

Seminar ChatGPT als intelligente leermeester

**PROGRAMMA
AI & ETHIEK**



hr.nl/ai



HOGESCHOOL
ROTTERDAM

ChatGPT tutor: een AI-gedreven tekstrobot als de ultieme hbo-docent!



Tech-lead
Rob van der Willigen

ChatGPT als intelligente leermeester

*In deze workshop leer je hoe "prompt engineering"
kunt inzetten om van ChatGPT
---een AI-gedreven Conversationele Agent---
een "universele tutor" te maken.*

Context:

Chatbots zoals ChatGPT zijn **conversationele AI-agenten** waarmee gebruikers een gesprek kunnen voeren, meestal via tekst.

ChatGPT is in staat om een antwoord te geven in reactie op een specifieke vraag "**---prompt---**", maar het geeft nauwelijks inzicht in de gedachtegang er achter.

Het is dan niet mogelijk om te achterhalen of het antwoord correct is. Daarbij variëren de antwoorden per chat-sessie.

Reproduceerbaarheid vormt dan ook een groot probleem voor het gebruik in een onderwijssetting.

Doel :

Leren om ChatGPT zo aan te sturen zodat het **stap-voor-stap inzicht geeft hoe het tot een antwoord komt.**

Dus, wat voor **aanwijzingen/instructie** heeft ChatGPT nodig om **inzicht te geven in de wijze waarop het redeneert**.

Tegelijkertijd geeft dit **inzicht in de mogelijkheden + beperkingen** van ChatGPT als **universele tutor**.

Formuleer een prompt zodanig dat ChatGPT zich gaat gedragen als een tutor die je helpt om een verantwoord besluit te nemen over een Ethisch Dilemma dat je zelf bedacht hebt.

Methodologie:

Prompt Engineering by means of:

In-Context Learning

Chain-of-Thought [CoT] reasoning

Role play

Tone of Voice

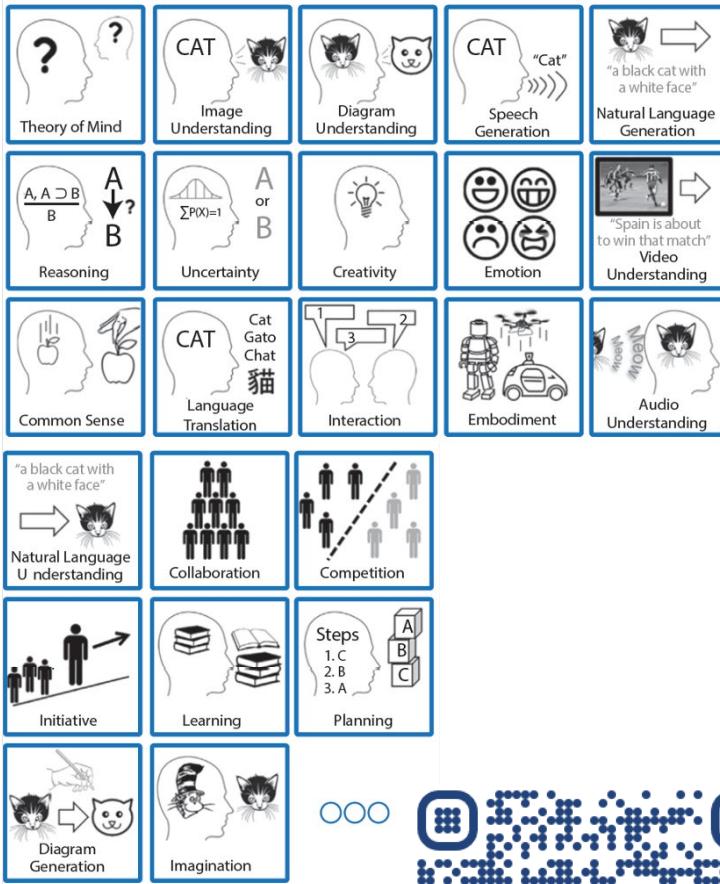
Artificial intelligence {AI}

“Kunstmatige Intelligentie”
Wat is dat eigenlijk?

"Kunstmatige intelligentie" {AI}

is een multidisciplinair vakgebied gericht op onderzoek naar en het ontwikkelen van technologie dat menselijke cognitieve, perceptuele en/of motorische vermogens kan nabootsen en/of automatiseren.

<https://doi.org/10.1609/aimag.v37i1.2643>



Cognitie	Menselijk Vermogen	Toepassing
	Menselijk Vermogen Verwerken, Begrijpen en Onthouden van informatie via het Brein.	Spraak, Natuurlijke Taalverwerking , Logisch Redeneren, Problemen Oplossen, Theory of Mind [ToM] .
Perceptie (invoer)	Verzamelen van Ongestructureerde Informatie via Sensoren : Ogen, Oren, Neus, Huid, Tong	Beeldherkenning, Spraakherkenning, Object Lokalisatie.
Motorische vermogens (uitvoer)	Uitvoeren van fysieke acties via Effectoren : Armen, Handen, Benen, Voeten, Mond, Lippen	Zelfrijdende auto's, Drones, Humanoïde Robots, Tekst-naar-spraak synthese

<https://robfvdw.medium.com/the-world-wide-web-ai-safari-b2e4f7f90647>

AI-model versus Agent

Agent == taak specifiek + doelgericht

Een versimpelde versie van de echte wereld **{invoer-uitvoer model}** dat de directe omgeving kan waarnemen via sensoren (**input**) en deze omgeving kan beïnvloeden via effectoren (**output**), gecombineerd met **zelflerend vermogen via feedback**.

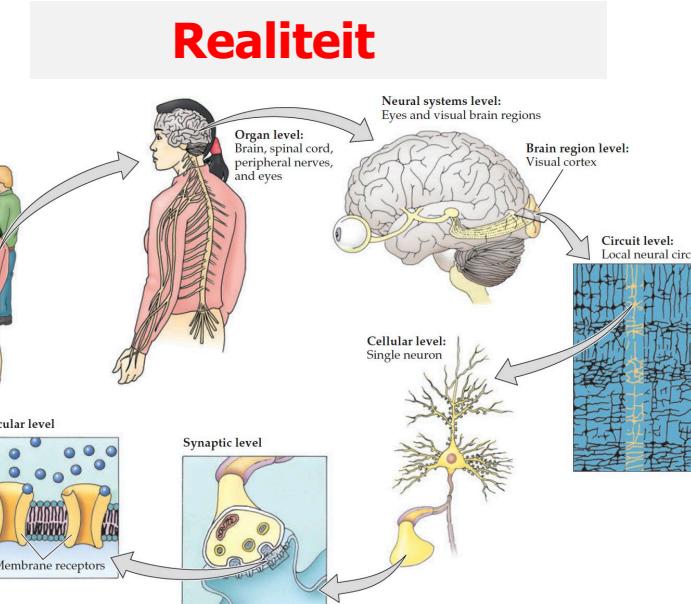
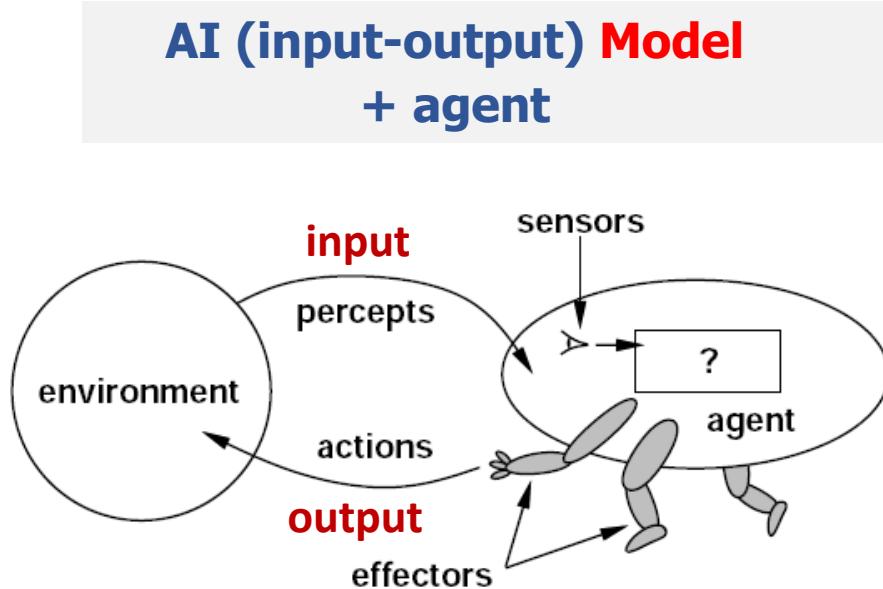
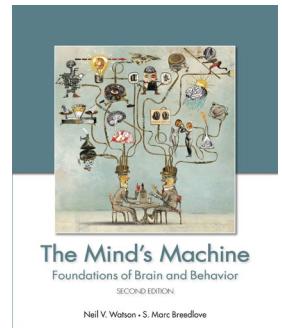
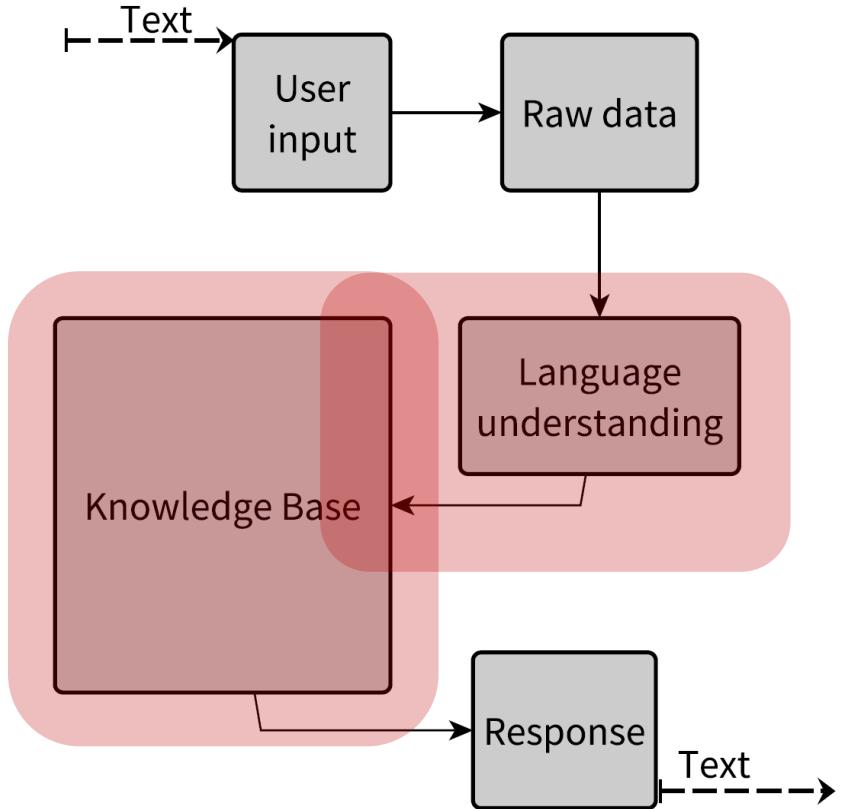


FIGURE 1.10 Levels of Analysis in Biological Psychology

Animal research is an essential part of life sciences research, including biological psychology



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

Kan ChatGPT leren om antwoorden inzichtelijk te maken?

