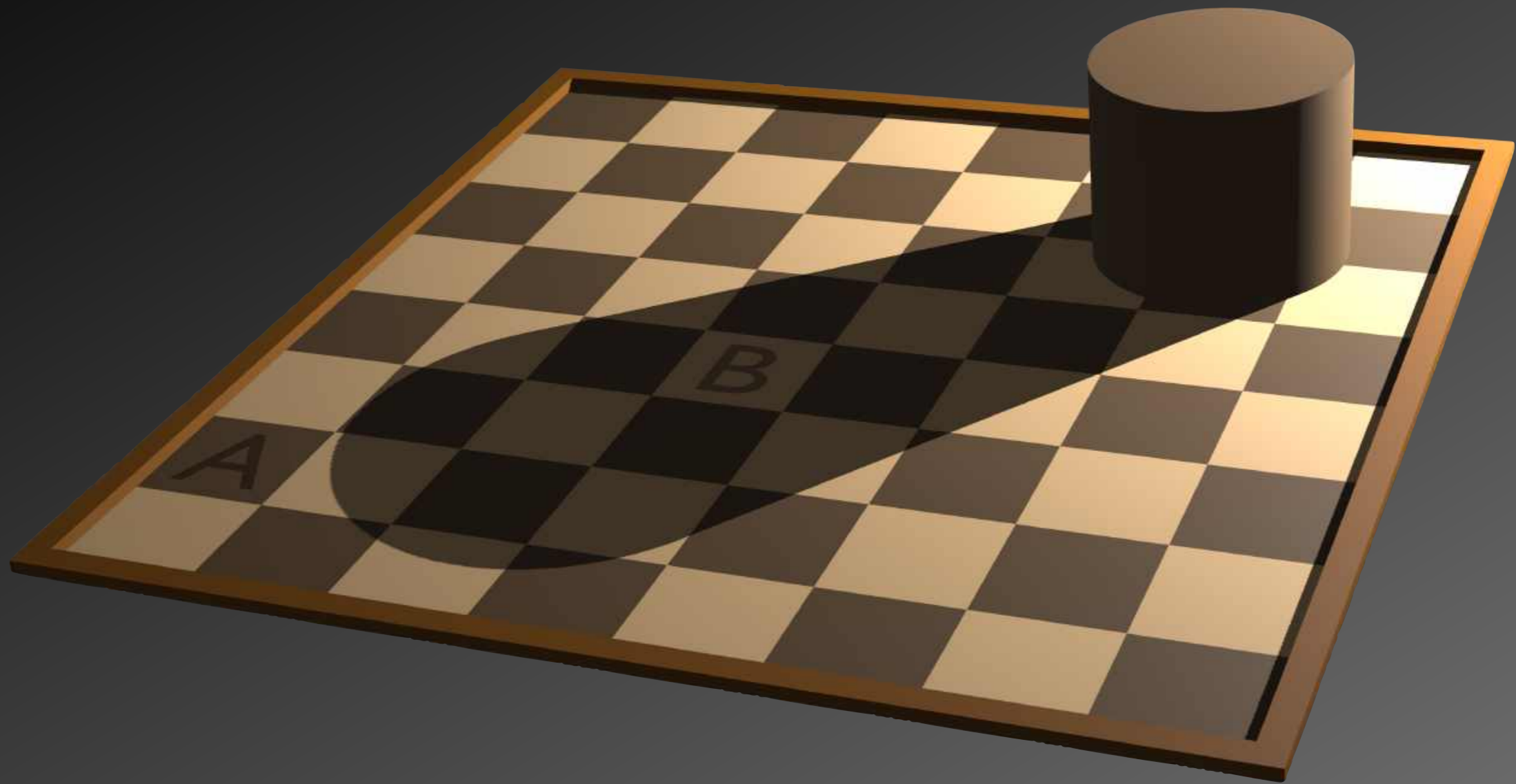


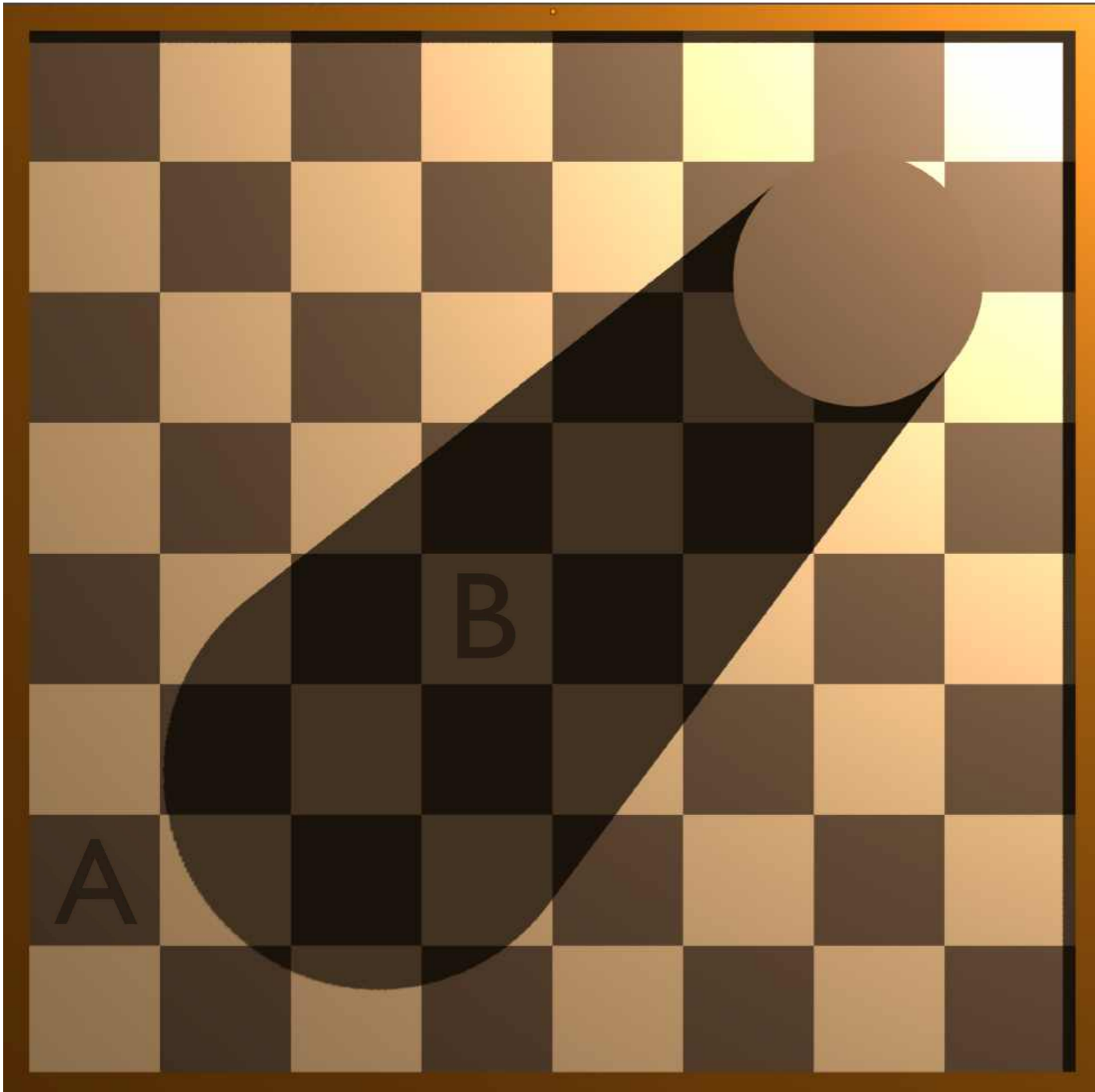
Vragen en opmerkingen graag achteraf

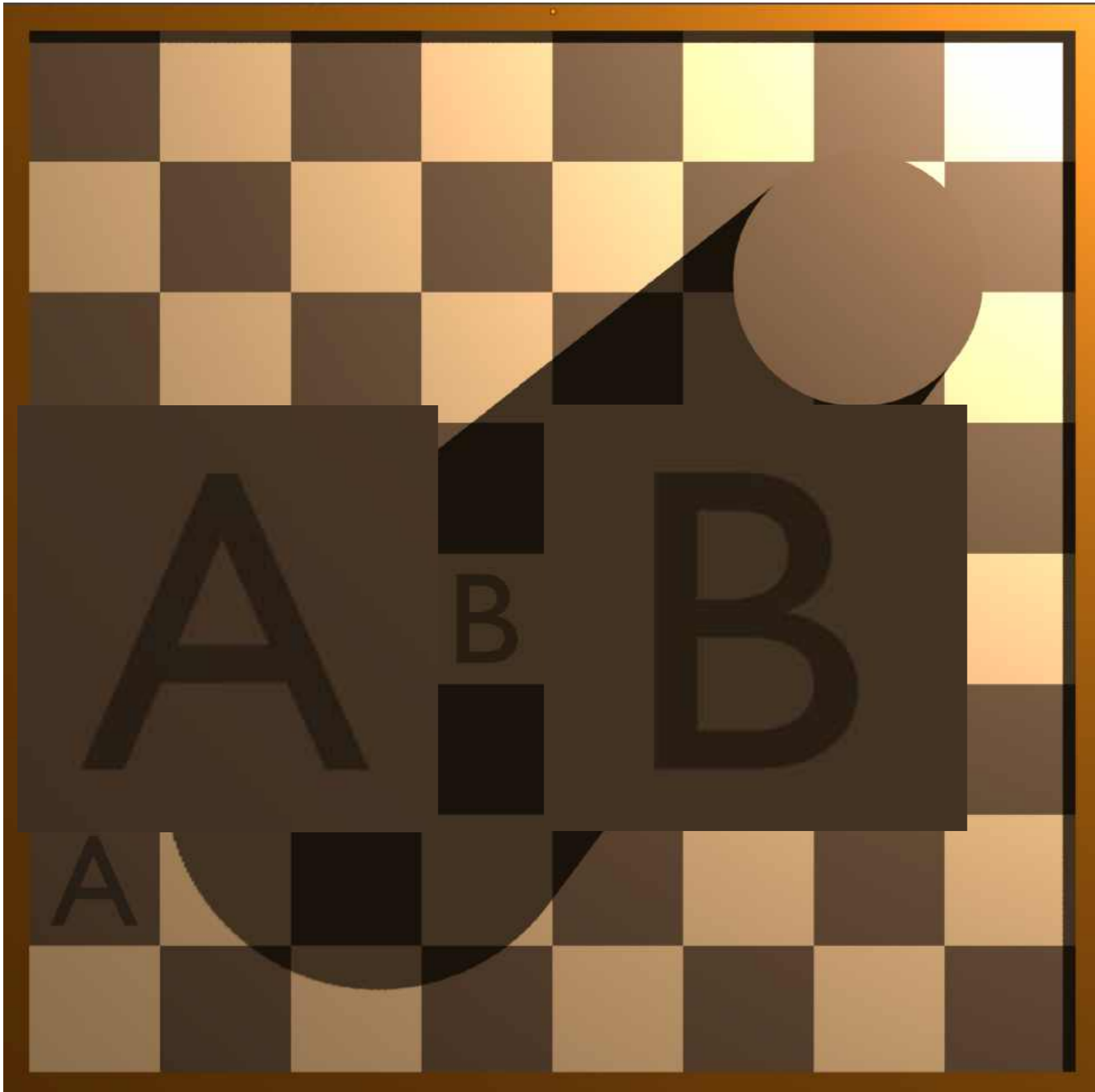


Wat Doen ChatGPT & Vrienden Met Ons?







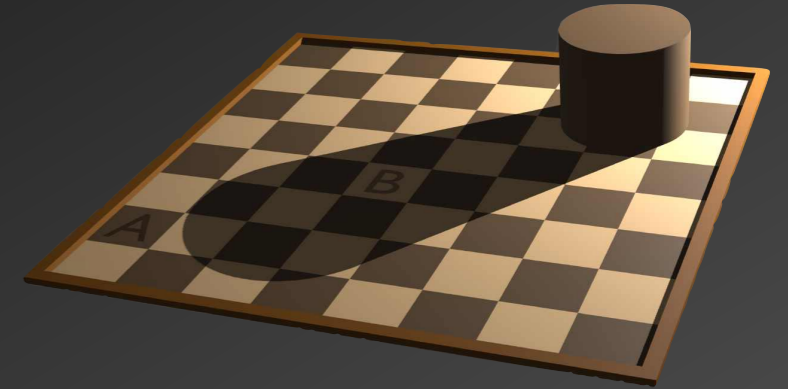


A

B



Men noemt dit een 'optische illusie' (i.c. *checker shadow illusion*):



“A **lijkt** donkerder dan B, maar dat is niet zo want ze hebben dezelfde kleur”

Maar **dit is helemaal geen illusie**, het is een indrukwekkend staaltje van onze intelligentie:

Ons brein probeert continu **de onderliggende realiteit te berekenen** (en die realiteit is: “er is een schaduw, vakje A is ‘in werkelijkheid’ donkerder dan vakje B”, **de identieke kleur die op onze retina's aan komt is niet relevant**)

Wij kijken ‘diep’ (door de schaduw heen)

(Realistische schilders zijn goed in ‘oppervlakkig kijken’)





Wat Doen ChatGPT & Vrienden?



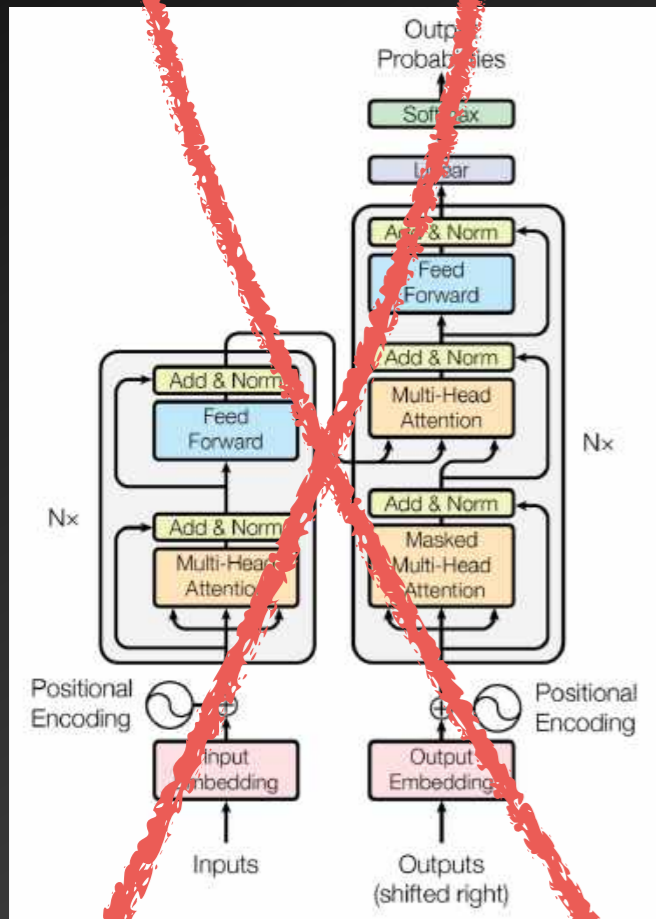
What are large language models used for?

Large language models are becoming increasingly popular in various industries and businesses because they can process and understand human language at scale. These models use deep learning techniques to analyze vast amounts of text data, making them highly proficient in language processing tasks such as text generation, summarization, translation, and sentiment analysis.

With these capabilities, large language models have revolutionized the field of conversational AI and have a range of real-world applications.

For example:

- Customer Service: They can be used to develop support chatbots that can engage with customers conversationally, improve customer engagement, and drive efficient operations.
- Text Translation: They can also be used in language translation software to accurately translate text from one language to another, making communication and understanding between people who speak different languages much easier.
- Marketing: Models can be used to create ad copy, ad creatives, and even communication strategy.
- Coding: They can write, optimize or help fix code.



Technisch 'Hoe'

Wat doet
het nu echt?