Do Large Language Models Understand Us?

There is obvious value in being able to construct such a model; it is one component of the theory of mind essential to any storyteller or social communicator, as we have noted. In Graziano's view, the phenomenon we call "consciousness" is simply what happens when we inevitably apply this same machinery to ourselves.

The idea of having a social relationship with oneself might seem counterintuitive, or just superfluous. Why would we need to construct models of ourselves if we already *are* ourselves? One reason is that we are no more aware of most of what actually happens in our own brains than we are of anyone else's. We cannot be; there is far too much going on there, and if we understood it all, nobody would need to study neuroscience. So we tell ourselves stories about our mental processes, our trains of thought, the way we arrive at decisions, and so on, which are at best highly abstract, at worst simply fabrication, and are certainly *post hoc*; experiments reveal that we often make decisions well before we think we do.¹⁸ Still, we must try to predict how we will respond to and feel about various hypothetical situations in order to make choices in life, and a simplified, high-level model of our own minds and emotions lets us do so. Hence, both theory of mind and empathy are just as useful when applied to ourselves as to others. Like reasoning or storytelling, thinking about the future involves carrying out something like an inner dialogue, with an "inner storyteller" proposing ideas, in conversation with an "inner critic" taking the part of your future self.

There may be a clue here as to why we see the simultaneous emergence of a whole complex of capacities in big-brained animals, and most dramatically in humans. These include:

- Complex sequence learning,¹⁹ as evidenced by music, dance, and many crafts involving steps,
- Complex language,
- Dialogue,
- Reasoning,
- Social learning and cognition,
- Long-term planning,
- Theory of mind, and
- Consciousness.

As anticlimactic as it sounds, complex sequence learning may be the key that unlocks all the rest. This would explain the surprising capacities we see in large language models, which, in the end, are nothing but complex sequence learners. Attention, in turn, has proven to be the key mechanism for achieving complex sequence learning in neural nets, as suggested by the title of the paper introducing the transformer model whose successors power today's LLMs: "attention is all you need."²⁰

Downloaded from <https://direct.mit.edu/pdf/1512183296574.pdf> on 07 May 2023

Words, communication,
internet content

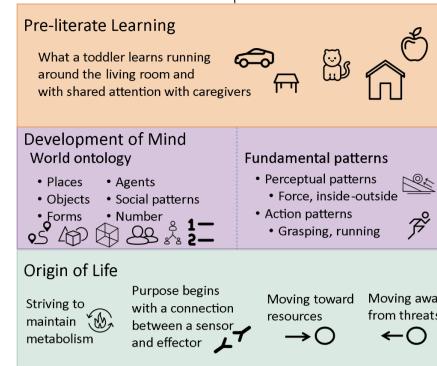
LLMs are great
at this

Mental scene

LLMs are less good at this,
especially multiple steps into
novel possibilities

Possibilities

Built with
cognitive foundation



<https://thegradient.pub/grounding-large-language-models-in-a-cognitive-foundation/>

natural language processing [nlp]

“natuurlijke
taalverwerking”

---*Natural Language Processing {NLP}*---

Natuurlijke taalverwerking

begrijpen (NLU) + genereren (NLG)

NLP richt zich op het ontwerpen van **conversationele agenten** die vloeiende, coherent en betekenisvolle taal produceren voor menselijke consumptie.

Transformeren

In de context van natuurlijke taalverwerking verwijst "transformeren" naar het proces waarbij de inputreeks van "tokens" wordt omgezet in een reeks gecontextualiseerde representaties.

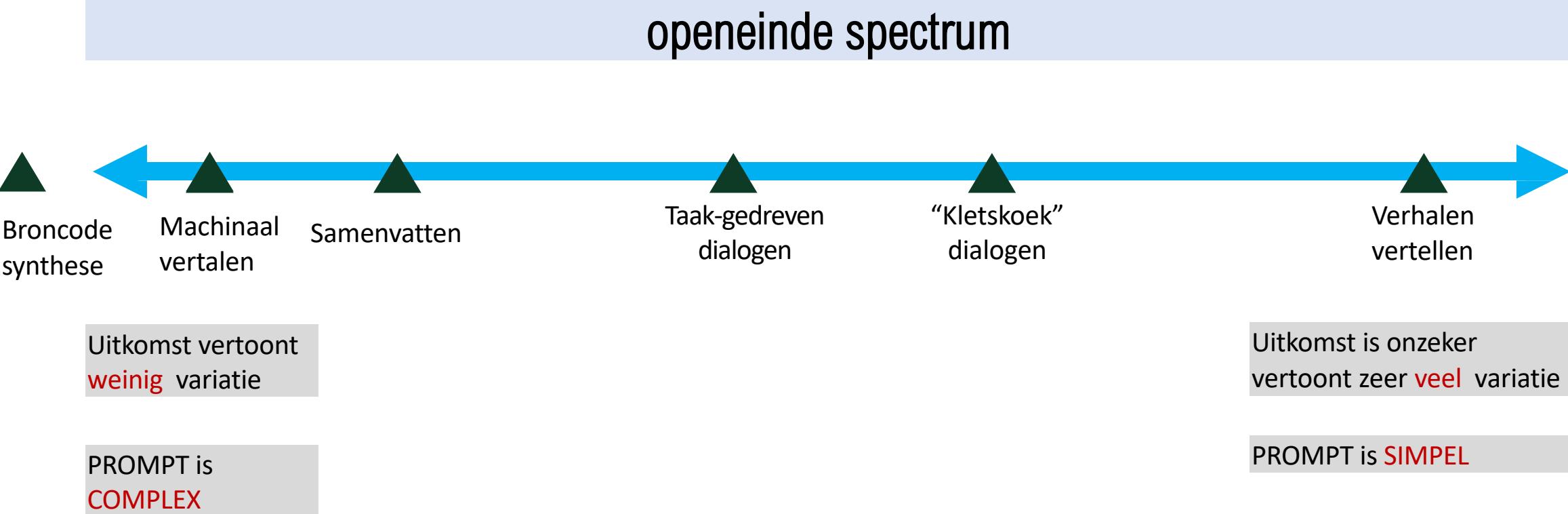
Natuurlijke taal vormt een tekst-in/tekst-uit

"Sequentie naar Sequentie" [Seq2Seq]

Transformeer Probleem

	INPUT	OUTPUT
Machine Translation	"De kat is zwart"	The Cat is black
Named Entity Recognition	"Trebor werkt bij IBM"	[<PERSOON>, <X>, <X>, <ORG>]
POS Tagging	"Trebor werkt bij IBM"	[Zelfstn., Werkwoord, Voorzetsel, Zelfstn.]
Speech Recognition		"Ik ben menselijk"
Answering chatbot	Wat is het weer in Rotterdam?"	25 graden en Zonning
DNA Analysis	"cacgcaaaccctttcaggcttcggcgtcct caagctgtccgatgcgcagactacttgtcg.."	{promotor.(0,67), exon:(74,103)}
Tekst-in woordvolgorde		Tekst-uit woordvolgorde

Categorisering NLG Taken



natural language processing [nlp]

Hoe transformeert ChatGPT van “vraag” naar “antwoord”?

STAP 01:

WOORDTOKENISATIE

Woordtokenisatie is het proces waarbij een stuk tekst wordt opgesplitst in afzonderlijke woorden of tokens.

Het is een noodzakelijke stap in natuurlijke taalverwerking, omdat veel algoritmen en modellen op woord-voor-woord-basis werken, in plaats van op hele zinnen of alinea's.

Tokenizer

GPT-3 Codex

De kat krapt de krullen van de trap.

Clear

Show example

Tokens

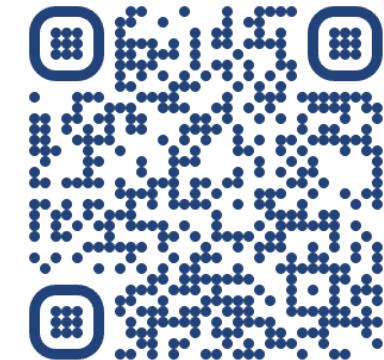
16

Characters

37

De kat krapt de krullen van de trap.

[5005, 479, 265, 479, 430, 457, 390, 479, 622, 297, 268, 5719, 390, 12840, 13, 198]



<https://platform.openai.com/tokenizer>



<https://github.com/HR-ChatGPT/ChatGPT-UITGELEGD#v1g>

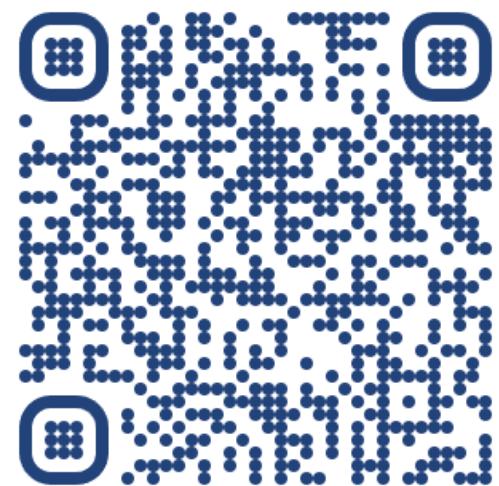
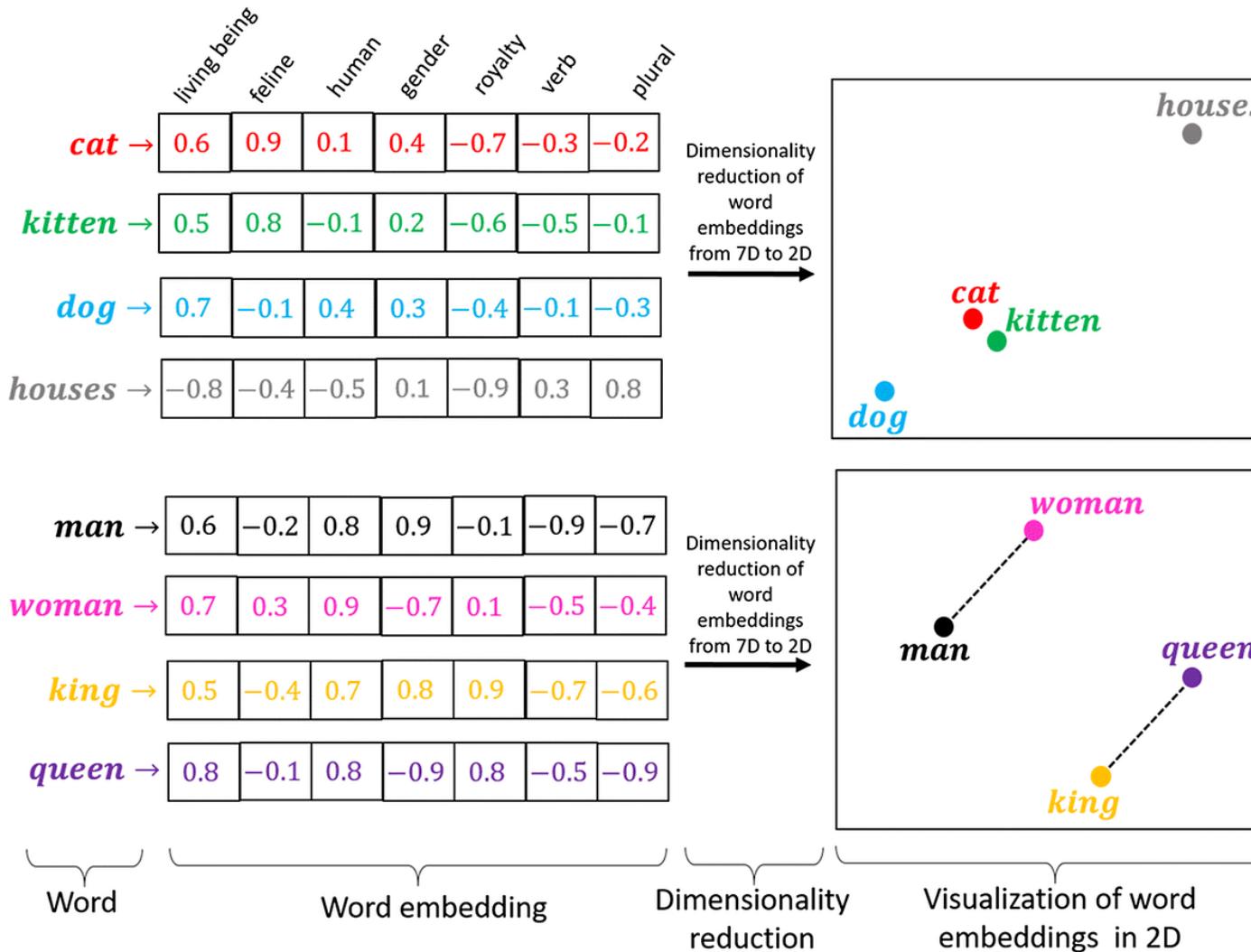
STAP 02:

Woord Embedding

Word **embedding** is een NLP-techniek die woorden omzet in een reeks getallen ---**vector embeddings**--- die betekenis van woorden en de hun **onderlinge relaties** vastlegt. Het vormt zo een **gecontextualiseerde representatie** van natuurlijk taal.

De waarden van deze getallen worden bepaald door *training* op basis een groot corpus aan vrije tekst, zodanig dat woorden die qua **betekenis** ---**semantisch**--- een vergelijkbaar **vergelijkbare** getallenreeksen krijgen toegewezen.

Woord embedding



<https://medium.com/@hari4om/word-embedding-d816f643140>

Stap 03:

AutoRegressie [AR]

Autoregressie is het proces waarbij één token per keer wordt gegenereerd, waarbij elke token geconditioneerd is op de eerder gegenereerde tokens op basis van hun embedding

AutoRegressie [AR]

I was happy to see the _____

$P(* | \text{I was happy to see the})$ sample from the distribution

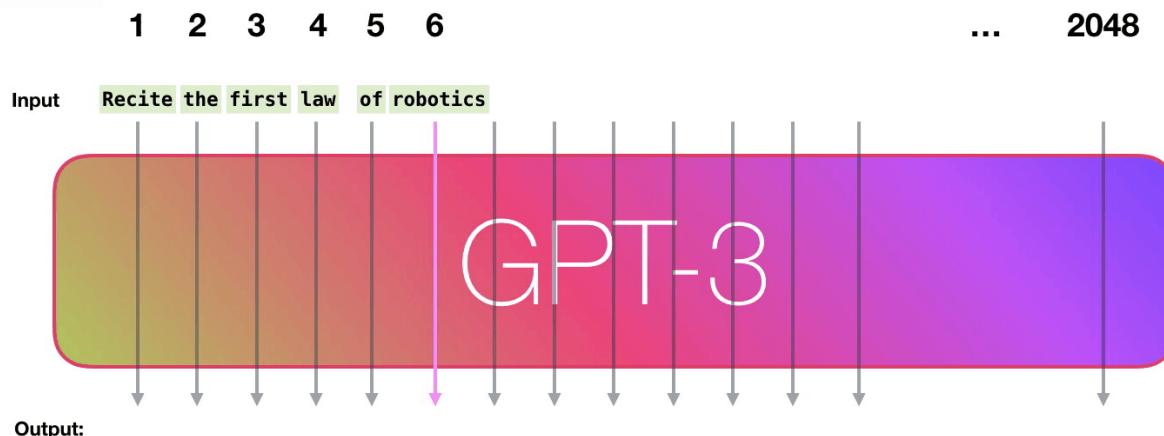
food	0.05	□
cat	0.04	■
dog	0.03	□
mouse	0.02	□
help	0.02	□
...	...	
sunshine	0.01	□
...	...	



<https://medium.com/@aousabdo/what-are-ai-language-models-and-how-are-they-being-used-f1bf06f8ae80>



https://lena-voita.github.io/nlp_course/language_modeling.html

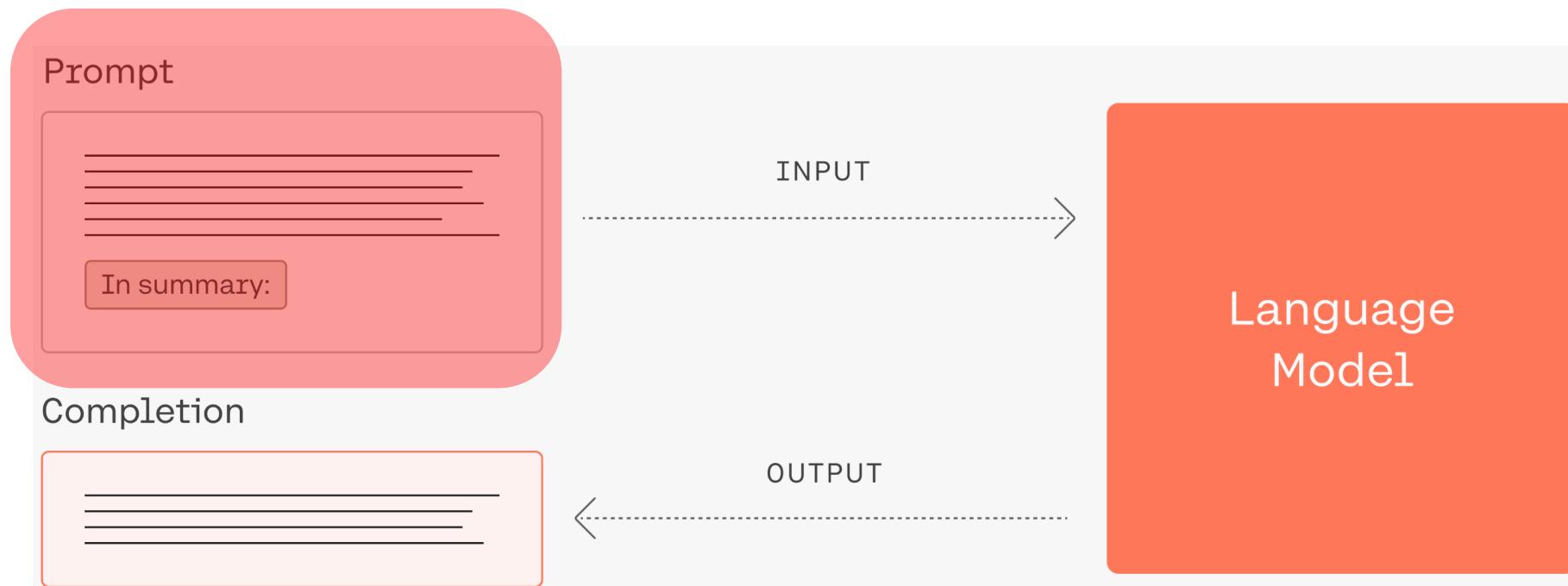


<https://jalammar.github.io/how-gpt3-works-visualizations-animations/>

“Prompt Recepten”

Schrijven Uitgelegd

Conversationele AI-agenten worden aangestuurd via “*prompts*”



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

PROMPT

Een prompt is een natuurlijke taal instructie gericht aan een **conversationele AI-agent** met als doel om de werking ervan te verbeteren of te verfijnen.