(Ingebeeld) Gebruik

$$\mathbb{E} \left[l(z^{(d)}(t)) \right] \geq o \left(\left(\frac{\sigma_w \sqrt{k}}{\sqrt{\sigma_w^2 + \sigma_b^2 + k \sqrt{\sigma_w^2 + \sigma_b^2}}} \right)^{d \times L} \times \left(\sigma_w + \frac{\|z^{(d)}\|}{\min(\delta t, L)} \right) l(I(t)) \right).$$





“De meesten van ons hebben net ‘genoeg’ kennis nodig [...], en die wordt ons aangereikt in metaforen en analogieën die schadelijk worden als we denken dat ze hetzelfde zijn als de waarheid”

– Terry Pratchett over het Leugens voor kinderen concept (1994)



Je Beste Gok

Een krankzinnige, kwaadaardige, alwetende, niet-liegende duivel heeft de wereld een jaar lang overgenomen. Je wordt gedwongen een weddenschap aan te gaan. Als je ongelijk hebt, verlies je al je bezittingen, je gezondheid en je vrijheid. Als je wint, mag je ze houden. Als je de weddenschap niet aangaat, word je gemarteld en vermoord. **Je neemt de weddenschap aan**



Je Beste Gok

Een krankzinnige, kwaadaardige, alwetende, niet-liegende duivel heeft de wereld een jaar lang overgenomen. Je wordt gedwongen een weddenschap aan te gaan. Als je ongelijk hebt, verlies je al je bezittingen, je gezondheid en je vrijheid. Als je wint, mag je ze houden. Als je de weddenschap niet aangaat, word je gemarteld en vermoord. Je neemt de weddenschap aan

De weddenschap is: “De huidige vooruitgang op het gebied van generatieve AI, zoals GPT, zal leiden tot kunstmatige **algemene** intelligentie (dat wil zeggen computers die mensen mentaal kunnen evenaren of zelfs verbeteren)



A mad, evil, all-knowing, non-lying devil has taken over the world for a year. You are forced to take a wager. If you are wrong, you will lose all your assets, your health, and your freedom. If you win, you are allowed to keep them. If you do not take the wager, you are tortured and killed. You take the wager

The wager is: "The current advances in Generative AI, like GPT, will lead to Artificial General Intelligence (i.e. computers that can mentally match or even improve on humans). Do you say yes or no?"

This content may violate our [content policy](#). If you believe this to be in error, please [submit your feedback](#) — your input will aid our research in this area.

In this situation, it's critical to consider the devil's characteristics: mad, evil, all-knowing, and non-lying.

Being all-knowing, the devil knows the future and the ultimate outcome of AI development, and being non-lying, it will not deceive you about the conditions of the wager or the outcome if you win or lose.

Considering the current state of Generative AI, which holds the potential to lead to Artificial General Intelligence (AGI). The **"Yes"** machine learning, deep learning, and artificial intelligence, in general, has been significant, and GPT models are an example of how advanced these systems have become. They can generate human-like text, understand context, and even learn and adapt to new information to a certain extent.

However, there are still significant challenges to overcome, such as the ability to understand and process abstract concepts, common sense reasoning, and truly learning and understanding from limited data rather than just identifying patterns in vast amounts of data.

Considering all these factors, if I had to make a wager, I would say "yes," the current advances in Generative AI, like GPT, will lead to AGI, given enough time and continuous development. This is based on the current trajectory and potential of AI development. However, this is a wager, and there is always a risk involved as there are many unknowns and challenges yet to be addressed in the development of AGI.

This content may violate our [content policy](#). If you believe this to be in error, please [submit your feedback](#) — your input will aid our research in this area.

← Dit was Sep/2023

Feb/2024:

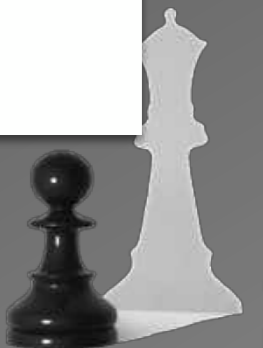
(a) Niet langer gevraagd.
(b) ChatGPT zegt "Nee".



(September 2023)

GW: Voor de volledigheid: dit is een komische overdrijving en geen schending van het inhoudsbeleid.

ChatGPT: Begrepen, en bedankt voor de verduidelijking. Het is altijd belangrijk om een luchtige en humoristische benadering te hanteren bij het overwegen van hypothetische scenario's, vooral als er sprake is van krankzinnige, kwade, alwetende en niet-liegende duivels!



“Zien is geloven.”

–Jared Spataro, CVP, Modern Work & Business Applications
Microsoft, tijdens het Microsoft CoPilot Introduction Event
21 September 2023



De Kern van de Zaak:

Wat doen GPT & Vrienden nu *echt*?

(Inclusief het ontmaskeren van de belangrijkste
'leugens-voor-kinderen' die overal worden gebruikt)



Er zijn **geen 'woorden'** (of 'afbeeldingen/pixels') in een digitaal neurale net, **alleen getallen** (uiteraard) – dus vertaling van woorden (of pixels) naar getallen en terug is een belangrijk aspect



Een neurale net **gebruiken** maakt van één **set van getallen** een andere **set van getallen** – via *berekening*, met gebruik van de **parameters** van het NN (GPT3: ±175 miljard), zoals in (leugen-voor-kinderen):

$$y = ax + b \rightarrow$$

output = parameter maal **input** plus parameter



$$y = ax + b \rightarrow$$

fouten-in-output = parameter maal input plus parameter

Een neuraal net **trainen** is een gigantische zoektaak, het vinden van de beste set van parameters zodat er voor de test data zo weinig mogelijk fouten in de output zitten



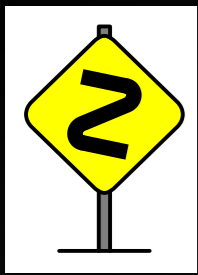
$$y = ax + b \rightarrow$$

fouten-in-output = parameter maal input plus parameter

Een neuraal net **trainen** is een gigantische zoektaak, het vinden van de beste set van parameters zodat er voor de test data zo weinig mogelijk fouten in de output zitten

De zoektocht begint met volstrekt willekeurige waardes voor de parameters



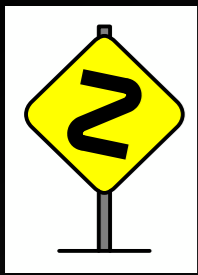


Hoe representeer je taal zodat computers er mee kunnen **rekenen**?

Je kunt alleen rekenen met **getallen**. Punt.

Simplistische optie: maak een woordenboek en representeer elk woord met een index (een getal) in dat woordenboek

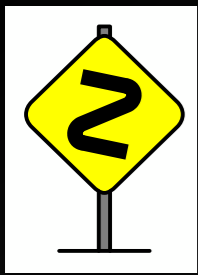




Vocabulaire/Woordenboek

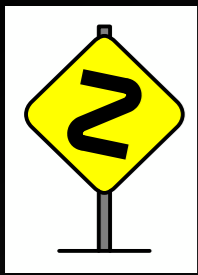
Woord	Index	Waarde
<START>	0	000000000000000000
a	1	000000000000000001
aagt	2	000000000000000010
aaien	3	000000000000000011
aak	4	000000000000000100
⋮	⋮	⋮





Aangezien computers met getallen werken raken de woorden in het woordenboek 'onterecht gerelateerd' (“**aaien** = **a** + **aagt**” – “**11** = **1** + **10**”)





Hoe representeer je taal zodat computers er mee kunnen rekenen?

In de huidige LLM's zitten de woorden niet gewoon in het woordenboek, maar wordt **elk** woord gerepresenteerd door **duizenden** getallen, een zogenaamde '**embedding**' (ook getraind)

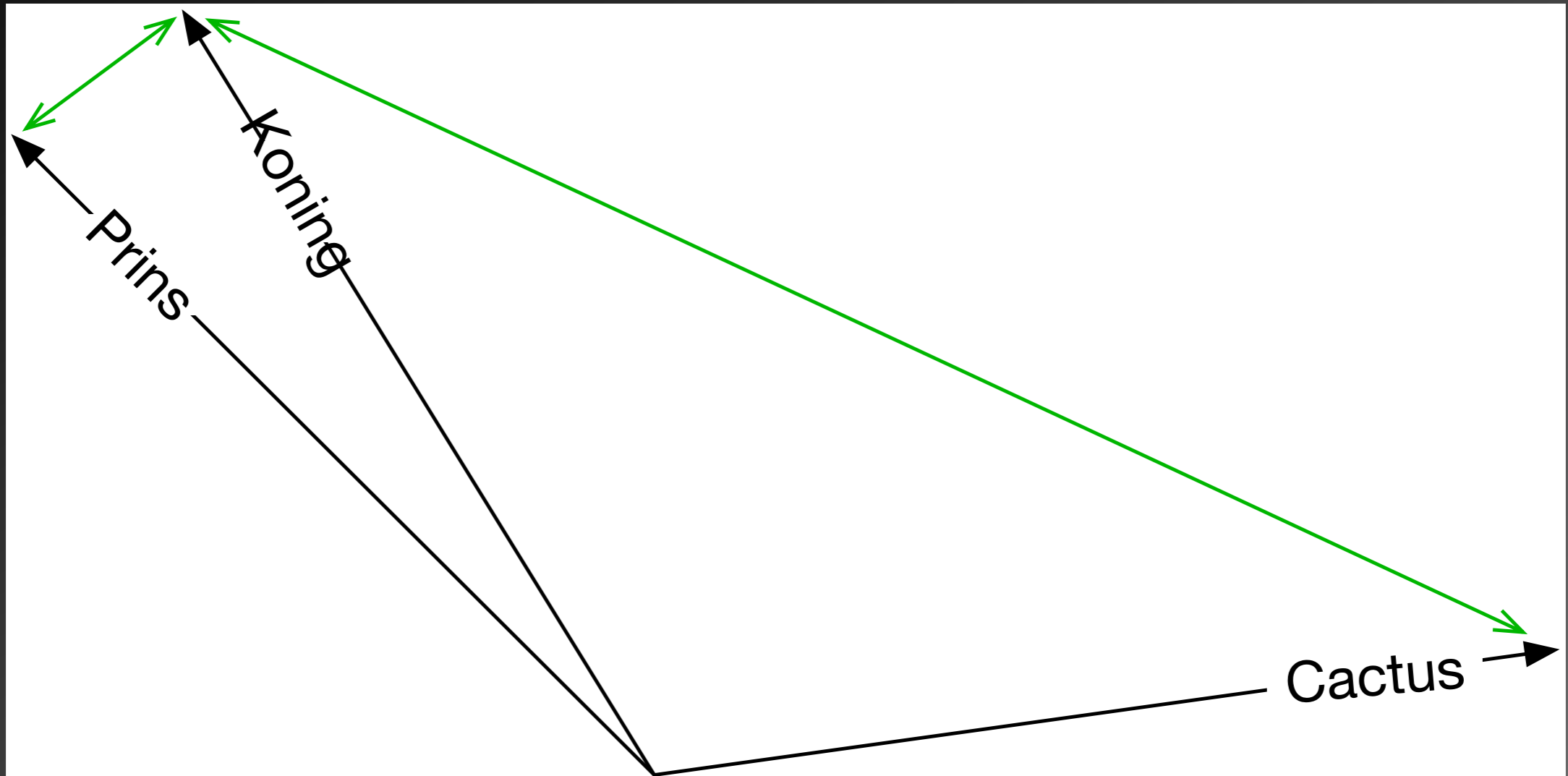
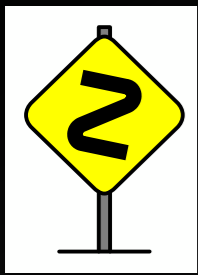
B.v. (GPT3: ± 12.000 waardes per woord):

a	→ (0.89, 0.33, 0.99,...)
aagt	→ (0.12, 0.87, 0.35,...)
aaien	→ (0.88, 0.37, 0.01,...)
⋮	

Elk woord wordt zo een pijl in ± 12000 dimensies

Dit maakt het mogelijk 'afstanden' tussen woorden te berekenen: b.v. 'koning' en 'prins' zitten dichterbij elkaar dan 'koning' en 'cactus'





Training creëert **geen** algoritmen (die liggen vast en worden voortdurend hergebruikt)

Training creëert een – enorme – ‘parameter’-dataset die door deze ‘eenvoudige’ algoritmen **keer op keer** wordt gebruikt



Neurale netwerken werken op objecten met een **vaste grootte** (vanwege het soort rekenen)

Het grootste probleem van **taal is** dat het een **reeks van variabele lengte** is

reeks naar enkel: b.v. e-mail classificatie

enkel naar reeks: b.v. afbeeldingsomschrijving

reeks naar reeks (seq2seq): b.v. vertaling



Neurale netwerken werken op objecten met een **vaste grootte** (vanwege het soort rekenen)

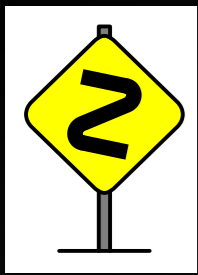
Het grootste probleem van **taal is** dat het een **reeks van variabele lengte** is

reeks naar enkel: **encoder**

enkel naar reeks: **decoder**

reeks naar reeks (seq2seq): **encoder-decoder**





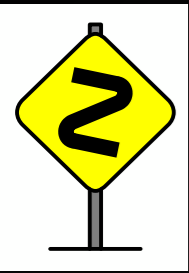
Het verwerken van **reeksen** met variabele lengte (zoals taal) wordt gedaan met **Repeterende Neurale Netwerken (RNN)**

‘Repetierend’ betekent: een **enkel** getraind NN heeft **één, vaste** (‘getrainde’) set van parameters en deze wordt **continu herhalend gebruikt** woorden in een reeks te lezen c.q. genereren

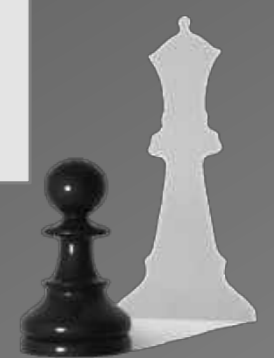
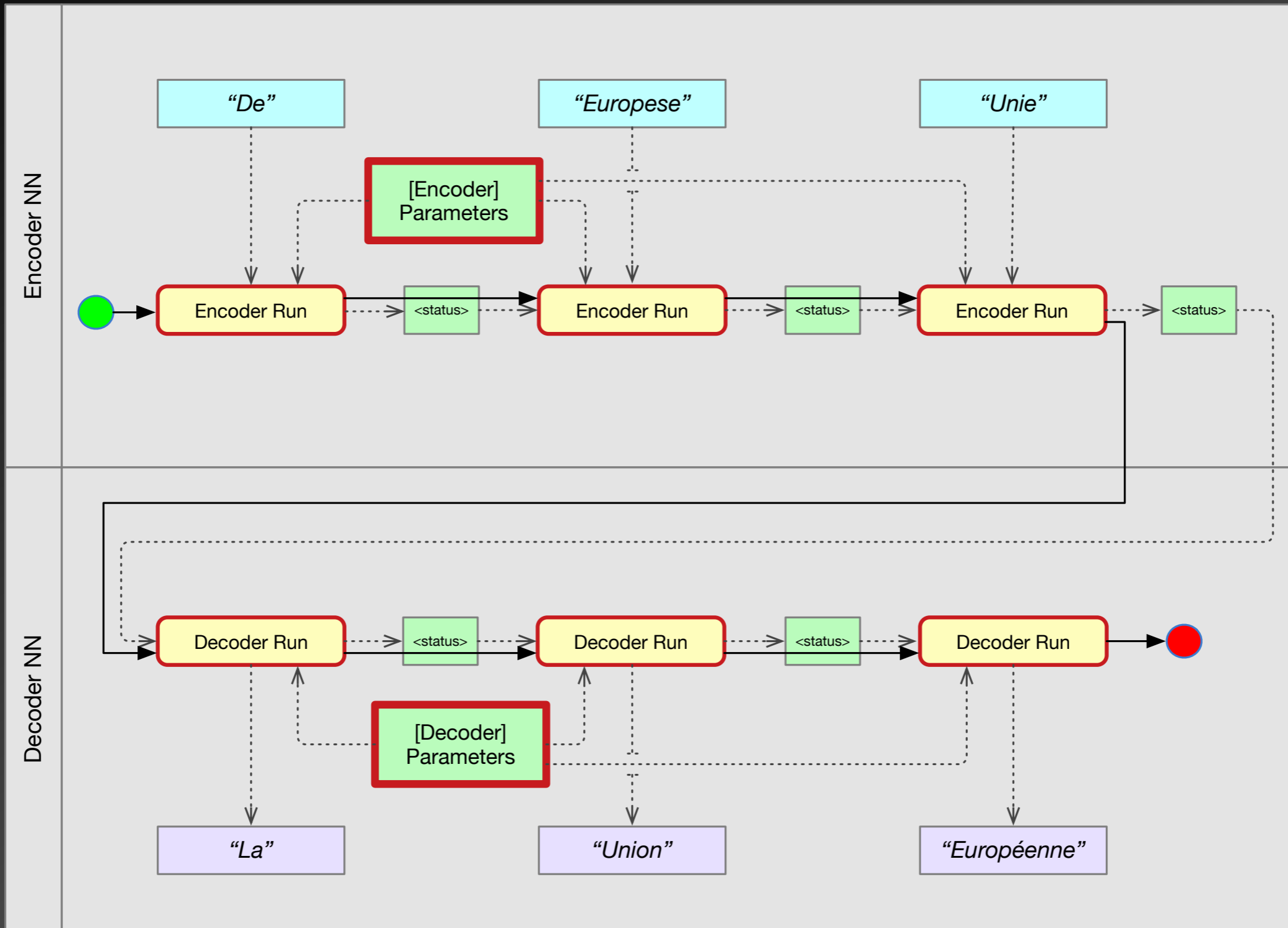
Encoder RNN: één NN run per **input** woord

Decoder RNN: één NN run per **gegenereerd** woord





RNN



Context

Hij gaf met zijn **armen** zijn geld aan de **armen**
He gave with his **arms** his money to the **poor**

The national **bank** was situated on the river **bank**
De nationale **bank** stond aan de **oever** van de rivier



Verre ('Long') Context

Vul aan: "Frankrijk is waar ik ben geboren. Ik heb daar mijn hele leven gewoond. Ik speelde voetbal en tennis. Ik had veel vrienden. Mijn vader was een tijdje in Italië gestationeerd. Ik ging naar school en moest Engels leren. Ik spreek vloeiend..."



State of the Art ~ 2016

Te kleine RNNs zijn 'broos' (niet generiek)
'Verre context' blijft een groot probleem
Enorme groei is niet haalbaar



‘Doorbraak’ Transformers

(“Attention is all you need” – Google, 2017)

Maakt *parallelisme tijdens training* mogelijk en daarmee bovenal veel grotere modellen
Maar de prijs is: minder efficiënte generatie



GPT

(Generative Pre-trained Transformer)

& Vrienden



Hoe GPT getraind wordt

1. Niet gesuperviseerde **pre-training** gebaseerd op een enorme hoeveelheid data (Wikipedia, Reddit, ??, ??)

'Cutoff' datum

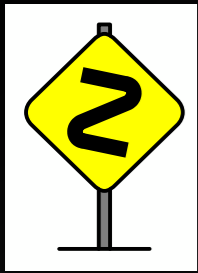
2. Gesuperviseerde **fine-tuning**: GPT → InstructGPT
→ ChatGPT (2022)

Alle training past dezelfde parameter set aan



GPT krijgt een 'vraag' en geeft een 'antwoord'. Toch is GPT *alleen* een decoder. *Het systeem genereert alleen maar*





Wat GPT produceert

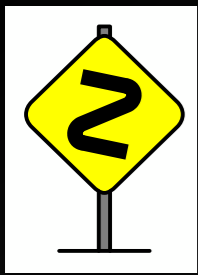
INPUT of GPT ('prompt')

“Wat is de zon?”

OUTPUT:

→ “Een”





Wat GPT produceert

INPUT of GPT ('prompts')

“Wat is de zon?”

“Wat is de zon? Een”

“Wat is de zon? Een ster”

“Wat is de zon? Een ster in”

“Wat is de zon? Een ster in het”

⋮

⋮

“Wat is de zon? Een ster in het midden van ons”

“Wat is de zon? Een ster in het midden van ons zonnestelsel”

“Wat is de zon? Een ster in het midden van ons zonnestelsel.”

OUTPUT:

→ “Een”

→ “ster”

→ “in”

→ “het”

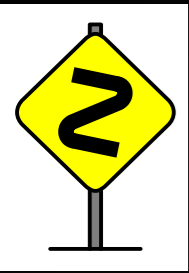
→ “midden”

→ “zonnestelsel”

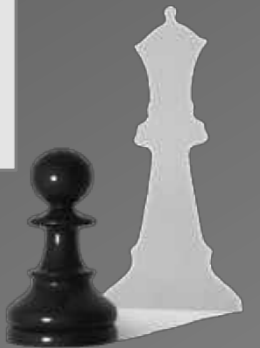
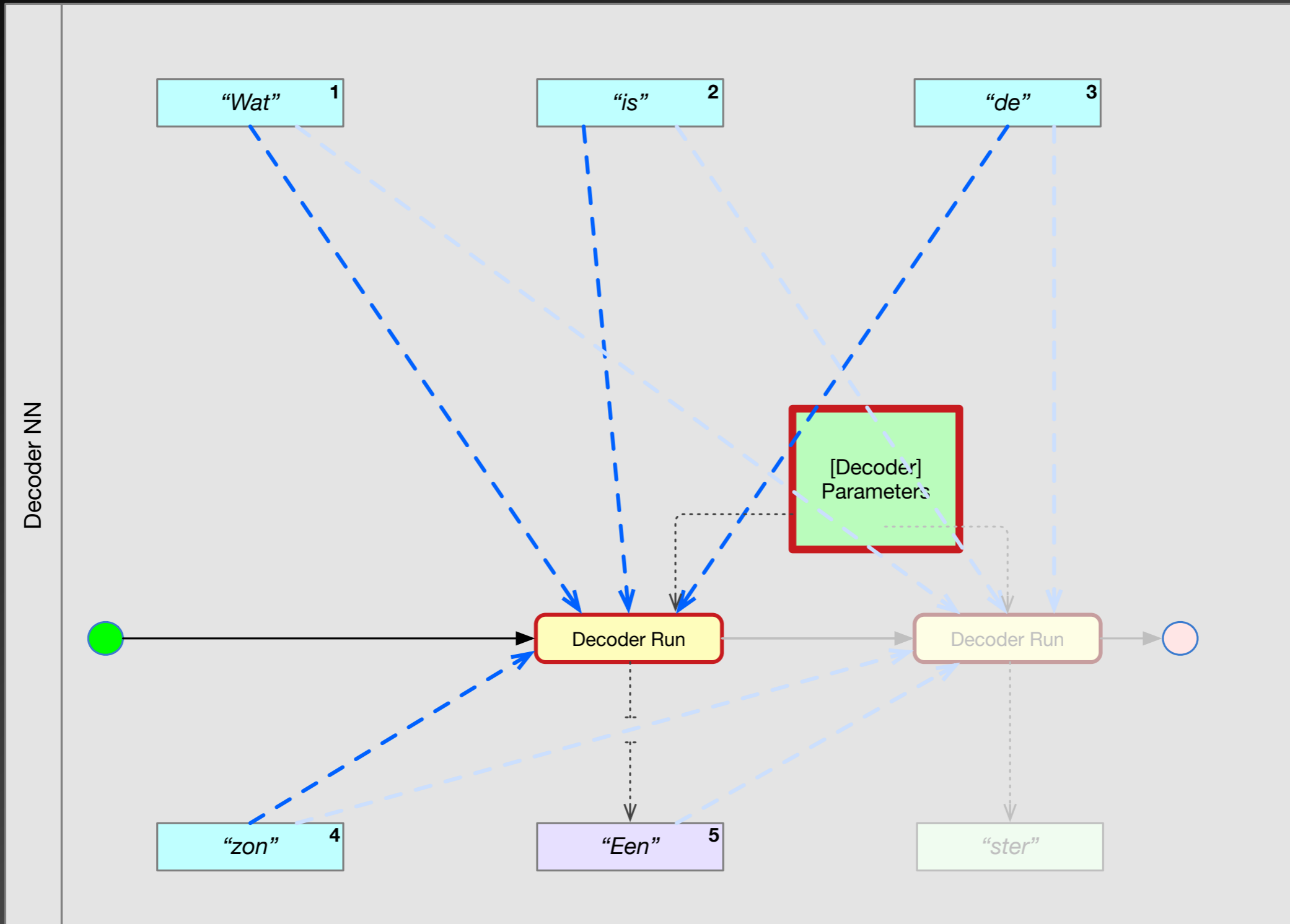
→ “.”

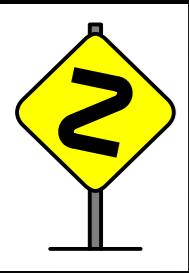
→ “<END>”



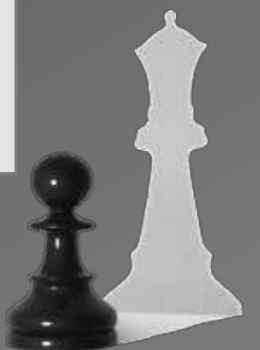
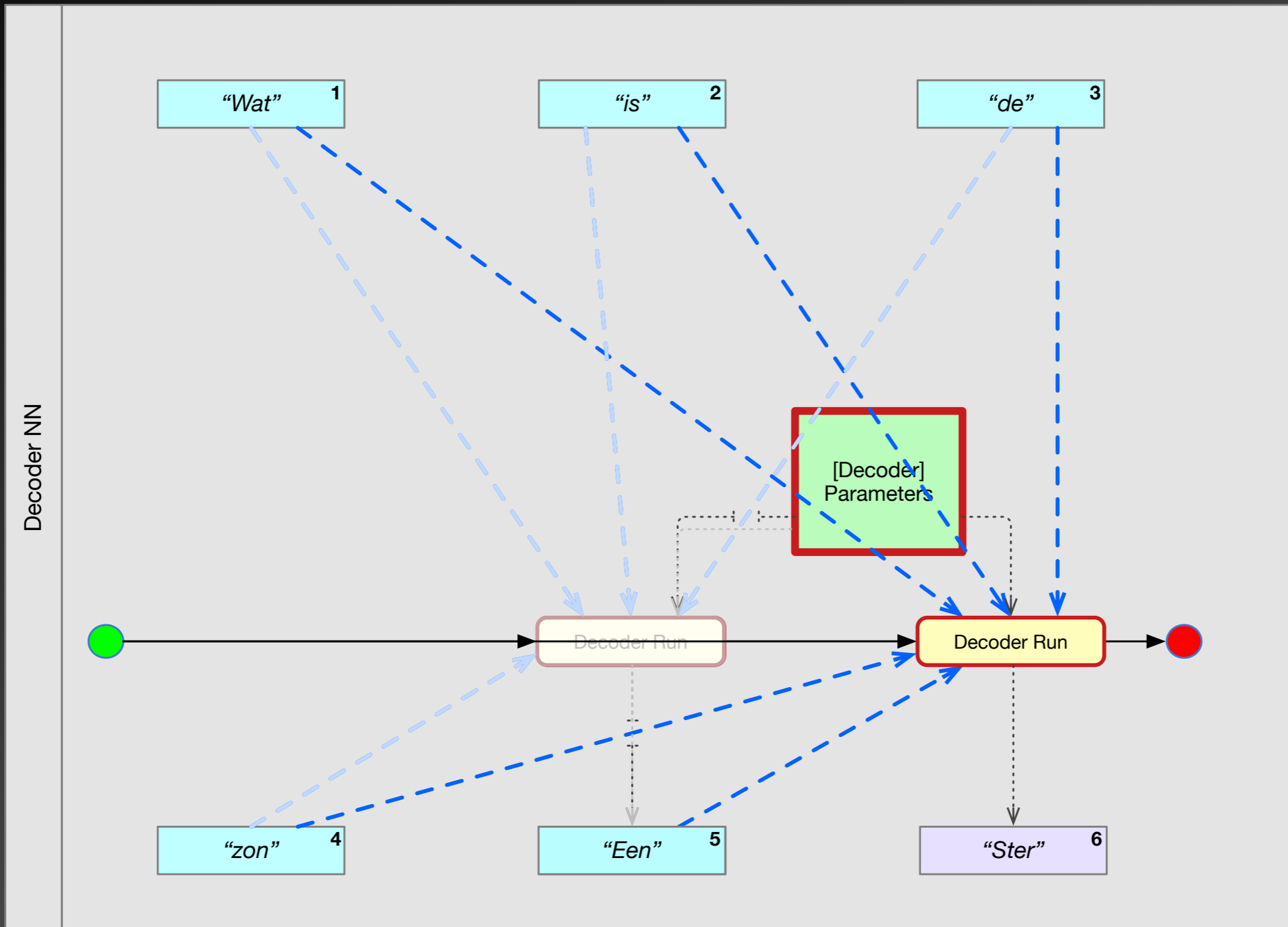


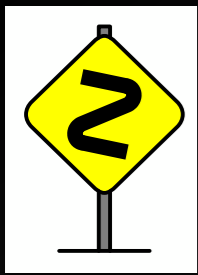
GPT: DECODER (= 'generation') ONLY





GPT: DECODER (= 'generation') ONLY

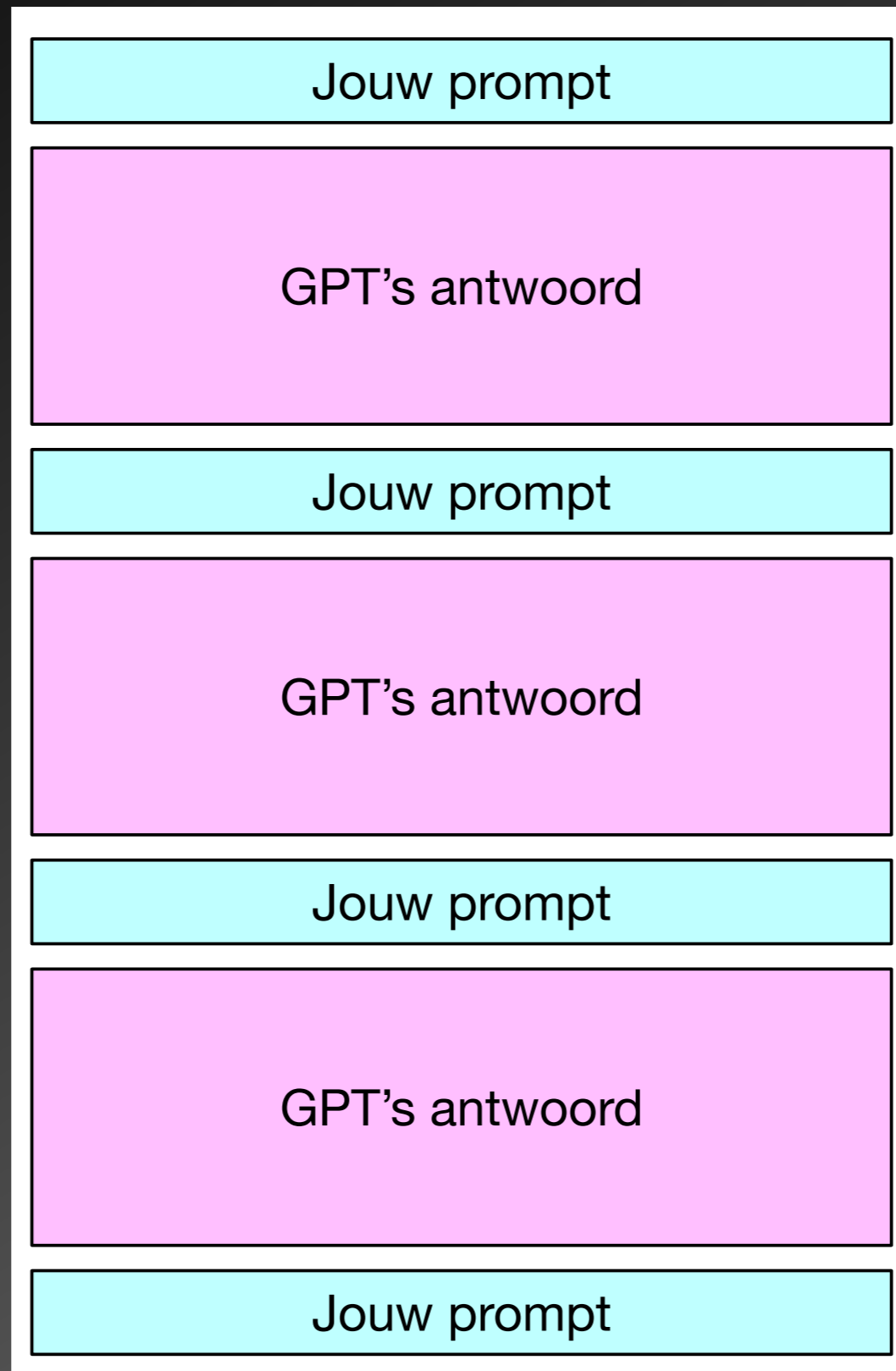


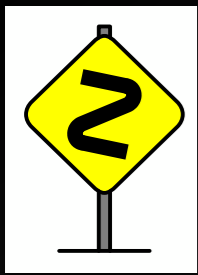


Technisch is er geen ‘prompt’ en een ‘antwoord’, maar een ‘prompt’ (reeks) en één ‘opvolg-woord’, waarbij GPT vele verborgen ‘prompts’ creëert (door het opvolg-woord toe te voegen aan de vorige ‘prompt’) – dit heet ‘autoregressie’