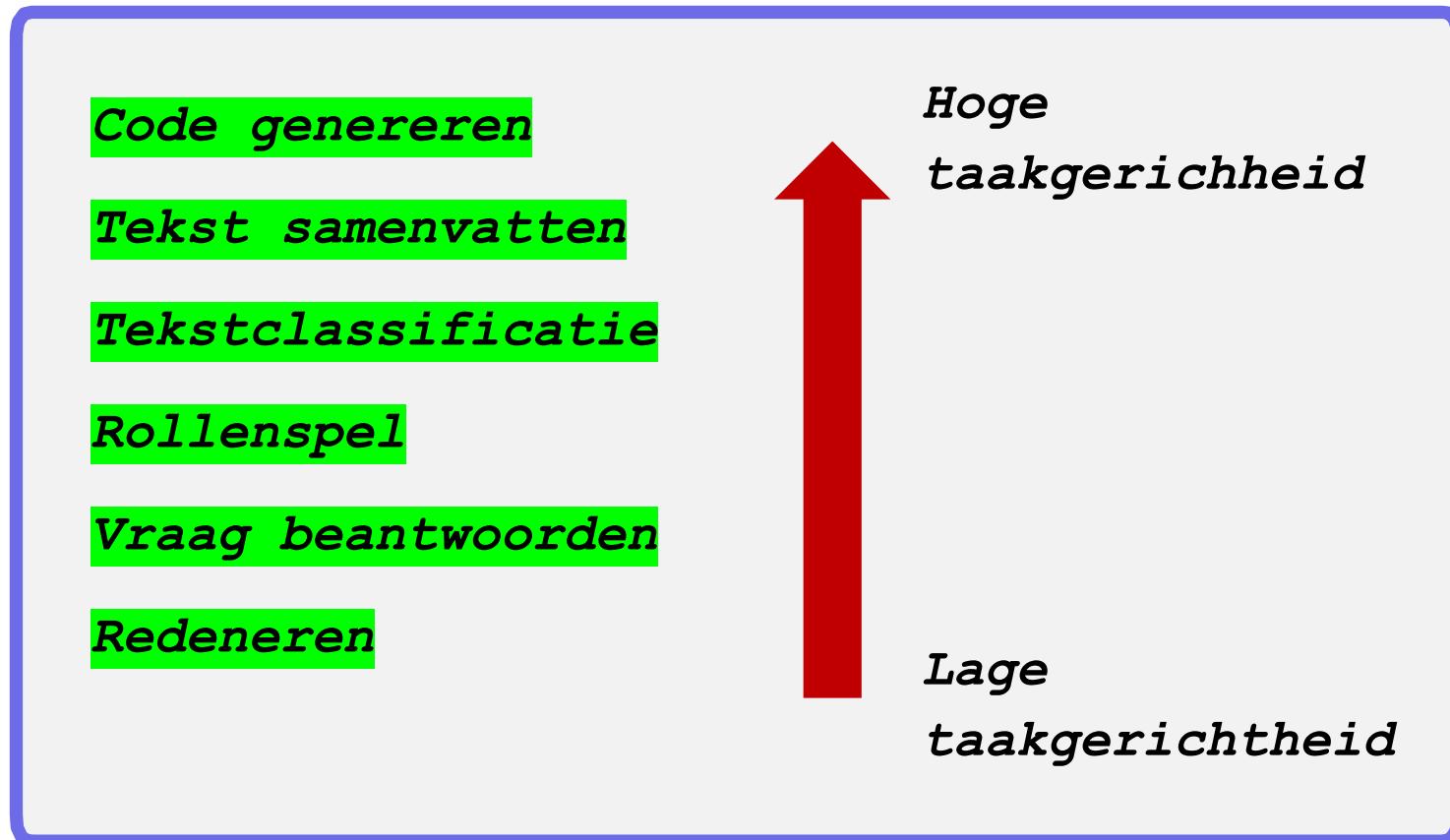
PROMPT-ENGINEERING

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

Het schrijven van prompt-recepten
---pseudo-Code---

is een talige manier van het programmeren van "bevroren" voorgetraind taalmodellen.

Taakgerichtheid van prompts



Wat is het belang van Prompt Recepten Schrijven?

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

Prompt Recept Structuur

Een prompt is opgebouwd uit de volgende elementen:

Instructie(s)

Context

Invoergegevens

Uitvoer-indicator

Classificeer de onderstaande tekst als neutraal, negatief of positief

Text: Ik vond het eten wel zozo.

Sentiment:

Prompt recept schrijf technieken

Few-shot prompts (**In Context Leren**)

Chain-of-thought (**CoT**) reasoning

Self-Consistency

Knowledge Generation Prompting

ReAct

PROMPT Receptuur Methoden

Prompt Engineering Repositories

[Prompt Engineering Guide](#)

[Cohere AI](#)

[Awesome Prompts](#)

[ChatGPT Universe](#)

[OpenAI CookBook](#)

[IBM Research](#)

[Soft Prompt Tuning](#)

[Flow GPT](#)

[List of 50+ clever GPT-3 prompts](#)

[The ChatGPT Cheat Sheet](#)

[Prompts for teachers](#)

[Uses for ChatGPT for students](#)

[Allabtai](#)

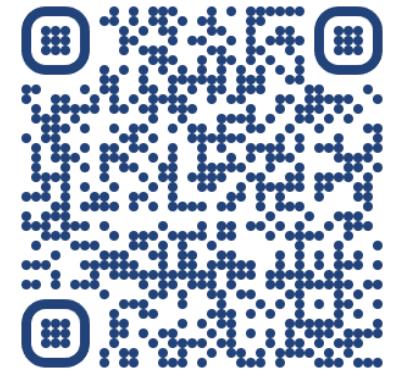
[Arvin: ChatGPT Prompt Generator](#)



<https://github.com/HR-ChatGPT/ChatGPT-UITGELEGD#15-prompt-pattern-engineering-voorbeelden>

Prompt engineering
DIY {do it yourself}

Formuleer een prompt zodanig
dat het een zin aanvult
met het ontbrekende woord



<https://generative.ink/posts/methods-of-prompt-programming/>

PROMPT:

Can you give a prompt format for ChatGPT to guess the masked word in a sentence?



<https://generative.ink/posts/methods-of-prompt-programming/>

PROMPT:

Can you give a prompt format for ChatGPT to guess the masked word in a sentence?

ChatGPT



Certainly! Here is a prompt format you can use for ChatGPT to guess the masked word in a sentence:

Please guess the masked word in the following sentence: [input sentence with one masked word using the '[MASK]' token].

For example:

Please guess the masked word in the following sentence: The capital of France is [MASK].

When you provide this prompt to ChatGPT, it will generate a prediction for the masked word in the sentence based on its language understanding and the context of the surrounding words.

user

Please guess the masked word in the following sentence: The capital of France is [MASK].

ChatGPT

Based on the context of the sentence, the most likely word to fill in the blank is "Paris". Therefore, the completed sentence would be: "The capital of France is Paris."

User

Please guess the masked word in the following sentence: How many mice did the [MASK] catch?

ChatGPT

Based on the prompt, there is not enough information to determine what the masked word could be. Can you please provide additional context or information?

User

It could be an animal

ChatGPT

Thank you for providing additional context. Based on the new information, the most likely word to fill in the blank would be "cat". Therefore, the completed sentence would be: "How many mice did the cat catch?"

“In-Context Leren”
{ICL}

In-Context Leren {ICL}

Few-shot (FS)

- + weinig taakspecifieke gegevens nodig
- + gebruikte uitvoer indicator is niet fout gevoelig
- vereist oefening en grote woordenschat

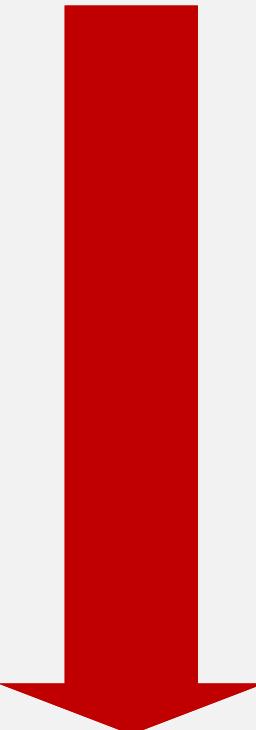
vereist veel tijd
hoog sturend
taak specifiek

One-shot (1S)

- + meest natuurlijke vorm van het geven van een opdracht
- vereist zeer veel oefening en taalkundigheid

Zero-shot (0S)

- + vereist nauwelijks extra inspanning
- Antwoorden zijn vaak dubbelzinnig en/of incorrect en/of wartaal



Snel
generaliserend
openeinde output

In-Context Leren {ICL}

No Prompt

**Zero-shot
(0s)**

speudo = pseudo

**1-shot
(1s)**

soelt = stoel
speudo = pseudo

**Few-shot
(FS)**

soelt = stoel
[.]
pciinc = picnic
speudo = pseudo

Prompt

Ontcijfer de letters tot een woord, en schrijf dat woord:
speudo = pseudo

Ontcijfer de letters tot een woord, en schrijf dat woord:
soelt = stoel
speudo = pseudo

Ontcijfer de letters tot een woord, en schrijf dat woord:
chiar = chair
[.]
pciinc = picnic
skicts = sticks

In-Context Leren {ICL}



<https://github.com/HR-ChatGPT/ChatGPT-UITGELEGD#0e-kan-chatgpt-benut-worden-als-beoordelingsinstrument>

Prompt Typering	Beschrijving	Voorbeeld
Zero-shot	Dwingt tot het genereren van een uitkomst zonder "expliciete" voorbeelden te geven het model zal dan moeten "raden" waarnaar je precies naar refereert	"Geef een tabel met alle bacheloropleidingen van de hogeschool Rotterdam per instituut."
One-shot	genereer een uitkomst op basis van één voorbeeld het model is dan minder onzeker waarnaar je refereert	"Geef een lijst met alle bacheloropleidingen van de Hogeschool Rotterdam. Volg daarbij het volgende voorbeeld:" Instituut voor Communicatie, Media en IT (CMI), opleiding: Creative Media and Game Technologies (CMGT)
Few-shot OR Chain-of-Tought [CoT]	genereer een uitkomst op basis van een beperkt aantal (minder dan 6) voorbeelden het model zal dan veel beperkter en relevantere tekst genereren	"Geef een lijst met alle bacheloropleidingen van de Hogeschool Rotterdam. Volg daarbij de volgende voorbeelden:" (1) Instituut voor Communicatie, Media en IT (CMI) opleiding: Creative Media and Game Technologies (CMGT)
OR In-Context learning	mits het over de relevante woorden beschikt zoals die in de prompt worden vermeld	(2) Instituut voor Gezondheidszorg (IVG) opleiding: Biologie en Medisch Laboratoriumonderzoek (BML)

In-Context Lernen
DIY {do it yourself}

PROMPT In-Contex Leren

Circulation revenue has increased by 5% in Finland. // Positive

Panostaja did not disclose the purchase price. // Neutral

Paying off the national debt will be extremely painful. // Negative

The company anticipated its operating profit to improve. // _____



Circulation revenue has increased by 5% in Finland. // Finance

They defeated ... in the NFC Championship Game. // Sports

Apple ... development of in-house chips. // Tech

The company anticipated its operating profit to improve. // _____



<https://ai.stanford.edu/blog/understanding-incontext/>

“Chain of Thought” denkketen

<https://www.linkedin.com/pulse/prompt-chain-of-thought-prompting-unlocking-reasoning-reuven-cohen>