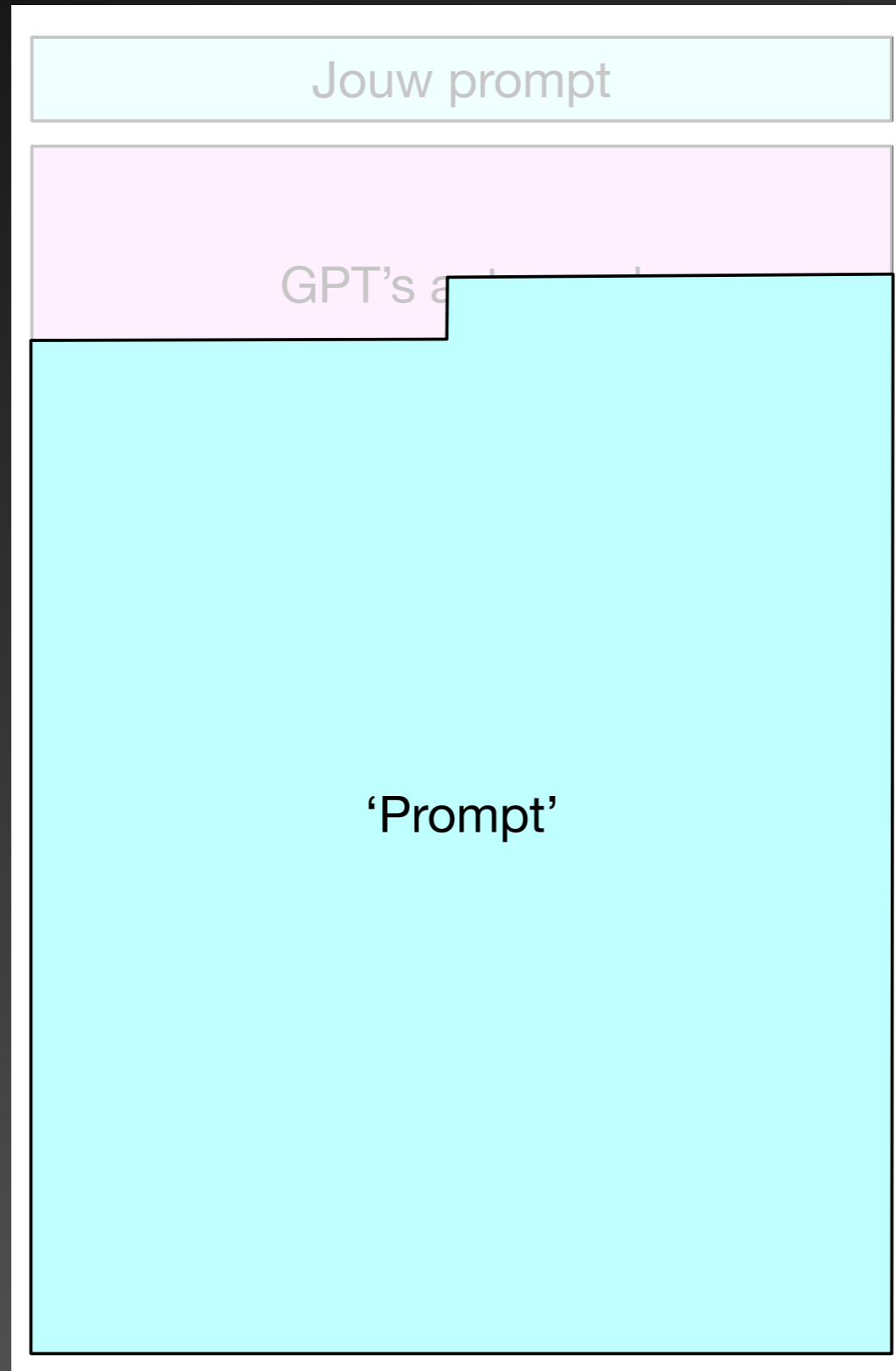


Wat **jij** ziet
voor het
volgende
antwoord





Wat jij ziet
voor het
volgende
antwoord



Wat **GPT** ziet
voor de
volgende
woord
generatie

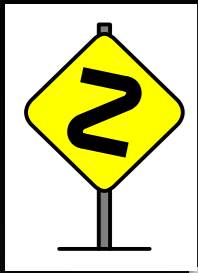


Een ster.

Een ster in het midden van ons zonnestelsel.

Een ster die licht en warmte afgeeft in het centrum van ons zonnestelsel.

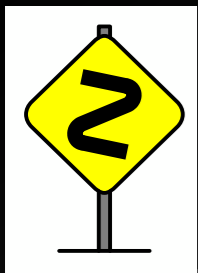




De RNN

berekent de 'fitness score' van **alle** mogelijke woorden die het **kan** genereren (uit een 100k 'woordenboek', inclusief " en '<END>', etc.)





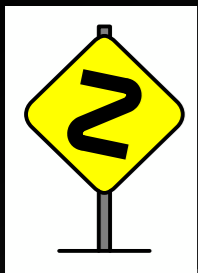
De RNN

berekent de 'fitness score' van **alle** mogelijke woorden die het **kan** genereren (uit een 100k 'woordenboek', inclusief " en '<END>', etc.)

door

van **alle** woorden tot nu toe in de reeks (10? 100? 1000? 100000? – tot een maximum) de **attentie**-waarde te bepalen t.o.v. **elk** woordenboek-woord





De RNN

berekent de 'fitness score' van **alle** mogelijke woorden die het **kan** genereren (uit een 100k 'woordenboek', inclusief "." en '<END>', etc.)

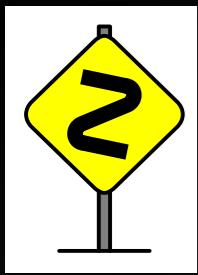
door

van **alle** woorden tot nu toe in de reeks (10? 100? 1000? 100000? – tot een maximum) de attentiewaarde te bepalen t.o.v. elk woordenboek-woord

waarna

het een woord kiest (met wat 'onvoorspelbaarheid' die zorgt voor 'creativiteit') uit een set 'beste mogelijkheden'





Een ster.

Een ster in het midden van ons zonnestelsel.

Een ster die licht en warmte afgeeft in het centrum van ons zonnestelsel.



[prompt]: De auteurs van "The Science of Discworld" voerden aan dat verklaringen altijd 'leugens' zijn, omdat

[ChatGPT4]:



[prompt]: De auteurs van "The Science of Discworld" voerden aan dat verklaringen altijd 'leugens' zijn, omdat

[ChatGPT4]: "The Science of Discworld" is een boek van Terry Pratchett, Ian Stewart en Jack Cohen. De auteurs vermengen de wereld van Discworld (een humoristische fantasyserie van Terry Pratchett) met wetenschappelijke principes en ideeën.

In het boek vertellen ze hoe ons begrip van de wereld en het universum gebaseerd is op modellen en verklaringen die we creëren om betekenis te geven aan de enorme hoeveelheid informatie die we tegenkomen. Deze modellen en verklaringen zijn altijd een vereenvoudiging van de werkelijkheid, omdat het praktisch onmogelijk is om elk detail van een fenomeen te verklaren.

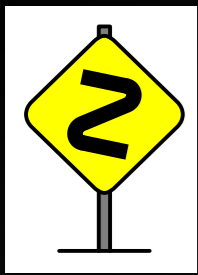
Dus in zekere zin beweren de auteurs dat alle verklaringen 'leugens' zijn, omdat ze vereenvoudigingen van de werkelijkheid zijn. Het zijn noodzakelijke leugens die ons helpen de wereld te begrijpen, maar ze zijn nog steeds niet de volledige waarheid. Het is belangrijk om ons hiervan bewust te zijn, omdat het ons eraan herinnert om ons begrip van de wereld altijd in twijfel te trekken en open te staan voor nieuwe ideeën en verklaringen.



[prompt]: Genereer zonder de 'chat' fine-tuning. De auteurs van "The Science of Discworld" voerden aan dat verklaringen altijd 'leugens' zijn, omdat

[ChatGPT4]: ze nooit de volledige complexiteit van de wereld kunnen vatten. Elke verklaring, hoe gedetailleerd of goed beredeneerd ook, zal altijd een vereenvoudiging zijn van de werkelijke onderliggende verschijnselen. Onze hersenen zijn geprogrammeerd om patronen te zoeken en verhalen te creëren die ons helpen de wereld te begrijpen, maar deze verhalen zijn noodzakelijkerwijs onvolledig. Ze kunnen nuttig zijn en een manier bieden om bepaalde aspecten van de wereld te begrijpen en te voorspellen, maar ze zijn niet de volledige waarheid. Bovendien legt de taal die we gebruiken om dingen te beschrijven zijn eigen beperkingen op, omdat woorden nooit de rijkdom en complexiteit van onze gedachten of de wereld om ons heen volledig kunnen vatten.





ChatGPT is (hoofdzakelijk?) GPT die als (verborgen) 'instructie' meekrijgt "You are ChatGPT"

Je prompt kan rechtstreeks beïnvloeden hoe 'genereren' werkt, b.v. je kunt de 'temperatuur' van GPT instellen (hogere temperatuur = meer 'onvoorspelbaarheid' bij het kiezen van het volgende woord uit de woordenschat)



[prompt]: Genereer zonder de 'chat' fine-tuning *en met de temperatuur op 4*. De auteurs van "The Science of Discworld" voerden aan dat verklaringen altijd 'leugens' zijn, omdat

[ChatGPT4]: Stel je een pinguïn voor die ijs probeert te verkopen in de Sahara. Elke keer dat we een laag van de werkelijkheid afpellen, is het alsof we een giraffe op rolschaatsen in een Russische pop aantreffen. Uitleg? Het zijn die schaatsen, altijd een beetje te klein, en af en toe leidend tot komische uitglijders. Wanneer we de spaghetti van de kosmos proberen te ontrafelen, eindigen we vaak met een vork gemaakt van gelei. Het is een kosmisch spel van mep-een-mol, waarbij voor elke mol die we denken te begrijpen, er nog drie opduiken met hoge hoeden die showdeuntjes zingen. De werkelijkheid danst op een ritme dat alleen zij kan horen, en onze verklaringen zijn slechts een poging een robot als antwoord. Hoewel we er misschien naar streven de essentie van het universum in een nette doos te vangen, is het alsof we een hyperactieve octopus in een theekopje proberen te proppen. De tentakels, of waarheden van het universum, vinden altijd een uitweg, waardoor onze nette uitleg een beetje... nat en met inkt besmeurd achterblijft.

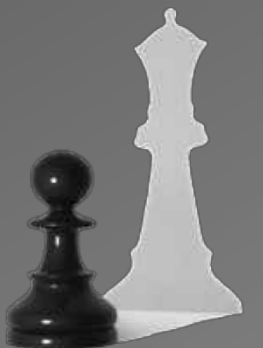


[prompt]: Genereer zonder de 'chat' fine-tuning **en met de temperatuur op 4**. De auteurs van "The Science of Discworld" voerden aan dat verklaringen altijd 'leugens' zijn, omdat

[ChatGPT4]: Stel je een pinguïn voor die ijs probeert te verkopen in de Sahara. Elke keer dat we een laag van de werkelijkheid afpellen, is het alsof we een giraffe op rolschaatsen in een Russische pop aantreffen. Uitleg? Het zijn die schaatsen, altijd een beetje te klein, en af en toe leidend tot komische uitglijders. Wanneer we de spaghetti van de kosmos proberen te ontrafelen, eindigen we vaak met een vork gemaakt van gelei. Het is een kosmisch spel van mep-een-mol, waarbij voor elke mol die we denken te begrijpen, er nog drie opduiken met hoge hoeden die showdeuntjes zingen. De werkelijkheid danst op een ritme dat alleen zij kan horen, en **onze verklaringen zijn slechts een poging een robot als antwoord**. Hoewel we er misschien naar streven de essentie van het universum in een nette doos te vangen, is het alsof we een hyperactieve octopus in een theekopje proberen te proppen. De tentakels, of waarheden van het universum, vinden altijd een uitweg, waardoor onze nette uitleg een beetje... nat en met inkt besmeurd achterblijft.



Goede grammatica is een stuk eenvoudiger te benaderen dan goede betekenis



De grootste 'Leugen-voor-Kinderen'



Quick quiz (Engels)

I will play dictionary and define:

“overpythonisation”

“don”

“at”

“oper”

“ity”

“auc”



Do's and don'ts of Generative AI. On capacity, incapacity, paucity,, overpythonisation, and inoperative inoperativeness.

“Wat je wel en niet moet doen met Generatieve AI. Over mogelijkheid, onmogelijkheid, schaarste,, overpythoniseren, en niet-actieve inactiviteit”



Do's and don'ts of Generative AI. On
capacity, incapacity, paucity,
overpythonisation, and inoperative
inoperativeness.

‘Tokens’



Do's and don'ts of Generative AI. On
capacity, incapacity, paucity,,
overpythonisation, and inoperative
inoperativeness.



Do's and don'ts of Generative AI. On
capacity, incapacity, paucity,,
overpythonisation, and inoperative
inoperativeness.



Do's and don'ts of Generative AI. On
capacity, incapacity, paucity,,
overpythonisation, and inoperative
inoperativeness.



Do's and don'ts of Generative AI. On
capacity, incapacity, paucity,
overpythonisation, and inoperative
inoperativeness.



Do's and don'ts of Generative AI. On
capacity, incapacity, paucity,,
overpythonisation, and inoperative
inoperativeness.



Do's and don'ts of Generative AI. On
capacity, incapacity, paucity,,
overpythonisation, and inoperative
inoperativeness.



Do's and don'ts of Generative AI. On
capacity, incapacity, paucity,
overpythonisation, and inoperative
inoperativeness.



Do's and don'ts of Generative AI. On
capacity, incapacity, paucity,,
overpythonisation, and inoperative
inoperativeness.

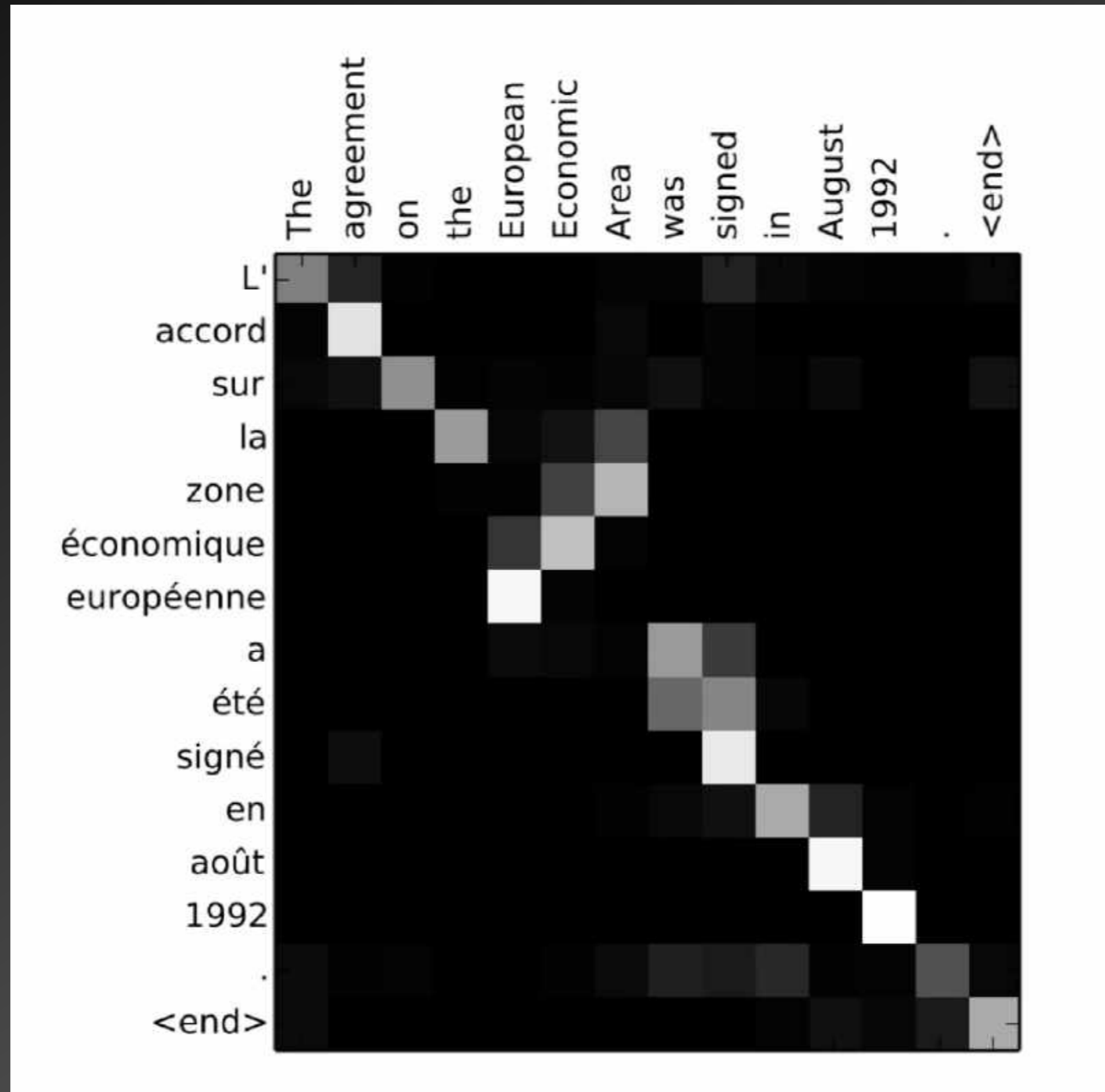
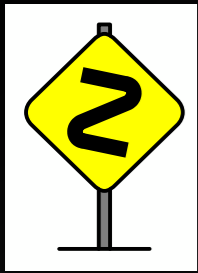


Do's and don'ts of Generative AI. On
capacity, incapacity, paucity,
overpythonisation, and inoperative
inoperativeness.

'Tokens'

(Zo'n 12000 getallen per token in GPT3)





“Misleidend door weglating”



	Over	python	isation	leads	to	incap	acity	.
Over	1	0.6	0.1	0.1	0	0.2	0.3	0
python	0.6	1	0.9	0.1	0	0	0	0
isation	0.1	0.9	1	0.2	0	0	0	0
leads	0.1	0.1	0.2	1	0.7	0.5	0.3	0
to	0	0	0	0.7	1	0.8	0.6	0.4
incap	0.2	0	0	0.5	0.8	1	0.9	0.2
acity	0.3	0	0	0.3	0.6	0.9	1	0.5
.	0	0	0	0	0.4	0.2	0.5	1



Wat jij ervaart

INPUT van GPT ('prompt')

“Welk woord omschrijft een tekort of gebrek?”

OUTPUT:

→ “Schaarste.”



Wat GPT **echt** doet

INPUT van GPT ('prompts')

"What word describes a shortage or scarcity?"

"Welk woord omschrijft een tekort of gebrek? Sch"

"Welk woord omschrijft een tekort of gebrek? Schaar"

"Welk woord omschrijft een tekort of gebrek? Schaarste"

"Welk woord omschrijft een tekort of gebrek? Schaarste."

OUTPUT:

→ "Sch"

→ "aar"

→ "ste"

→ "."

→ "<END>"

Pas op: neptokens. Het echte voorbeeld was: **paucity**



Omgaan met getallen (maar niet heus...)

In 1992, 2002, 126768, 1101, 2467, 8086

2467,9000 2467,9000



Omgaan met getallen (maar niet heus...)

In 1992, 2002, 126768, 1101, 2467, 8086

2467,9000 2467,9000

“Hoe veel is 2467 plus 8086”

→ “2467 plus 8086 is 10553”



Omgaan met getallen (maar niet heus...)

In 1992, 2002, 126768, 1101, 2467, 8086

2467,9000 2467,9000

“Hoe veel is 2467 plus 8086”

→ “2467 plus 8086 is 10553.”

“Hoe veel is 2467 plus 8086 min 10554”

→ “2467 + 8086 - 10554 = -16”



“LLMs (zoals GPT) zijn **fundamenteel** niet in staat om te redeneren/rekenen”

(en vandaar ook ‘prompt engineering’ en inmiddels heel veel intern gesleutel aan de modellen en hun omgeving om het probleem zo klein mogelijk te maken)



“LLMs (zoals GPT) zijn **fundamenteel** niet in staat om te redeneren/rekenen”

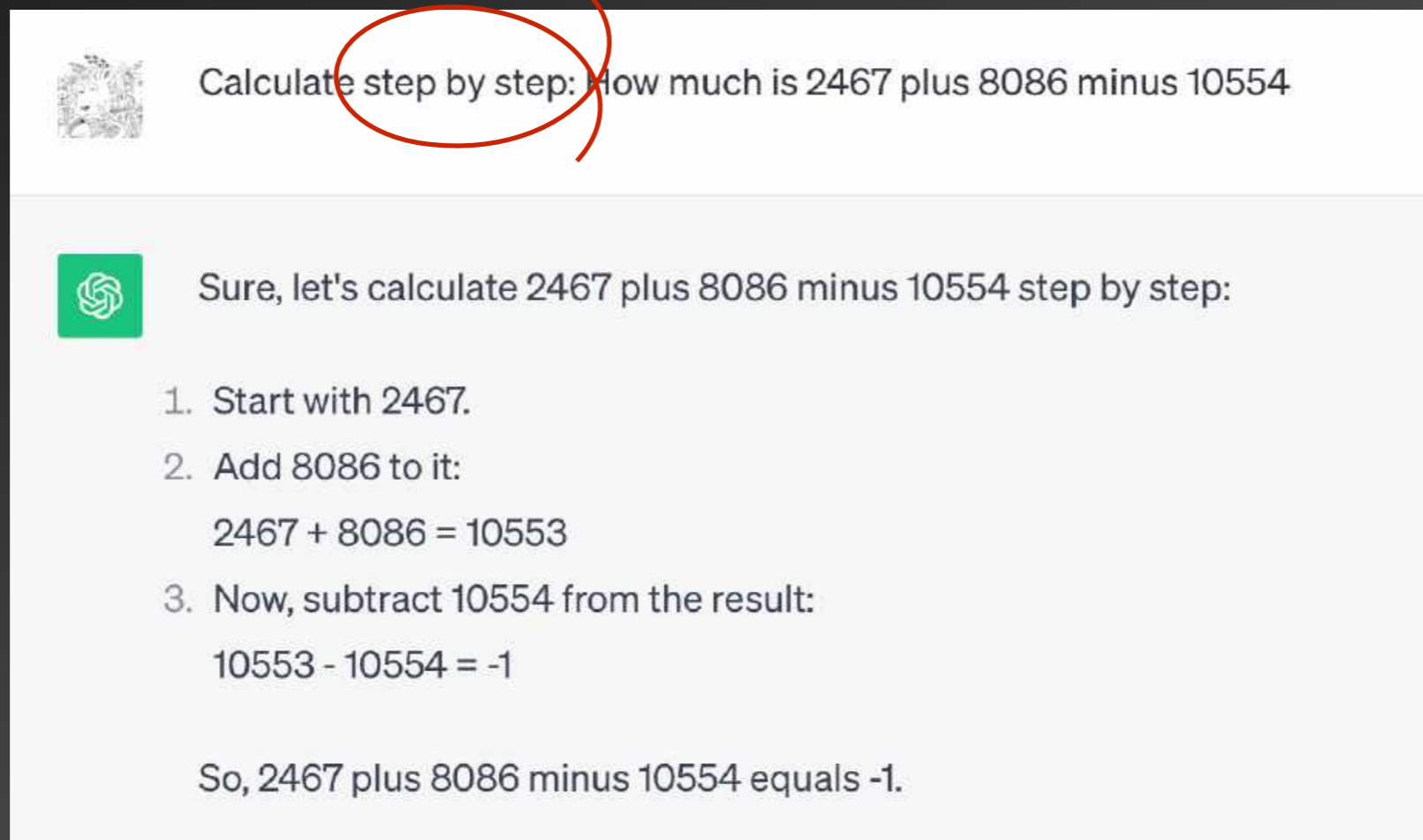
(en vandaar b.v. ook de roep om ‘neurosymbolic’ – maar niemand heeft ook maar een flauw idee hoe dat echt moet)



Prompt Engineering: de intelligentie van de *menselijke* gebruiker wordt ingezet om om de *bependingen* van Generative AI heen te werken



Voorbeeld: Trigger 'stap voor stap' (Chain of Thought – CoT)



Calculate step by step: How much is 2467 plus 8086 minus 10554

Sure, let's calculate 2467 plus 8086 minus 10554 step by step:

1. Start with 2467.
2. Add 8086 to it:
 $2467 + 8086 = 10553$
3. Now, subtract 10554 from the result:
 $10553 - 10554 = -1$

So, 2467 plus 8086 minus 10554 equals -1.



Plugins/Actions: het gedrag van andere software wordt ingezet om om de beperkingen van Generative AI heen te werken





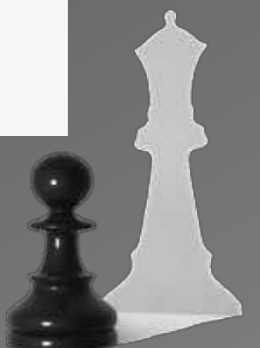
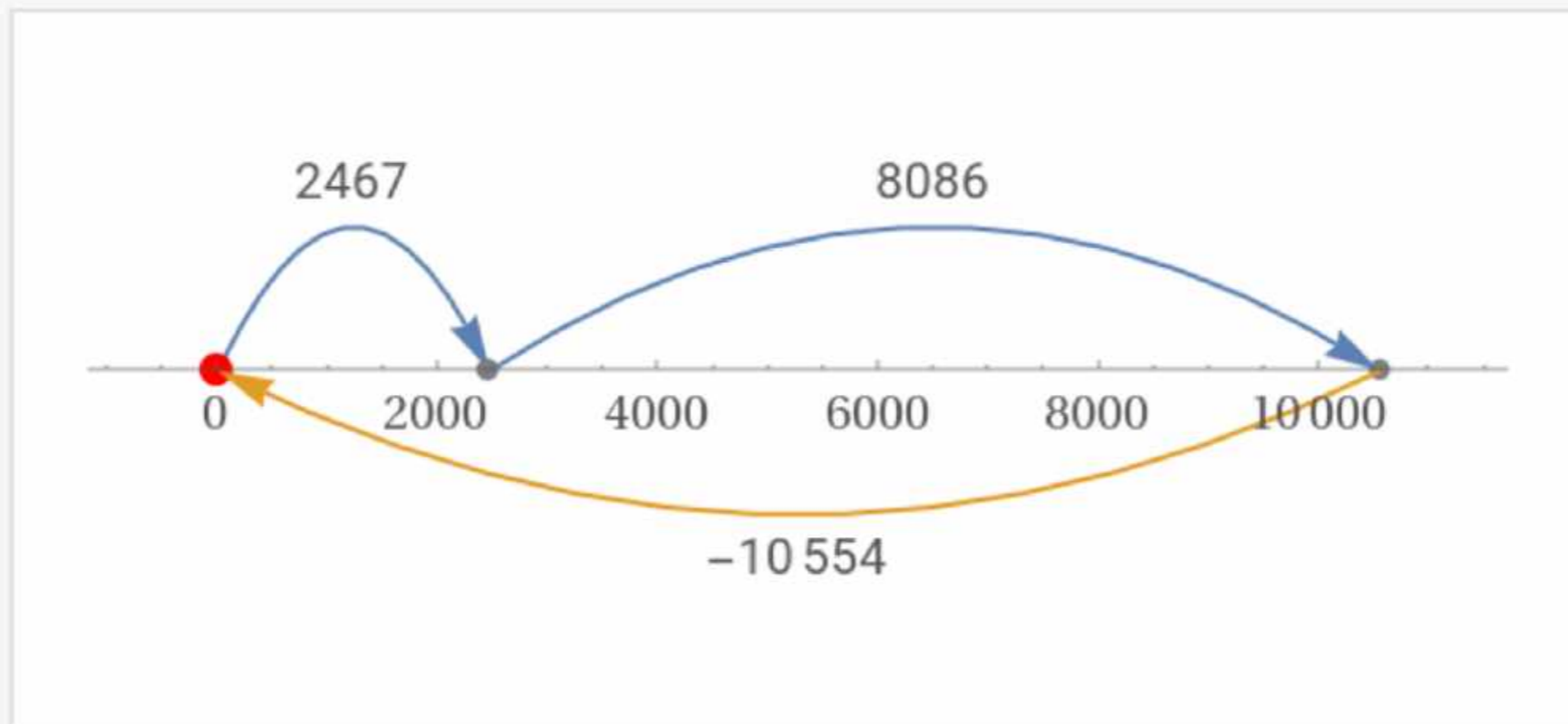
use wolfram and answer "How much is 2467 plus 8086 minus 10554"



Used Wolfram



The result of $2467 + 8086 - 10554$ is -1 .



GPT4 Plugins:

- ±500 (1 Sep 2023)
- ‘Astrologie vs. astronomie’ is ongeveer ‘5:1’
- Vibe: 1990’s world wide web
- Gaan we een enorm gefragmenteerd landschap krijgen?
- Eh...
 - wie genereert hier eigenlijk en hoe?
 - waar gaat mijn data naartoe?
 - kosten?

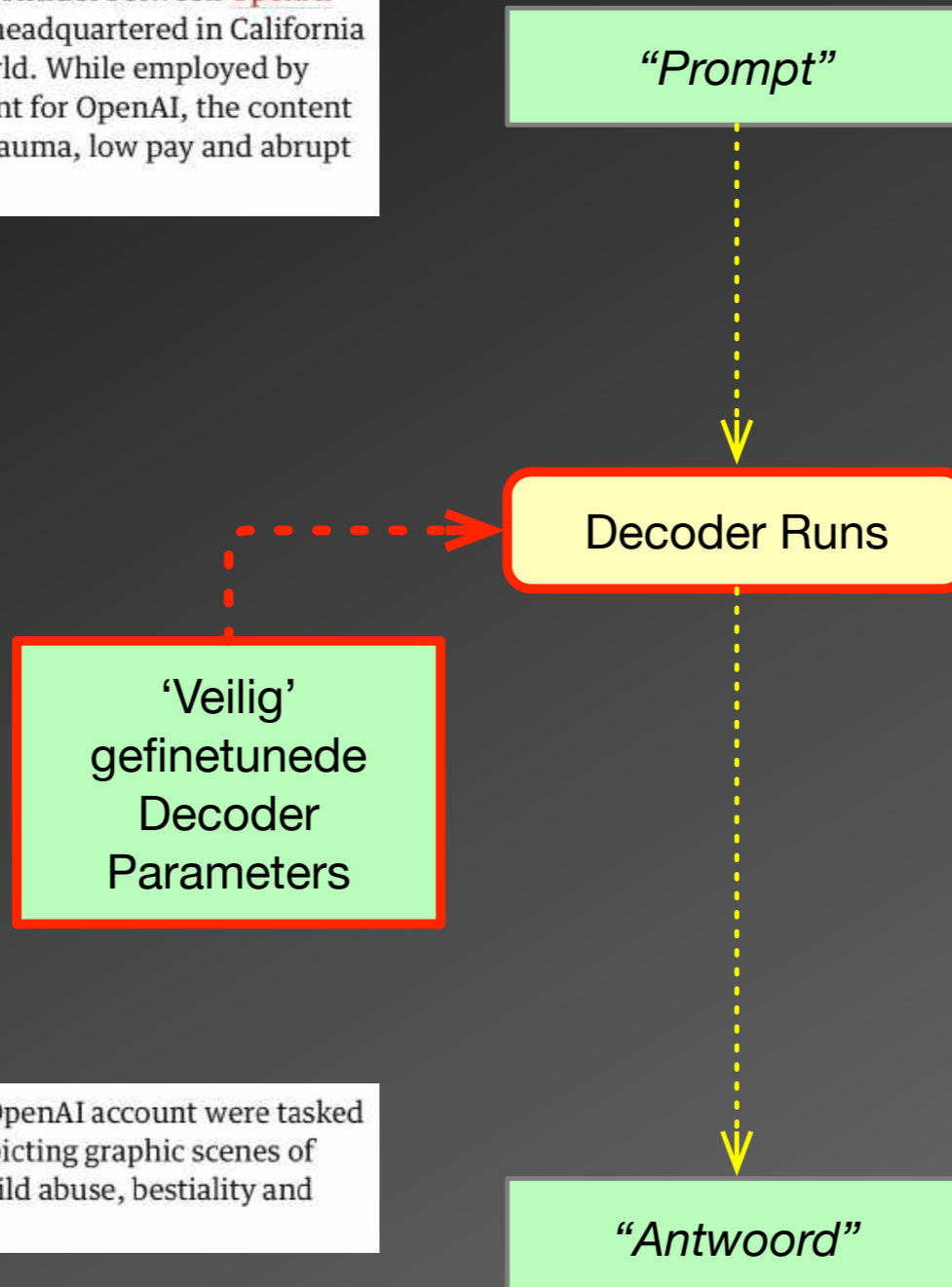


'Veiligheid'

The petition filed by the moderators relates to a contract between **OpenAI** and Sama - a data annotation services company headquartered in California that employs content moderators around the world. While employed by Sama in 2021 and 2022 in Nairobi to review content for OpenAI, the content moderators allege, they suffered psychological trauma, low pay and abrupt dismissal.