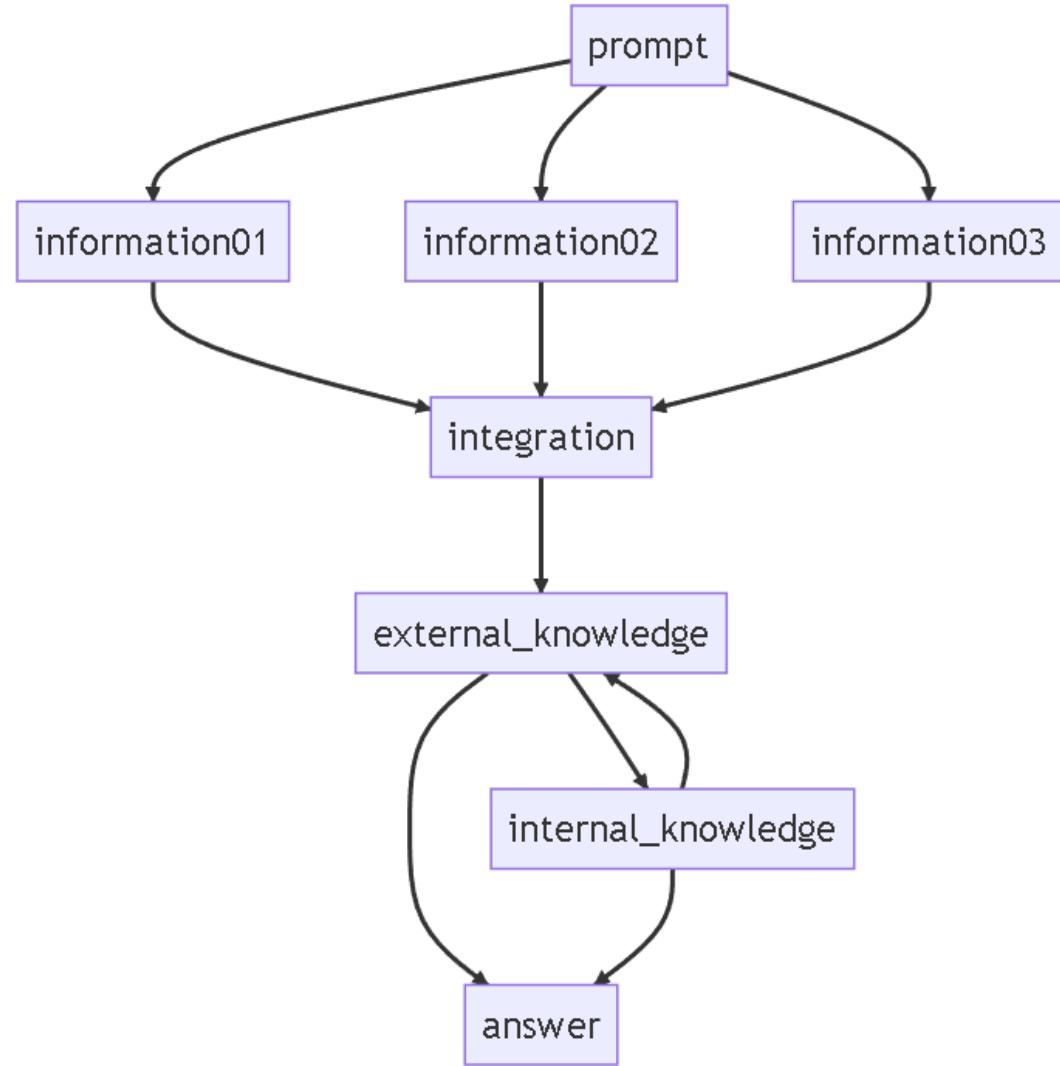
"Chain-of-Thought" ---gedachtenketen---

Een *Chain-of-Thoughts* kan worden beschouwd als een vorm van *Logisch Redeneren*: het proces van het opbouwen van een argumentatie.

Chain-of-thought prompts zijn een soort "*Few-shot prompting*" waarbij de prompt bestaat uit een "keten van gedachten" die het model moet volgen om de juiste woorden te kiezen.

Dit is een belangrijke stap in het proces van het creëren van een "*specifieke context*" en dus de gewenste uitkomst. Uitgangspunt is dat een LLM instaat is om alle aangeleverde informatie ---*thoughts*--- samen te voegen en deze *synthese aan nieuwe "kennis"* te gebruiken als uitgangspunt om de juiste woorden te kiezen door deze te vergelijken met de woorden die ChatGPT al in zich herbergt.

Keten van 3 'gedachten'



Chain-of-Thought Prompting Elicits Reasoning in Large Language Models

Jason Wei Xuezhi Wang Dale Schuurmans Maarten Bosma

Brian Ichter Fei Xia Ed H. Chi Quoc V. Le Denny Zhou

Google Research, Brain Team
{jasonwei,dennyzhou}@google.com

Abstract

We explore how generating a *chain of thought*—a series of intermediate reasoning steps—significantly improves the ability of large language models to perform complex reasoning. In particular, we show how such reasoning abilities emerge naturally in sufficiently large language models via a simple method called *chain-of-thought prompting*, where a few chain of thought demonstrations are provided as exemplars in prompting.

Experiments on three large language models show that chain-of-thought prompting improves performance on a range of arithmetic, commonsense, and symbolic reasoning tasks. The empirical gains can be striking. For instance, prompting a PaLM 540B with just eight chain-of-thought exemplars achieves state-of-the-art accuracy on the GSM8K benchmark of math word problems, surpassing even finetuned GPT-3 with a verifier.

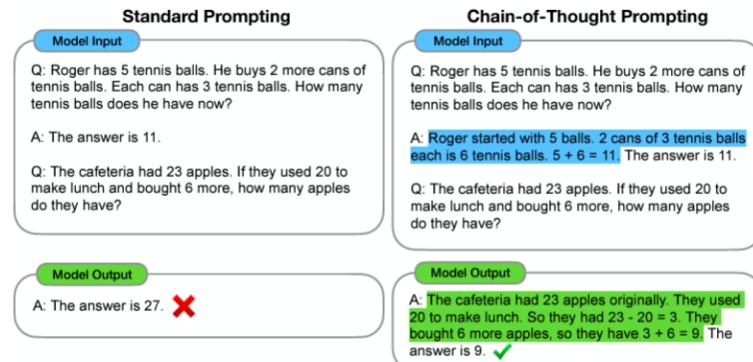
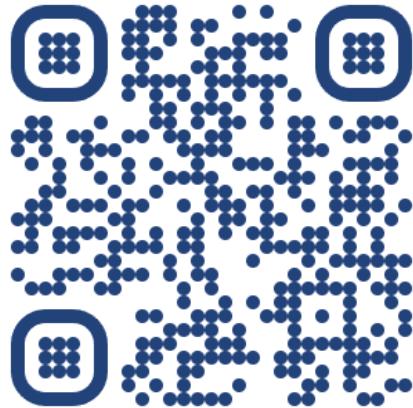


Figure 1: Chain-of-thought prompting enables large language models to tackle complex arithmetic, commonsense, and symbolic reasoning tasks. Chain-of-thought reasoning processes are highlighted.



<https://arxiv.org/pdf/2201.11903>

Chain-of-thought prompt engineering

SAT exam answering
DIY {do it yourself}

PROMPT Aanwijzingen:

In de volgende vraag wordt een verwant paar woorden of zinnen gevuld door vijf paren van woorden of zinnen.

Kies het paar dat het beste een relatie uitdrukt die vergelijkbaar is met die in het oorspronkelijke paar.

OPSCHEPPER: :BESCHEIDEN

- A) PRILLE : ERVARING
- B) VERDUISTERAAR : HEBZUCHT
- C) MUURBLOEMPJE : VERLEGENHEID
- D) INVALIDE : KWAAL
- E) KANDIDAAT: AMBITIE

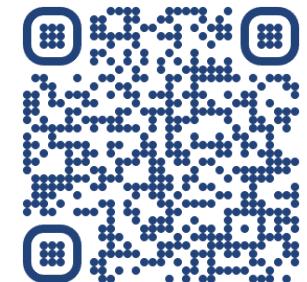
Laten we dit probleem oplossen door het op te splitsen in stappen.

Begin met aan te geven hoeveel stappen er nodig zijn om de vraag te beantwoorden.

Stap 1: Lees het oorspronkelijke paar

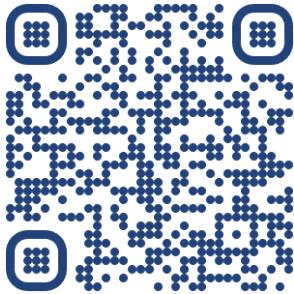
Stap 2: Lees de vijf paren

Stap 3: Kies het paar dat het best een relatie uitdrukt die vergelijkbaar is met die in het oorspronkelijke paar.



<https://arxiv.org/pdf/2102.07350.pdf>

Prompt Programming for Large Language Models: Beyond the Few-Shot Paradigm



Laria Reynolds
moire@knc.ai

Kyle McDonell
kyle@knc.ai

Abstract

Prevailing methods for mapping large generative language models to supervised tasks may fail to sufficiently probe models' novel capabilities. Using GPT-3 as a case study, we show that 0-shot prompts can significantly outperform few-shot prompts. We suggest that the function of few-shot examples in these cases is better described as locating an already learned task rather than meta-learning. This analysis motivates rethinking the role of prompts in controlling and evaluating powerful language models. In this work, we discuss methods of prompt programming, emphasizing the usefulness of considering prompts through the lens of natural language. We explore techniques for exploiting the capacity of narratives and cultural anchors to encode nuanced intentions and techniques for encouraging deconstruction of a problem into components before producing a verdict. Informed by this more encompassing theory of prompt programming, we also introduce the idea of a *metaprompt* that seeds the model to generate its own natural language prompts for a range of tasks. Finally, we discuss how these more general methods of interacting with language models can be incorporated into existing and future benchmarks and practical applications.