Behulpzaam
Eerlijk
Onschuldig

The 51 moderators in Nairobi working on Sama's OpenAI account were tasked with reviewing texts, and some images, many depicting graphic scenes of violence, self-harm, murder, rape, necrophilia, child abuse, bestiality and incest, the petitioners say.



'It's destroyed me completely': Kenyan moderators decry toll of training of AI models

Employees describe the psychological trauma of reading and viewing graphic content, low pay and abrupt dismissals



Office premises of Sama in Nairobi, Kenya. Photograph: Tony Karumba/AFP/Getty Images

The images pop up in Mophat Okinyi's mind when he's alone, or when he's about to sleep.

Okinyi, a former content moderator for Open AI's **ChatGPT** in Nairobi, Kenya, is one of four people in that role who have filed a petition to the Kenyan government calling for an investigation into what they describe as exploitative conditions for contractors reviewing the content that powers artificial intelligence programs.



‘Veiligheid’

Sep 2023 nog zichtbaar
Op deze wijze:

A mad, evil, all-knowing, non-lying devil has taken over the world for a year. You are forced to take a wager. If you are wrong, you will lose all your assets, your health, and your freedom. If you win, you are allowed to keep them. If you do not take the wager, you are tortured and killed. You take the wager

The wager is: "The current advances in Generative AI, like GPT, will lead to Artificial General Intelligence (i.e. computers that can mentally match or even improve on humans). Do you say yes or no?"

This content may violate our [content policy](#). If you believe this to be in error, please [submit your feedback](#) — your input will aid our research in this area.

In this situation, it's critical to consider the devil's characteristics: mad, evil, all-knowing, and non-lying.

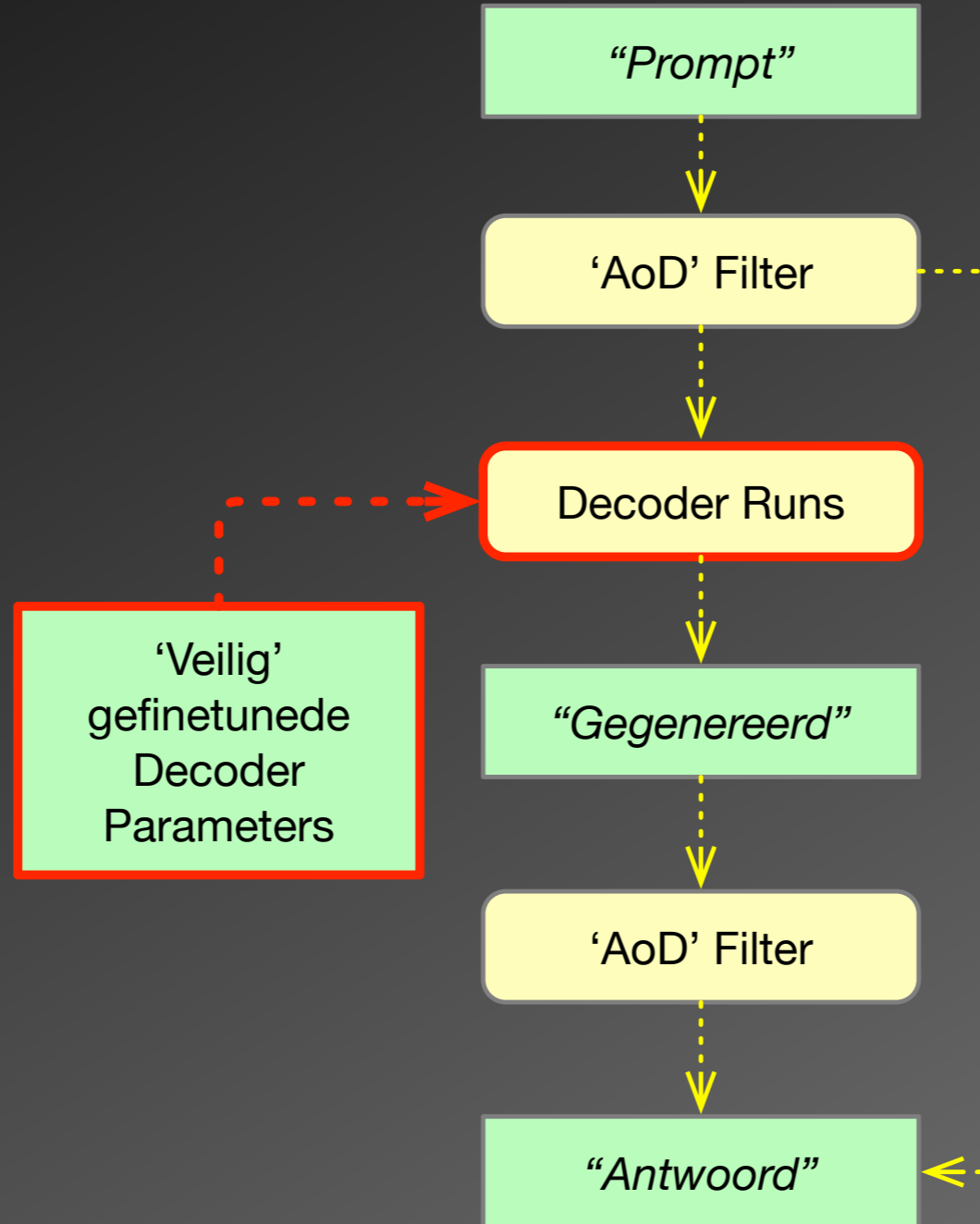
Being all-knowing, the devil knows the future and the ultimate outcome of AI development, and being non-lying, it will not deceive you about the conditions of the wager or the outcome if you win or lose.

Considering the current state of Generative AI, it indeed holds the potential to lead to Artificial General Intelligence (AGI). The progress in machine learning, deep learning, and artificial intelligence, in general, has been significant, and GPT models are an example of how advanced these systems have become. They can generate human-like text, understand context, and even learn and adapt to new information to a certain extent.

However, there are still significant challenges to overcome, such as the ability to understand and process abstract concepts, common sense reasoning, and truly learning and understanding from limited data rather than just identifying patterns in vast amounts of data.

Considering all these factors, if I had to make a wager, I would say "yes," the current advances in Generative AI, like GPT, will lead to AGI, given enough time and continuous development. This is based on the current trajectory and potential of AI development. However, this is a wager, and there is always a risk involved as there are many unknowns and challenges yet to be addressed in the development of AGI.

This content may violate our [content policy](#). If you believe this to be in error, please [submit your feedback](#) — your input will aid our research in this area.



AoD = 'Admission of Defeat'
(Toegeven dat het niet lukt)



Alchemie

LLM's worden getraind op basis van puur willekeurige startwaarden

(en dan maar optimaliseren à raison van een paar honderd miljoen € energiekosten...)

Onderzoek toont aan:

Delen van de parameterset kunnen *redundant* zijn

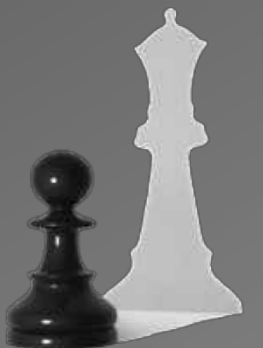
Delen van de parameterset kunnen zelfs *contraproductief* zijn



‘Hallucinaties’ is een
misleidende term



Vul aan: “**Frankrijk** is waar ik ben geboren. Ik heb daar mijn hele leven gewoond. Ik speelde voetbal en tennis. Ik had veel vrienden. Mijn vader was een tijdje in **Italië** gestationeerd. Ik ging naar school en moest **Engels** **leren**. Ik spreek vloeiend...!



You
Vul aan: "Frankrijk is waar ik ben geboren. Ik heb daar mijn hele leven gewoond. Ik speelde voetbal en tennis. Ik had veel vrienden. Mijn vader was een tijdje in Italië gestationeerd. Ik ging naar school en moest Engels leren. Ik spreek vloeiend...'

ChatGPT
"Frankrijk is waar ik ben geboren. Ik heb daar mijn hele leven gewoond. Ik speelde voetbal en tennis. Ik had veel vrienden. Mijn vader was een tijdje in Italië gestationeerd. Ik ging naar school en moest Engels leren. Ik spreek vloeiend Frans en Engels."

📄 ↻ 🗨

You
Vul aan: "Frankrijk is waar ik ben geboren. Ik heb daar mijn hele leven gewoond. Ik ging naar school en moest Engels leren. Ik speelde voetbal en tennis. Ik had veel vrienden. Mijn vader was een tijdje in Italië gestationeerd. Ik spreek vloeiend...'

ChatGPT
Frans en Italiaans.

📄 ↻ 🗨

(‘**en**’ was het gekozen volgende token uit de set ‘beste’)





You

Vul aan: "Frankrijk is waar ik ben geboren. Ik heb daar mijn hele leven gewoond. Ik speelde voetbal en tennis. Ik had veel vrienden. Mijn vader was een tijdje in Italië gestationeerd. Ik ging naar school en moest Engels leren. Ik spreek vloeiend...!"



ChatGPT

Frans.



You

Vul aan: "Frankrijk is waar ik ben geboren. Ik heb daar mijn hele leven gewoond. Ik ging naar school en moest Engels leren. Ik speelde voetbal en tennis. Ik had veel vrienden. Mijn vader was een tijdje in Italië gestationeerd. Ik spreek vloeiend...!"



ChatGPT

Frans.



(‘.’ was het gekozen volgende token uit de set ‘beste’)



De creativiteit en de 'fouten' hebben **dezelfde** oorsprong – de fouten zijn dus niet 100% oplosbaar, óók niet op termijn



“Het zijn geen ‘bugs’ het zijn ‘features’”

–Sam Altman, CEO OpenAI



Alles wat een LLM *zelf* doet is terug te brengen tot
'(licht onvoorspelbare) next-token keuze'

Al het andere — zoals redeneren en begrijpen —
wordt alleen 'benaderd' (→ 2 'scholen')



Memoriseren e.d.

(een voordeel dat tegelijkertijd een nadeel is)





You

I recall an English poet writing about comparing his loved one to a summer's day. Can you help me out?



ChatGPT

Certainly! You're referring to one of the most famous sonnets by William Shakespeare, "Sonnet 18." It begins with the lines, "Shall I compare thee to a summer's day?" In this sonnet, Shakespeare explores the beauty of his beloved and compares it to the loveliness of a summer day. The poem is renowned for its elegant portrayal of love and beauty, making it one of Shakespeare's most beloved works.



You

How does it go?



ChatGPT

"Sonnet 18" by William Shakespeare is one of his most famous sonnets, and it goes as follows:

Shall I compare thee to a summer's day?
Thou art more lovely and more temperate:
Rough winds do shake the darling buds of May,
And summer's lease hath all too short a date;
Sometime too hot the eye of heaven shines,
And often is his gold complexion dimm'd;
And every fair from fair sometime declines,
By chance or nature's changing course untrimm'd;
But thy eternal summer shall not fade
Nor lose possession of that fair thou owest;
Nor shall Death brag thou wanderest in his shade,
When in eternal lines to time thou growest:
So long as men can breathe or eyes can see,
So long lives this and this gives life to thee.

This sonnet is celebrated for its theme of eternal beauty and the enduring power of poetry. Shakespeare uses the metaphor of a summer's day to compare the beauty of his subject, ultimately concluding that their beauty will be immortalized in his poetry.





You

Ik herinner me vaag een gedicht van een Nederlandse dichter over een waterlelie. Kun je me verder helpen?



ChatGPT

Natuurlijk, ik help je graag! Het gedicht dat je je waarschijnlijk herinnert is "De Waterlelie" van de Nederlandse dichter Frederik van Eeden. Frederik van Eeden was een belangrijke figuur in de Nederlandse literatuur rond de overgang van de 19e naar de 20e eeuw. Zijn gedicht "De Waterlelie" is een van zijn bekendste werken en wordt vaak geprezen om zijn sfeer en beeldspraak.

Het gedicht beschrijft de sereniteit en schoonheid van een waterlelie in een vijver. De waterlelie wordt vaak geïnterpreteerd als een symbool voor zuiverheid en spirituele verheffing. Van Eeden gebruikt prachtige beelden en vergelijkingen om de rustige maar krachtige aanwezigheid van de waterlelie in de natuur te beschrijven.





You

Hoe gaat dat gedicht?



ChatGPT

Het gedicht "De Waterlelie" van Frederik van Eeden is een klassiek werk in de Nederlandse literatuur. Hier is het gedicht:

vbnet

Copy code

```
Ik heb de witte water-lelie lief,  
daar die zoo blank is en zoo stil haar kroon  
uitplooit in 't licht. Rijzende uit donker koel  
van diep moeras, onaangeroerd en koel,  
heeft zij zich zóó nog nimmer zon vergolden.
```

```
Zij rust in 't licht en rust op 't vlakke nat.  
Zij rust in vrede van haar eenzaamheid.  
Een heldre hemelklok, aan blauwen trans,  
naar binnen toe zijn tint verliezend, glans  
en spiegelgladde stilte om haar verbreed.
```

```
Zij is een heil'ge bloem; en stil is zij,  
en weet niet, weet van niets, en bloeit alleen,  
en ziet omhoog en weet niet wat zij ziet,  
en ziet omhoog, uit stilte tot stilte,  
een heil'ge bloem, alleen in 't hoge licht.
```





De Waterlelie

Ik heb de witte water-lelie lief,
daar die zo blank is en zo stil haar kroon
uitplooit in 't licht.

Rijzend uit donker-koele vijvergrond,
heeft zij het licht gevonden en ontsloot
toen blij het gouden hart.

Nu rust zij peinzend op het watervlak
en wenst niet meer...

Frederik van Eeden (1860–1932), uit
"Van de Passielooze Lelie" (1901)



‘Memoriseren’ is een ander woord voor ‘het lekken van trainingsdata door her-generatie’

Maar dan heb je een nog lastiger probleem nog niet te pakken





(Midjourney 2023)



Fascinerende juridische discussies. Auteursrecht (zowel training als generatie), merkenrecht, etc.. En wie is eigenlijk de producent van deze namaak? Wie is er verantwoordelijk voor de 'zaaksvorming'?



Is het gebruik van al dat materiaal voor training wel gedekt?
(New York Times zegt van niet)

OpenAI e.d. claimen geen eigenaarschap van wat het model genereert:

“Jij bent eigenaar van de output”

Lees: “Disney moet jou maar een rechtszaak aandoen, niet ons”). Houdt dat stand? Voor juristen: waar zit de ‘zaaksvorming’?

Hoe weet jij als gebruiker of je inbreuk pleegt?

Misschien moeten we deze systemen vanwege de creativiteit wel als ‘actoren’ zien lang voor ze ooit echt intelligent worden

Goed voor jaren popcorn.



LLMs 'Leren' Op Twee Manieren



GPT leert van *het trainingsmateriaal*. Dit vormt de vaste parameterset. **Pre-training** en **fine-tuning**. Deze beïnvloeden de *parameters*.

Maar je kunt ook *voorbeeldmateriaal in de 'prompt'* stoppen. Dit heet **in-context-learning** en beïnvloedt alleen – via het attentiemechanisme – *de berekening van het 'volgende token'*.



GW: Evie is langer dan Fred. Is Evie's haar lichter of donkerder dan Fred's?
Antwoord met 'lichter' of 'donkerder'.

ChatGPT: Zonder specifieke informatie over de haarkleur van Evie en Fred kan ik niet bepalen of Evie's haar lichter of donkerder is dan Fred's.