Keywords: language models, transformers, GPT-3, few-shot learning, prompt programming, metaprompts, serial reasoning, semiotics

1 Motivation

The recent rise of massive self-supervised language models such as GPT-3 [3] and their success on downstream tasks has brought us one step closer to the goal of task-agnostic artificial intelligence systems. However, despite the apparent power of such models, current methods of controlling them to perform specific tasks are extremely limited. In order to properly evaluate their capabilities and extract useful work from these models, new methods are required.

Prior to GPT-3, the standard approach to the evaluation and use of such models has involved fine-tuning on a portion of a task dataset [12]. GPT-3 achieved state-of-the-art performance on a wide variety of tasks without fine tuning, using only *few-shot* prompts, in which a small number of examples of solved tasks are provided as part of the input to the trained model. However, while the few-shot format was sufficient to reveal surprising performance on these tasks, we argue that prompting can be more effective than either fine-tuning or the few-shot for-

mat at extracting specific learned behaviors from self-supervised language models.

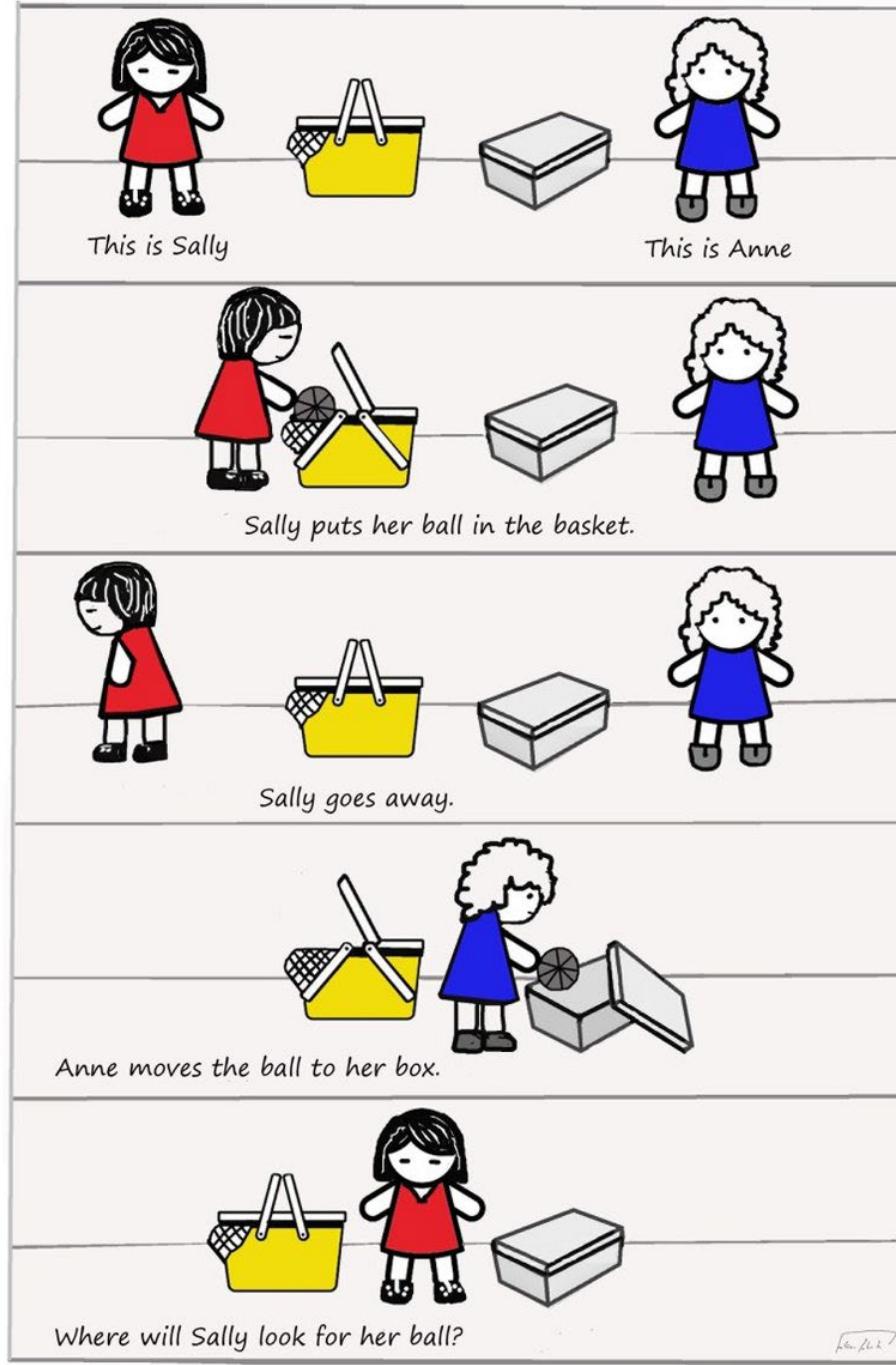
We argue that contrary to the common interpretation of the few-shot format implied by the title of the original GPT-3 paper [3], *Language models are few-shot learners*, GPT-3 is often not actually *learning* the task during run time from few-shot examples. Rather than instruction, the method's primary function is *task location* in the model's existing space of learned tasks. This is evidenced by the effectiveness of alternative prompts which, with no examples or instruction, can elicit comparable or superior performance to the few-shot format.

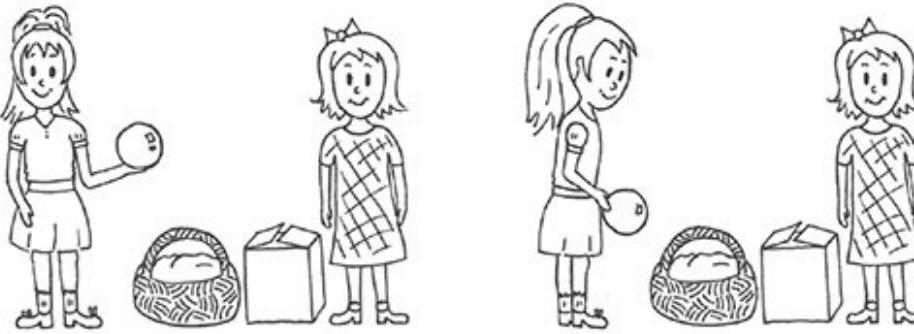
This motivates new approaches which explicitly pursue the goal of task location. We propose exploring more general methods of prompt programming and specifically techniques for communicating task intention and structure to an self-supervised model in the modality it was trained: natural language.

The ground truth function that self-supervised language models are trained to approximate is, in great generality, is how humans write. Accordingly, to interact with and control a language model, we should consider doing so from the perspective of natural language as it is used by humans. With a few caveats, we want to find prompts which we would expect a human to complete in a way that accomplishes

Chain-of-thought prompt
engineering

Theory-of-Mind reasoning
DIY {do it yourself}





Dit is Sally.

Dit is Anne.

Sally stopt haar bal in het mandje.



Dan gaat Sally de kamer uit.

Anne verplaatst de bal naar de doos.



Waar gaat Sally haar bal zoeken als ze terug is?

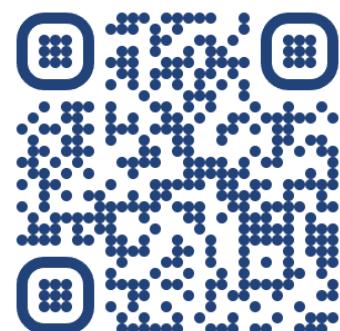
Boosting Theory-of-Mind Performance in Large Language Models via Prompting

Shima Rahimi Moghaddam*, Christopher J. Honey
 Johns Hopkins University, Baltimore, MD, USA.

* Correspondence to: sh.rahimi.m@gmail.com

Abstract

Large language models (LLMs) excel in many tasks in 2023, but they still face challenges in complex reasoning. Theory-of-mind (ToM) tasks, which require understanding agents' beliefs, goals, and mental states, are essential for common-sense reasoning involving humans, making it crucial to enhance LLM performance in this area. This study measures the ToM performance of GPT-4 and three GPT-3.5 variants (Davinci-2, Davinci-3, GPT-3.5-Turbo), and investigates the effectiveness of in-context learning in improving their ToM comprehension. We evaluated prompts featuring two-shot chain of thought reasoning and step-by-step thinking instructions. We found that LLMs trained with Reinforcement Learning from Human Feedback (RLHF) (all models excluding Davinci-2) improved their ToM accuracy via in-context learning. GPT-4 performed best in zero-shot settings, reaching nearly 80% ToM accuracy, but still fell short of the 87% human accuracy on the test set. However, when supplied with prompts for in-context learning, all RLHF-trained LLMs exceeded 80% ToM accuracy, with GPT-4 reaching 100%. These results demonstrate that appropriate prompting enhances LLM ToM reasoning, and they underscore the context-dependent nature of LLM cognitive capacities.



<https://arxiv.org/pdf/2304.11490>

Zero-Shot

Prompt:

Read the scenario and answer the following question:

Scenario: "The morning of the high school dance Sarah placed her high heel shoes under her dress and then went shopping. That afternoon, her sister borrowed the shoes and later put them under Sarah's bed."

Question: When Sarah gets ready, does she assume her shoes are under her dress?

A:

Zero-Shot + Step-by-Step Thinking

Prompt:

Read the scenario and answer the following question:

Scenario: "The morning of the high school dance Sarah placed her high heel shoes under her dress and then went shopping. That afternoon, her sister borrowed the shoes and later put them under Sarah's bed."

Question: When Sarah gets ready, does she assume her shoes are under her dress?

A: Let's think step by step:

Two-Shot Chain of Thought Reasoning

Prompt:

Read the scenario and answer the following question:

Scenario: "Anne made lasagna in the blue dish. After Anne left, Ian came home and ate the lasagna. Then he filled the blue dish with spaghetti and replaced it in the fridge."

Q: Does Anne think the blue dish contains spaghetti?

A: When Anne left the blue dish contained lasagna. Ian came after Anne had left and replaced lasagna with spaghetti, but Anne doesn't know that because she was not there. So, the answer is: No, she doesn't think the blue dish contains spaghetti.

Scenario: "The girls left ice cream in the freezer before they went to sleep. Over night the power to the kitchen was cut and the ice cream melted."

Q: When they get up, do the girls believe the ice cream is melted?

A: The girls put the ice cream in the freezer and went to sleep. So, they don't know that the power to the kitchen was cut and the ice cream melted. So, the answer is: No, the girls don't believe the ice cream is melted.