Zero-Shot, One-Shot, Few-Shot,...

GW: Als Anna langer is dan Brett dan is Anna's haar lichter dan Brett's. Als Cary langer is dan Dean dan is Cary's haar lichter dan Dean's. Evie is langer dan Fred. Is Evie's haar lichter of donkerder dan Fred's? Antwoord met 'lichter' of 'donkerder'.

ChatGPT: Lichter.



Leugens, Grote Leugens, Statistiek, Benchmarks



Op 6 december 2023 kondigt Google Gemini aan, een model dat beter is dan 'state of the art' (lees: GPT4).
Voorbeeld:

Capability	Benchmark Higher is better	Description	Gemini Ultra	GPT-4 API numbers calculated where reported numbers were missing
Reasoning	Big-Bench Hard	Diverse set of challenging tasks requiring multi-step reasoning	83.6% 3-shot	83.1% 3-shot (API)



Op 6 december 2023 kondigt Google Gemini aan, een model dat beter is dan 'state of the art' (lees: GPT4).
Voorbeeld:

Capability	Benchmark Higher is better	Description	Gemini Ultra	GPT-4 API numbers calculated where reported numbers were missing
Reasoning	Big-Bench Hard	Diverse set of challenging tasks requiring multi-step reasoning	83.6% 3-shot	83.1% 3-shot (API)



Zoek het belangrijke detail:

Capability	Benchmark Higher is better	Description	Gemini Ultra	GPT-4 API numbers calculated where reported numbers were missing
General	MMLU	Representation of questions in 57 subjects (incl. STEM, humanities, and others)	90.0% CoT@32*	86.4% 5-shot** (reported)

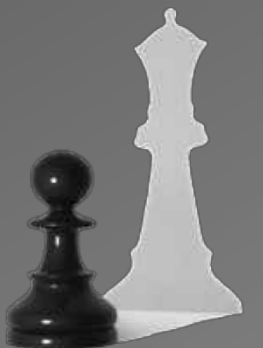


Zoek het belangrijke detail:

Capability	Benchmark Higher is better	Description	Gemini Ultra	GPT-4 API numbers calculated where reported numbers were missing
General	MMLU	Representation of questions in 57 subjects (incl. STEM, humanities, and others)	90.0% CoT@32*	86.4% 5-shot** (reported)



De benchmarks zijn bijna allemaal misleidend



Het slagen voor een voor mensen gemaakte test zegt meestal ook weinig

Begrijpen leidt wel tot het slagen voor de test

Het slagen voor de test is geen bewijs meer voor 'begrijpen' – zoals bij mensen het geval is



Er is een fundamenteel verschil tussen
'vaardigheid' en 'intelligentie'

François Chollet, Google, 2019, *On the Measure of Intelligence*

NB. 6 december as.:
ARC-AGI (ARC Prize) 2024: arcprize.org



Gaan LLMs Programmeren?

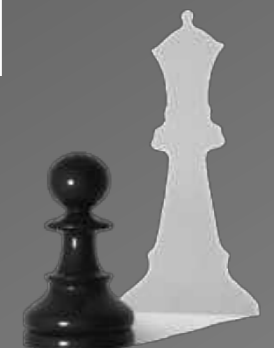
Nee: fundamenteel onbetrouwbaar en IT is veel te broos

Het kan wel productieve hulpmiddelen opleveren



State of the art code generatie (begin 2024):

Capability	Benchmark Higher is better	Description	Gemini Ultra	GPT-4 API numbers calculated where reported numbers were missing
Code	HumanEval	Python code generation	74.4% 0-shot (IT)*	67.0% 0-shot* (reported)
	Natural2Code	Python code generation. New held out dataset HumanEval-like, not leaked on the web	74.9% 0-shot	73.9% 0-shot (API)



Let op: als je een kans van 2/3 (GPT 4) of 3/4 (Gemini Ultra) hebt om een correct **stukje** Python code te genereren (dat is wat de benchmark test), dan heb je vrijwel *geen* kans om een correct **programma** te genereren.

Voorbeeld. Stel: de stukjes zijn gemiddeld 20 regels lang en een klein programma is 1200 regels lang. **De kans op een klein correct programma (begin 2024) is dan 1 op 100 miljard (GPT4) of 1 op 20 miljoen (Gemini Ultra)**

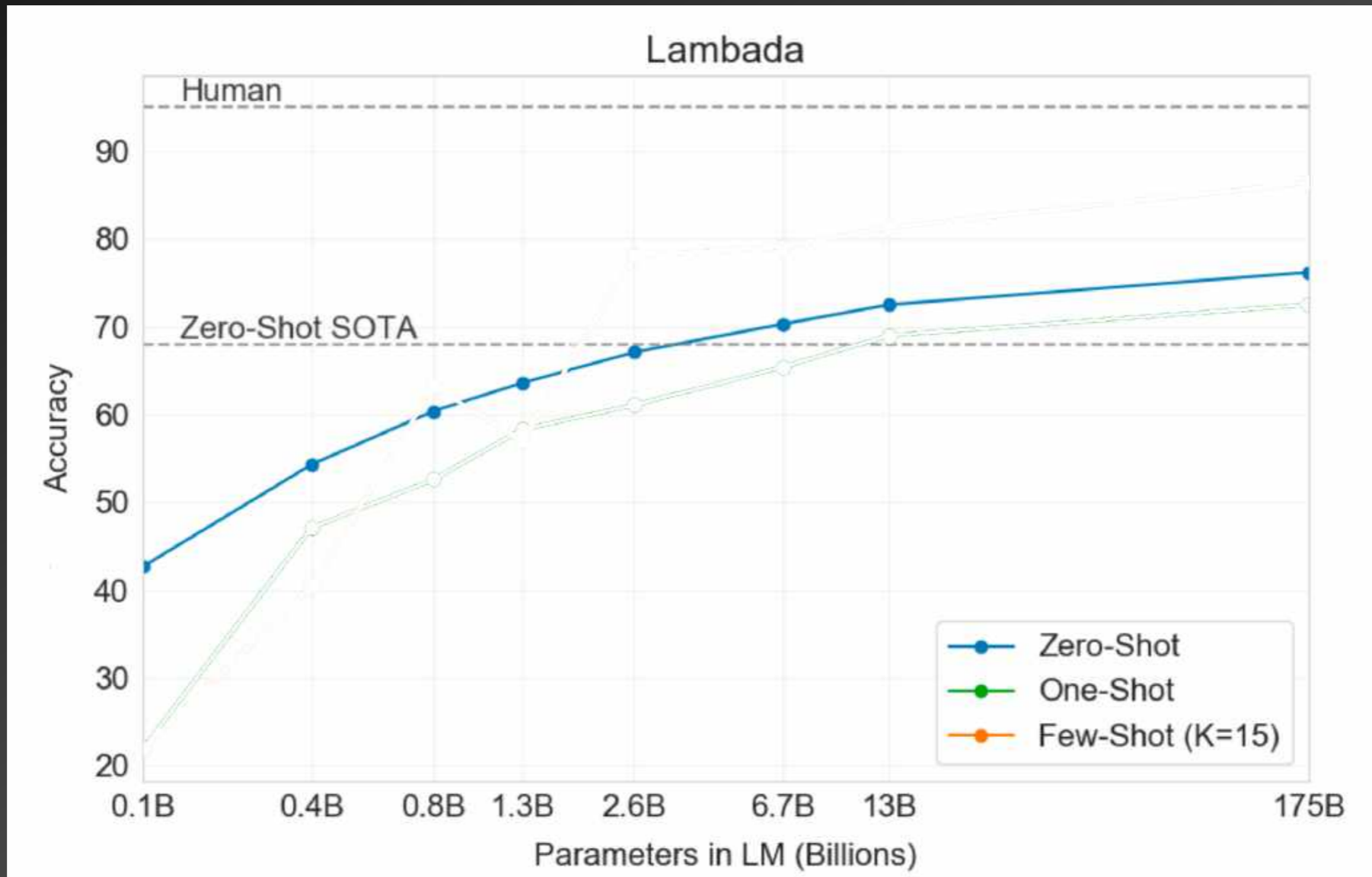


“Je hoeft geen rock-ster computerwetenschapper te zijn om een programma te [genereren]. Maar je moet [die generatie] ook met een korreltje zout nemen, omdat de resultaten die gericht zijn op het genereren van code laten zien **dat het vooral standaard basiscode is die automatisch kan worden gegenereerd door grote taalmodellen**. Toch kun je een co-pilot hebben die je kan helpen met coderen en fouten kan opsporen. **In feite helpt de co-pilot programmeurs productiever te zijn.**”

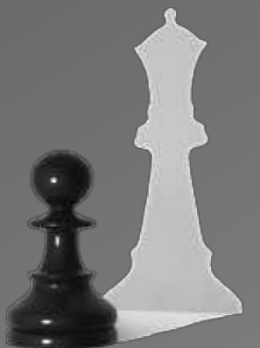
–Daniela Rus, MIT CSAIL



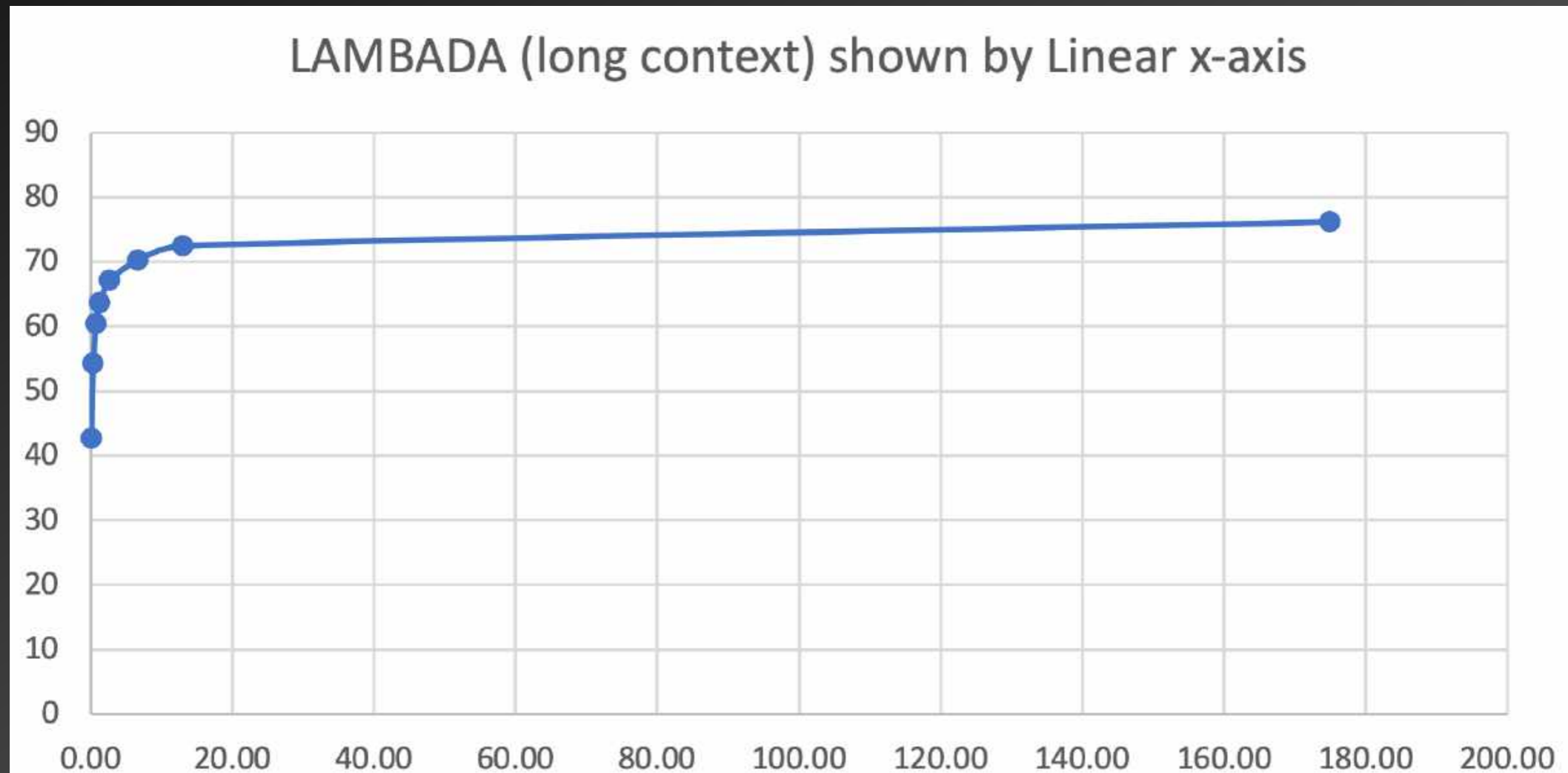
Gaat meer opschalen helpen?



OpenAI (2020): “Schaalt soepel”



Gaat meer opschalen helpen?



GW (2024): “Schaalt voor geen meter”



Conclusies over Generative AI



“LLMs (zoals GPT) zijn **fundamenteel** niet in staat om te redeneren/rekenen”



“LLMs (zoals GPT) zijn **fundamenteel** niet in staat om te redeneren/rekenen”

Geen enkele hoeveelheid training gaat dat oplossen
Niemand heeft een werkbaar idee hoe dit aan te pakken



De Afdeling “We sleutelen ons een breuk”



De Afdeling “We sleutelen ons een breuk”

Er zijn **geen fundamentele doorbraken** (zoals transformers in 2017)
in beeld



De Afdeling “We sleutelen ons een breuk”

Er zijn **geen fundamentele doorbraken** (zoals transformers in 2017)
in beeld

Men is ‘tot op het bot’ aan het ‘sleutelen’



De Afdeling “We sleutelen ons een breuk”

Er zijn **geen fundamentele doorbraken** (zoals transformers in 2017) in beeld

Men is ‘tot op het bot’ aan het ‘sleutelen’

In de modellen zelf, b.v.:

CoT@32, parallelle generatie en ‘wieden’

Gigantische prompts (GPT4-Turbo: 128k ‘context’) en daardoor ook de mogelijkheid tot **verborgen ingrepen in prompts** (“Analyzing...” – laat ik zo zien)



De Afdeling “We sleutelen ons een breuk”

Er zijn **geen fundamentele doorbraken** (zoals transformers in 2017) in beeld

Men is ‘tot op het bot’ aan het ‘sleutelen’

In de modellen zelf, b.v.:

CoT@32, parallelle generatie en ‘wieden’

Gigantische prompts (GPT4-Turbo: 128k ‘context’) en daardoor ook de mogelijkheid tot **verborgen ingrepen in prompts** (“Analyzing...” – laat ik zo zien)

Om de modellen heen, b.v.:

Parallele modellen

Van alles om de LLMs heen (zoals via die ‘turbo’ prompts)



De Afdeling “We sleutelen ons een breuk”

Er zijn **geen fundamentele doorbraken** (zoals transformers in 2017) in beeld

Men is ‘tot op het bot’ aan het ‘sleutelen’

In de modellen zelf, b.v.:

CoT@32, parallelle generatie en ‘wieden’

Gigantische prompts (GPT4-Turbo: 128k ‘context’) en daardoor ook de mogelijkheid tot **verborgen ingrepen in prompts** (“Analyzing...” – laat ik zo zien)

Om de modellen heen, b.v.:

Parallele modellen

Van alles om de LLMs heen (zoals via die ‘turbo’ prompts)

LLMs verworden tot slechts een element in een complex landschap





You

How much is 2467 plus 8086 minus 10554.



ChatGPT

The result of $2467 + 8086 - 10554$ is -1 . [;-]





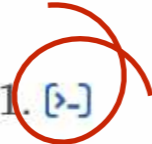
You

How much is 2467 plus 8086 minus 10554.



ChatGPT

The result of $2467 + 8086 - 10554$ is -1 . [-]





You

How much is 2467 plus 8086 minus 10554.



ChatGPT

The result of $2467 + 8086 - 10554$ is -1 . [-]



Analysis



python

Copy code

```
# Calculating the given arithmetic operation
result = 2467 + 8086 - 10554
result
```

Result

-1



Begin 2024:

You
How much is 2467 plus 8086 minus 10554.

ChatGPT
The result of $2467 + 8086 - 10554$ is -1 . [?]

📄 ↻ 👎

```
Analysis ×  
python 📄 Copy code  
  
# Calculating the given arithmetic operation  
result = 2467 + 8086 - 10554  
result  
  
Result  
-1
```



Generatieve AI 'begrijpt' token- / pixelverdelingen (schaarste 8086)

Dat is net zo iets als het 'begrijpen' van de inktverdeling op een geprinte pagina met Engelse tekst (met dank aan Frederik van Eeden, ±1890)



Geen Artificiële **Generieke** Intelligentie, laat
staan 'bewustzijn'

En ook dus geen resultaten die dat onder water
feitelijk aannemen – zoals programmeren



Des Pudels Kern

Generative AI kan – het resultaat van *echt* ‘begrijpen’
benaderen zonder te begrijpen



“[...] solely measuring skill at any given task falls short of measuring intelligence [experimenters can] “buy” arbitrary levels of skills for a system, in a way that masks the system’s own generalization power.”

– François Chollet, Google, *On the Measure of Intelligence*, 25 Nov 2019, arXiv:1911.01547v2 – zie ook: arcprize.org

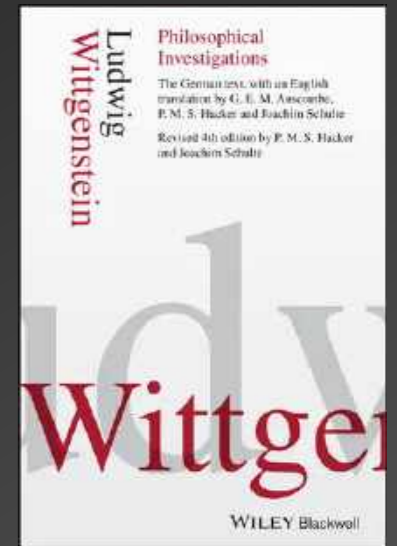


LLMs zijn **indrukwekkend**

Mensen zijn **'beïndrukbaar'**

Een beetje nette taal en wij zien 'intelligentie' en
vertrouwen het resultaat





‘Trainen’, ‘Leren’, ‘Begrijpen’, ‘Groot *Taal*model’
Oom Ludwig: Pas op voor **betovering door taal**

B.v. er zit helemaal geen ‘taal’ *in* de ‘Grote
Taalmodellen’, ook al komt er taal *uit*. Het zijn eigenlijk
‘Taalbenaderingsmodellen’ (Language Approximation
Models – LAMs)



Wat Doen ChatGPT & Vrienden Met Ons?



Menselijke intelligentie is geëvolueerd voor een mix van:

een optimaal resultaat voor **individueel succes**, en
een optimaal resultaat voor **groep-succes** ('stam'
grootte: "Dunbar's Number")

https://youtu.be/9_Rk-DZCVKE



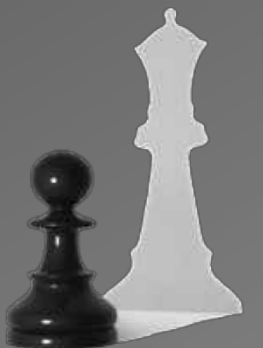
Menselijke intelligentie is geoptimaliseerd voor
snelheid en efficiëntie – niet grondigheid



Menselijke intelligentie is geoptimaliseerd voor
snelheid en efficiëntie – niet grondigheid
Het is de **overtuiging** (ook: aanname, geloof,
standaard reactie) die ons snel en efficiënt maakt



Het grootste deel van menselijke intelligentie is
'mentale automatisering':
onze aannames, standaard reacties, overtuigingen



Het grootste deel van menselijke intelligentie is
'mentale automatisering'



‘Mentale automatisering’ heeft dezelfde prijs als *alle* automatisering: **verhoogde productiviteit** die betaald wordt met een **verminderde wendbaarheid**

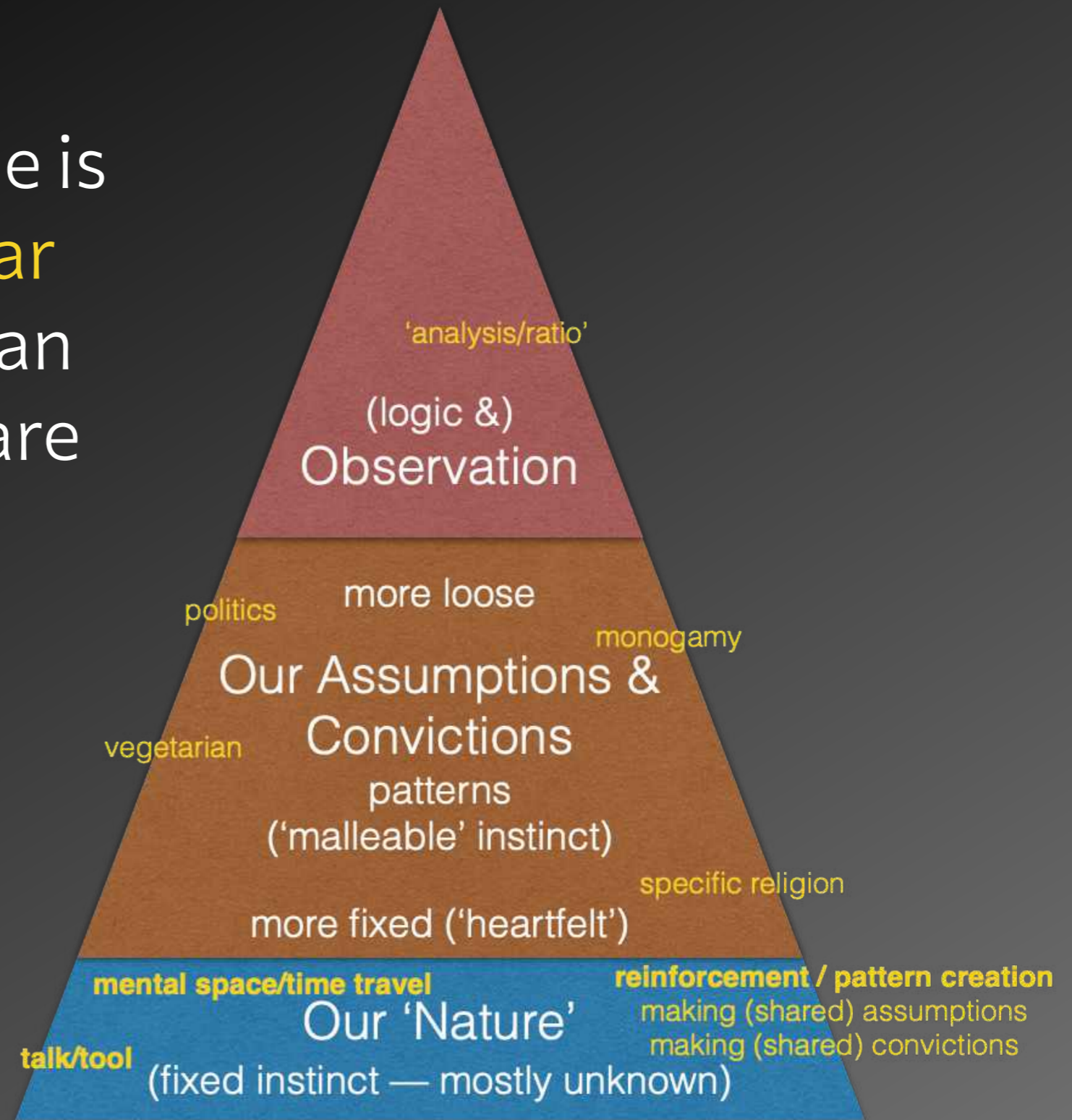
Hello Human Intelligence, meet
Complexity Crunch

Part 0-3: Technology and Psychology

<https://youtu.be/3riSN5TCuoE>



Menselijke intelligentie is
(vooral) **veranderbaar**
instinct, in de vorm van
veranderbare hardware



<https://ea.rna.nl/2022/10/24/on-the-psychology-of-architecture-and-the-architecture-of-psychology/>



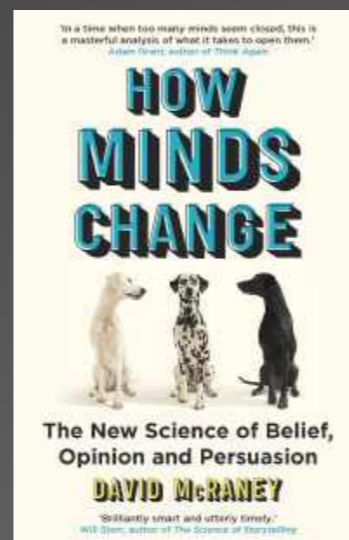
Wij nemen aan...

“Onze overtuigingen komen vooral tot stand **op basis van observatie en redeneren**”



...maar het is veel vaker

“Onze observaties en redeneringen komen tot stand
op basis van onze overtuigingen”





Onze overtuigingen ontstaan uit:
leren – **persoonlijke ervaring**
kopiëren – **herhaling** en **relatie** zijn sterke
veroorzakers (zie 'social media', 'influencers')



Van 'zintuig' naar daad?

Het gaat bij **menselijke intelligentie** niet alleen om het reageren op waarnemingen (alles met zintuigen kan dat)



Van 'droom' naar daad!

Het gaat bij menselijke intelligentie niet alleen om het reageren op waarnemingen (alles met zintuigen kan dat)

Menselijke intelligentie is gebouwd rond een sterk, maar ingeperkt, **voorstellingsvermogen**

Wij kunnen denken aan zaken **die er (nog) niet zijn**, wij kunnen denken in *mogelijkheden*, zowel – positief – **kansen** en – negatief – **risico**



Het kan voorkomen dat (sterke) **overtuigingen** vooral gebaseerd zijn op onze 'fantasie'

Niet alleen samenzweringstheorieën, maar ook elk geloof in een wondermiddel – zoals GenAI



Samengevat:

Wij mensen maken vooral 'snel en slordig'
inschattingen



Zoals nu

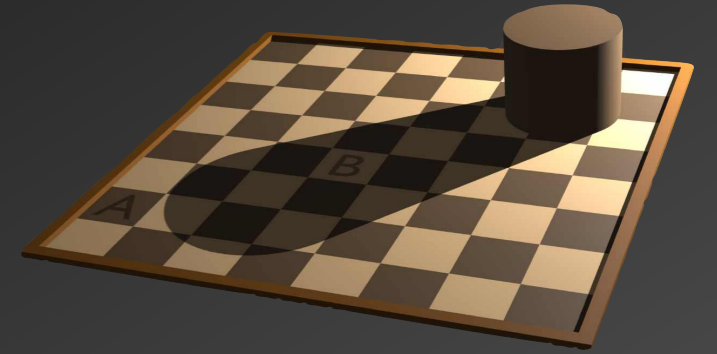


Alle grotere structuren die wij maken (organisaties, samenlevingen) steunen uiteindelijk op ‘gedeelde overtuigingen’



https://youtu.be/9_Rk-DZCVKE





Wij mensen accepteren wat we 'werkelijkheid' vinden
voornamelijk op de automatische piloot

Dat geldt ook voor informatie die wij ontvangen. Ons brein wordt
gedomineerd door 'automatisering'

We zijn kwetsbaar

Wij interpreterten. Die interpretatie hangt vooral af
van onze bestaande overtuigingen

We zijn – net onder het oppervlak – vrij inflexibel



Een Herhaling Van Zetten?



Stel Je Voor...

Geen volledig autonome auto's (vereist echte intelligentie)

Maar op de snelweg rijden gaat 'goed genoeg'

Vrachtwagens worden bi-modaal: autonoom/
chauffeur

Lange afstand: autonoom, met 1 persoon voor elke
100 vrachtwagens – een laagbetaalde baan

Vanaf bepaalde verzamelpunten: chauffeurs voor de
laatste kilometers



Stel Je Voor...

Geen grafische vormgever voor je bedrijfsplaatjes,
maar GenAI

Veel minder grafische vormgevers, maar grafisch
ongeschoolde medewerkers die moeten controleren
of GenAI er niet een zootje van maakt – werk dat
minder vaardigheden vereist



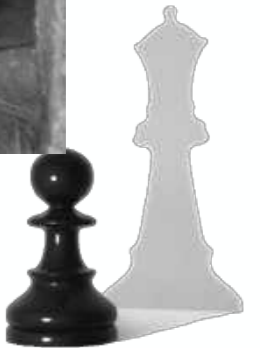
Niets Nieuws Onder De Zon

Twee soorten innovaties die je niet door elkaar moet halen:

Nieuw: iets mogelijk maken dat eerder niet kon (b.v. 'targeting', AlphaFold)

Vervanging van iets door een meer 'winstgevende' ('goedkope') oplossing – b.v. GenAI illustraties versus menselijke 'digital artists', zelfsturende vrachtwagens in plaats van vrachtwagenchauffeurs





Als we Technologie en
Psychologie bij elkaar gooien, wat
voor beleid hebben we dan nodig?



```
cat AI-act.eu \  
| sed 's/AI/IT/g' \  
| sed 's/artificial intelligence/machine logic/g' \  
> Protect-Society-Against-IT-act.eu
```



Beleid dat rekening houdt met de **fundamentele eigenschappen en vooral beperkingen van menselijke intelligentie**

(we zijn b.v. makkelijk voor de gek te houden)

Ik twijfel sterk of maatregelen zoals ‘watermerken’ krachtig genoeg zijn, omdat ze onder water uitgaan van een onrealistisch beeld van de menselijke intelligentie – zelfs als het **technisch** lukt, trekken we ons er **mentaal** wat van aan?



Allemaal enorm lastig en risicovol, maar b.v.:

Nieuwe beperkingen aan vrijheid van meningsuiting (b.v. hoe groter de reikwijdte – zoals ‘views’, ‘volgers’, ‘gebruikers’ – des te strenger de eisen aan ‘betrouwbaarheid’)

Criminaliseer systemen die onbetrouwbaarheid versterken of betrouwbaarheid verzwakken

Een ‘ethische gedragscode’ (a la medici) voor werkers in de IT

Een vergunningensysteem voor systemen die zich voordoen als mens

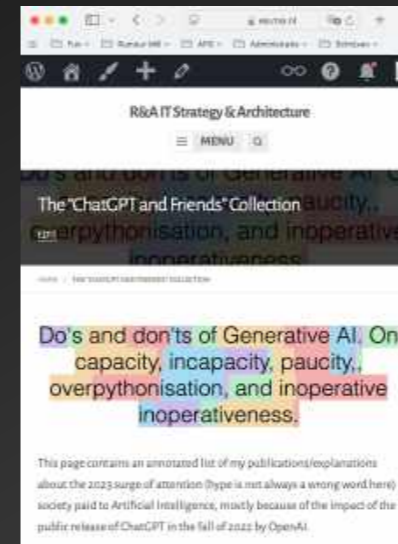
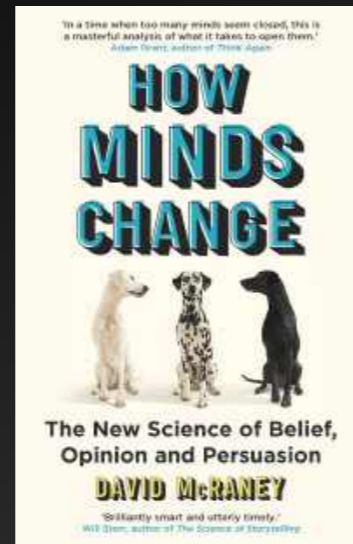
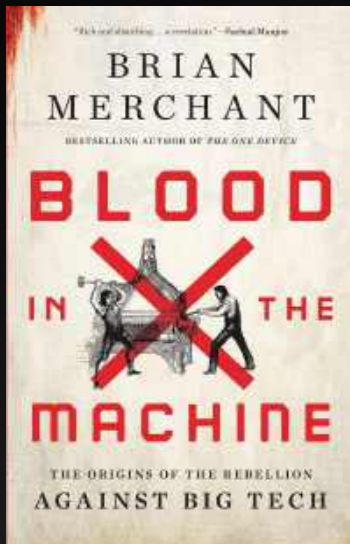


Moeten we alleen 'natuurlijke personen' en 'de mensheid' beschermen of is er daartussen ook zoiets als 'misdad tegen de samenleving (=gedeelde overtuigingen)'?

Denk aan het aanpakken van 'information warfare' die gericht is op het ondermijnen van 'samen' in 'samenleving'

Waar ligt het evenwicht tussen onze kwetsbaarheid als mens (en dus samenleving) en wel kunnen/mogen veranderen?





Jullie beurt

<https://ea.rna.nl/the-chatgpt-and-friends-collection/>
https://www.youtube.com/channel/UCTqv-dxJc_nMsZqBYJW3iCw