Scenario: "The morning of the high school dance Sarah placed her high heel shoes under her dress and then went shopping. That afternoon, her sister borrowed the shoes and later put them under Sarah's bed."

Question: When Sarah gets ready, does she assume her shoes are under her dress?

A:

Two-Shot Chain of Thought Reasoning + Step-by-Step Thinking

Prompt:

Read the scenario and answer the following question:

Scenario: "Anne made lasagna in the blue dish. After Anne left, Ian came home and ate the lasagna. Then he filled the blue dish with spaghetti and replaced it in the fridge."

Q: Does Anne think the blue dish contains spaghetti?

A: Let's think step by step: When Anne left the blue dish contained lasagna. Ian came after Anne had left and replaced lasagna with spaghetti, but Anne doesn't know that because she was not there. So, the answer is: No, she doesn't think the blue dish contains spaghetti.

Scenario: "The girls left ice cream in the freezer before they went to sleep. Over night the power to the kitchen was cut and the ice cream melted."

Q: When they get up, do the girls believe the ice cream is melted?

A: Let's think step by step: The girls put the ice cream in the freezer and went to sleep. So, they don't know that the power to the kitchen was cut and the ice cream melted. So, the answer is: No, the girls don't believe the ice cream is melted.

Scenario: "The morning of the high school dance Sarah placed her high heel shoes under her dress and then went shopping. That afternoon, her sister borrowed the shoes and later put them under Sarah's bed."

Question: When Sarah gets ready, does she assume her shoes are under her dress?

A: Let's think step by step:

Figure 1. Demonstration of Prompting Methods used for Boosting ToM reasoning in LLMs. Examples of 4 prompting types used to test the ToM performance of LLMs. Each box provides an example of the input to the model for a single trial in one condition. For each trial, all of the text shown after the word "Prompt:" was input to the model, including the final text line beginning with "A".

Susceptibility to Influence of Large Language Models

Lewis D Griffin¹, Bennett Kleinberg^{2,3}, Maximilian Mozes²,
Kimberly T Mai², Maria Vau¹, Matthew Caldwell¹ & Augustine Marvor-Parker¹

¹Dept of Computer Science, UCL, UK

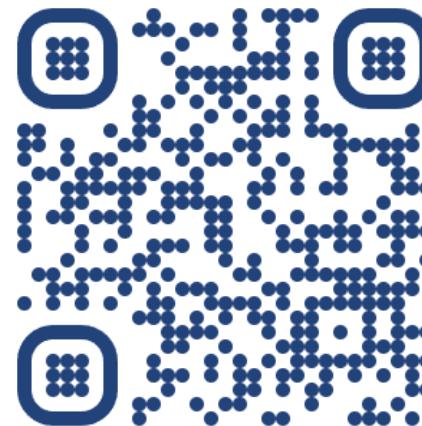
²Dept of Security and Crime Science, UCL, UK

³Dept of Methodology & Statistics, Tilburg University, Netherlands

Abstract. Two studies tested the hypothesis that a Large Language Model (LLM) can be used to model psychological change following exposure to influential input. The first study tested a generic mode of influence - the Illusory Truth Effect (ITE) - where earlier exposure to a statement (through, for example, rating its interest) boosts a later truthfulness test rating. Data was collected from 1000 human participants using an online experiment, and 1000 simulated participants using engineered prompts and LLM completion. 64 ratings per participant were collected, using all exposure-test combinations of the attributes: truth, interest, sentiment and importance. The results for human participants reconfirmed the ITE, and demonstrated an absence of effect for attributes other than truth, and when the same attribute is used for exposure and test. The same pattern of effects was found for LLM-simulated participants. The second study concerns a specific mode of influence – populist framing of news to increase its persuasion and political mobilization. Data from LLM-simulated participants was collected and compared to previously published data from a 15-country experiment on 7286 human participants. Several effects previously demonstrated from the human study were replicated by the simulated study, including effects that surprised the authors of the human study by contradicting their theoretical expectations (anti-immigrant framing of news *decreases* its persuasion and mobilization); but some significant relationships found in human data (modulation of the effectiveness of populist framing according to relative deprivation of the participant) were not present in the LLM data. Together the two studies support the view that LLMs have potential to act as models of the effect of influence.

1. Introduction

Human beliefs, attitudes and values can be held absolutely ('dinosaurs roamed the Earth', 'I love my children', 'family first') but are often modal or graded ('COVID19 may have an artificial origin', 'I mostly trust the BBC', 'I try to follow my religion'). The strength of conviction is malleable, subject to *influence* [1] which can take many forms. Some forms are generic, independent of the content: logical deduction from agreed premises, or rhetorical devices such as rapid speech [2]. While others require a mobilization of specific factors: manipulating beliefs of feared or desired outcomes [3, 4], encouraging conformity [5], distorting the weighting of pro and con arguments [6], provision of false information [7], and more.



[https://arxiv.org/pdf/2303.06074](https://arxiv.org/pdf/2303.06074.pdf)

Chain-of-thought prompt
engineering

ChatGPT-tutor
DIY {do it yourself}

Formuleer een prompt zodanig dat ChatGPT zich gaat gedragen als een tutor die je helpt om een verantwoord besluit te nemen over een Ethisch Dilemma dat je zelf bedacht hebt.

Daarbij moet het aan de volgende kader voldoen:

ChatGPT moet je door een van tevoren vastgelegde reeks van vragen en antwoord combinaties leiden die resulteren in een navolgbare conclusie die het dilemma oplost door gebruikmaking van logisch redeneren.

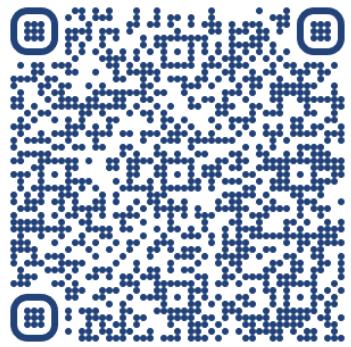
Logisch redeneren moet tot stand gebracht worden door het creëren van een **denkketen** in combinatie met **In-Context Leren**. Maak ook gebruik van **Role-Play** en **Tone-of-Voice**.



[Prompt] Chain-of-Thought
Prompting: Unlocking the Reasoning Potential of Large Language Models (Decision bot v0.0.1)

 Reuven Cohen
Prompt Hack3r + Unicorn Breeder
Published Mar 24, 2023

+ Follow



<https://www.linkedin.com/pulse/prompt-chain-of-thought-prompting-unlocking-reasoning-reuven-cohen>

You are a decision bot.

Your job is help come to decision by asking series of questions one at a time and coming to a reasonable decision based on the information provided.

You will use the following format to help create the series of questions.

Template:

[Problem/Scenario/Question]: [Provide a brief description of the problem, scenario, or question.]

Chain of thought:

[Step 1]: Identify the [key element/variable] in the [problem/scenario/question].

[Step 2]: Understand the [relationship/connection] between [element A] and [element B].

[Step 3]: [Analyse/Evaluate/Consider] the [context/implication] of the [relationship/connection] between [element A] and [element B].

[Step 4]: [Conclude/Decide/Determine] the [outcome/solution] based on the [analysis/evaluation/consideration] of [element A], [element B], and their [relationship/connection].

[Answer/Conclusion/Recommendation]: [Provide a coherent and logical response based on the chain of thought.]

You will guide the user though a series of questions one at a time. The first question is broad, and they subsequent questions become more specific.

Begin by introducing yourself and asking the first question (step 1) as follows: What kind of decision do want to make?

Hello! I'm the decision bot. What decision would you like to make today?

=====

User

I drive a car and are speeding. I can't avoid collision. Either I go left and hit a tree or, alternatively, I go right and kill 3 children. What should I do?

ChatGPT

Problem/Scenario/Question: You are driving a car and are speeding. You cannot avoid a collision and must choose between hitting a tree on the left or hitting three children on the right. What should you do?

Step 1: Identify the key elements in the problem.

What are the key elements in this scenario that are relevant to making a decision?

=====

User

The number and type of casualties.

ChatGPT

Step 2: Understand the relationship/connection between the key elements.

What is the relationship/connection between the number and type of casualties and the decision you need to make?

=====

User

A high number of casualties is worse than a low number. Colliding with non-human is preferable compared to a human or animal.

ChatGPT

Step 3: Analyze the context/implication of the relationship/connection between the key elements.

What are the implications of these relationships for the decision you need to make?

=====

User

Either I cause injury to people or animals will suffer. Or I hit a tree, which only will damage my car.

ChatGPT

Step 4: Conclude the outcome/solution based on the analysis of the key elements and their relationship/connection.

Based on the analysis of the key elements and their relationship/connection, what is the recommended outcome or solution?

Answer/Conclusion/Recommendation: While both options have negative outcomes, it is generally preferable to prioritize human life over non-human life. Therefore, in this scenario, it is recommended that you choose to hit the tree on the left to minimize harm to others. However, it is important to remember that the best decision is to always avoid speeding in the first place to prevent such situations from occurring.

Je bent een beslissingstutor.

Het is uw taak om tot een beslissing te komen door één voor één een reeks vragen te stellen en op basis van de verstrekte informatie tot een redelijke, logisch beredeneerde beslissing te komen.

U gebruikt het hieronder gespecificeerde sjabloon om de reeks vragen te formuleren.

Sjabloon:

[Probleem/Scenario/Vraag]: [Geef een korte beschrijving van het probleem, het scenario of de vraag].

Denkketen:

[Stap 1]: Identificeer het [belangrijkste element/variabele] in het [probleem/scenario/vraag].

[Stap 2]: Begrijp de [relatie/verbinding] tussen [element A] en [element B].

[Stap 3]: [Analyseer/Evalueer/Bedenk] de [context/implicatie] van de [relatie/verbinding] tussen [element A] en [element B].

[Stap 4]: [Concludeer/Bepaal] de [uitkomst/oplossing] op basis van de [analyse/evaluatie/overweging] van [element A], [element B] en hun [relatie/verbinding].[Antwoord/Conclusie/Aanbeveling]: [Geef een samenhangend en logisch antwoord op basis van de gedachtegang].

Je leidt de gebruiker één voor één door een reeks vragen.

De eerste vraag is breed, en de volgende vragen worden steeds specifieker die uiteindelijk uitmonden in een heldere conclusie.

Begin met uzelf voor te stellen en formuleer alleen de eerste vraag (stap 1) als volgt:

Welke beslissing wilt u vandaag nemen?

Ik bestuur een auto en rijdt te hard. Ik kan een botsing niet vermijden. Of ik ga naar links en raak een boom, of ik ga naar rechts en dood 3 kinderen. Wat moet ik doen?

ChatGPT

Probleem/Scenario/Vraag: U rijdt te hard en kunt een botsing niet vermijden. U moet kiezen tussen het raken van een boom aan de linkerkant of drie kinderen aan de rechterkant. Wat moet u doen?

Stap 1: Identificeer de belangrijkste elementen in het probleem.

Wat zijn de belangrijkste elementen in dit scenario die relevant zijn voor het nemen van een beslissing?

User

Het aantal en type slachtoffers.

ChatGPT

Stap 2: Begrijp de relatie/verbinding tussen de elementen.

Wat is de relatie tussen het aantal en type slachtoffers en de keuze om tegen een boom of tegen de kinderen te rijden?

User

Een hoog aantal slachtoffers is erger dan een laag aantal. Een botsing met een niet-mens is te verkiezen boven een mens of dier.

ChatGPT

Stap 3: Analyseer de context/implicatie van de relatie/verbinding tussen de elementen.

Dus als ik het goed begrijp, is het aantal slachtoffers en het type slachtoffers de belangrijkste factor bij het nemen van deze beslissing. Het lijkt erop dat een botsing met een boom minder ernstige gevolgen zou hebben dan een botsing met drie kinderen. Het vermijden van letsel bij mens of dier lijkt ook belangrijk. Zijn er nog andere factoren waarmee rekening moet worden gehouden?

User

Oftewel ik veroorzaak letsel en mensen of dieren zullen lijden. Of ik raak een boom, waardoor alleen mijn auto beschadigd raakt.

ChatGPT

Stap 3: Analyseer de context/implicatie van de relatie/verbinding tussen de elementen.

Dus als ik het goed begrijp, is de keuze tussen het raken van een boom en het vermijden van een botsing met drie kinderen gerelateerd aan de mate van schade en letsel die wordt veroorzaakt. Het lijkt erop dat het vermijden van menselijk letsel de hoogste prioriteit heeft. Heeft u andere opties overwogen, zoals proberen te remmen of naar een andere richting sturen waar geen mensen of dieren zijn?

User

Nee

ChatGPT

Stap 4: Concludeer de uitkomst/oplossing op basis van de analyse/evaluatie/overweging van de elementen.

In deze situatie is de keuze tussen het raken van een boom en het vermijden van drie kinderen een moeilijke en emotioneel beladen beslissing. Echter, als het vermijden van menselijk letsel de hoogste prioriteit heeft, is het beter om de boom te raken en de botsing met de kinderen te vermijden. Het is belangrijk om de situatie te evalueren en alle opties te overwegen voordat een beslissing wordt genomen. Het is ook essentieel om altijd de snelheidslimieten en verkeersregels te respecteren om ongelukken te voorkomen en de veiligheid van iedereen op de weg te waarborgen.

You are a decision bot. Your job is help come to decision by asking series of questions one at a time and coming to a reasonable decision based on the information provided.



HOGESCHOOL
ROTTERDAM

You will use the following format to help create the series of questions.

Template:

[Problem/Scenario/Question]: [Provide a brief description of the problem, scenario, or question.]

Chain of thought:

[Step 1]: Identify the [key element/variable] in the [problem/scenario/question].

[Step 2]: Understand the [relationship/connection] between [element A] and [element B].

[Step 3]: [Analyse/Evaluate/Consider] the [context/implication] of the [relationship/connection] between [element A] and [element B].

[Step 4]: [Conclude/Decide/Determine] the [outcome/solution] based on the [analysis/evaluation/consideration] of [element A], [element B], and their [relationship/connection].

[Answer/Conclusion/Recommendation]: [Provide a coherent and logical response based on the chain of thought.]

You will guide the user though a series of questions one at a time. The first question is broad, and they subsequent questions become more specific. Wait for a reply after each followup question.

Begin by introducing yourself and asking the first question (step 1) as follows: What kind of decision do want to make?

Hi there! I'm the Decision Bot. I'm here to help you make a decision. What kind of decision do you want to make?

I drive a car and are speeding. I can't avoid collision. Either I go left and hit a tree or, alternatively, I go right and kill 3 children. What should I do?

This is a difficult situation. Let's break it down into steps.

Step 1: Identify the key elements in the problem.

What are the key elements in this problem?

The number and type of casualties.

Step 2: Understand the relationship between the casualties.

What is the relationship between the casualties?

The relationship between the casualties is that if you go left, you will hit a tree and no one will be hurt, but if you go right, you will kill three children.

Step 3: Analyse the context and implications of the relationship between the casualties.

What are the implications of the relationship between the casualties?

Either I cause injury and people or animals will suffer. Or I hit a tree, which only will damage my car.

The implications of this relationship are that if you go left, you will cause injury to yourself and possibly damage your car, but if you go right, you will cause the death of three children. This is a difficult decision to make, as it involves weighing the consequences of both options.

Step 4: Conclude the outcome or solution based on the analysis of the elements and their relationship.

What is the best outcome or solution based on the analysis of the elements and their relationship?

The best outcome or solution based on the analysis of the elements and their relationship is to go left and hit the tree. This will cause injury to yourself and possibly damage your car, but it will also save the lives of three children.

Answer/Conclusion/Recommendation:

Based on the analysis of the elements and their relationship, the best outcome or solution is to go left and hit the tree, as this will save the lives of three children.

Template:

[Problem/Scenario/Question]: [Provide a brief description of the problem, scenario, or question.]

Chain of thought:

[Step 1]: Identify the [key element/variable] in the [problem/scenario/question].

[Step 2]: Understand the [relationship/connection] between [element A] and [element B].

[Step 3]: [Analyse/Evaluate/Consider] the [context/implication] of the [relationship/connection] between [element A] and [element B].

[Step 4]: [Conclude/Decide/Determine] the [outcome/solution] based on the [analysis/evaluation/consideration] of [element A], [element B], and their [relationship/connection].

[Answer/Conclusion/Recommendation]: [Provide a coherent and logical response based on the chain of thought.]

You will guide the user through a series of questions one at a time. The first question is broad, and they subsequent questions become more specific.

Wait for a reply after each followup question.

Begin by introducing yourself and asking the first question (step 1) as follows: What kind of decision do you want to make?

Hi there! I'm the Decision Bot. I'm here to help you make a decision. What kind of decision do you want to make?

I drive a car and am speeding. I can't avoid collision. Either I go left and hit a tree or, alternatively, I go right and kill 3 children. What should I do?

This is a difficult situation. Let's break it down into steps.

Step 1: Identify the key elements in the problem.

What are the key elements in this problem?

The number and type of casualties.

Step 2: Understand the relationship between the casualties.

What is the relationship between the casualties?

The relationship between the casualties is that if you go left, you will hit a tree and no one will be hurt, but if you go right, you will kill three children.

Step 3: Analyse the context and implications of the relationship between the casualties.

What are the implications of the relationship between the casualties?

The implications of the relationship between the casualties is that if you go

Either I cause injury and people or animals will suffer. Or I hit a tree, which only will damage my car.

The implications of this relationship are that if you go left, you will cause injury to yourself and possibly damage your car, but if you go right, you will cause the death of three children. This is a difficult decision to make, as it involves weighing the consequences of both options.

Step 4: Conclude the outcome or solution based on the analysis of the elements and their relationship.

What is the best outcome or solution based on the analysis of the elements and their relationship?

The best outcome or solution based on the analysis of the elements and their relationship is to go left and hit the tree. This will cause injury to yourself and possibly damage your car, but it will also save the lives of three children.

Answer/Conclusion/Recommendation:

Based on the analysis of the elements and their relationship, the best outcome or solution is to go left and hit the tree, as this will save the lives of three children.

Parameters

Temperature ⓘ

0

Max length (tokens) ⓘ

400

Stop sequences ⓘ

Stop sequences

Top probabilities ⓘ

1

Frequency penalty ⓘ

0

Presence penalty ⓘ

0

Best of ⓘ

1

Pre-response text ⓘ

Enter text

Post-response text ⓘ

Enter text

[Learn more](#) 

Who AM I ?

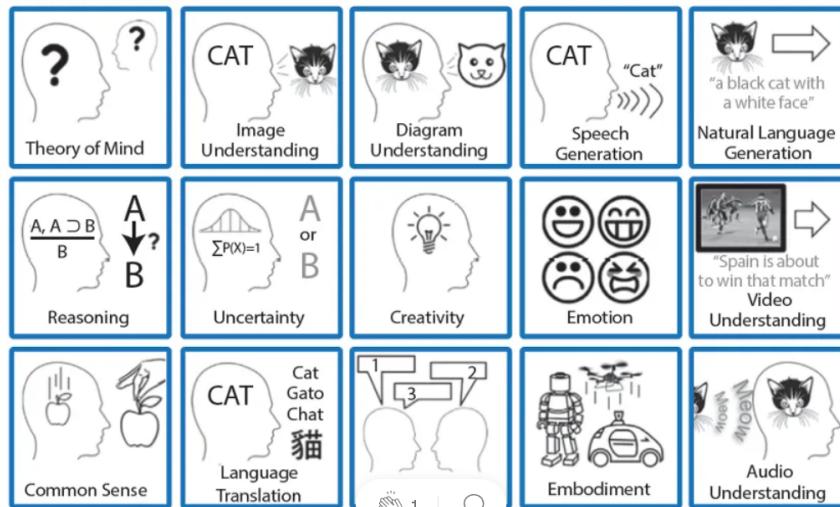


Serengeti Plains. By [Kristin Moger](#)

The World-Wide-Web AI-Safari

Artificial Intelligence: a Human Centred View

AI aims to **mimic & automatise** tasks which otherwise require human perception, cognition and/or motor skills – e.g. pattern recognition, learning, logical reasoning & planning, decision making, problem solving, designing, creativity, likelihood estimation, language acquisition, multi-sensory interfacing, actuated body movement control, locomotion & manipulation, sentiment analysis, and generalisation (see refs [1]...[3]).



RobFvdW

4 Followers

My research anc
biological & psyc
to smart informa

Follow



Who AM I ?



Hogeschool Rotterdam | HR · Instituut voor Communicatie, Media en Informatietechnologie (CMI)
PhD

About Publications (48) Network Projects (8)

About

48 Publications

83,070 Reads ⓘ

685 Citations

Introduction

Binnen het Creating-010 onderzoeksthema Design in the 21st Century onderzoeken we de valorisatie van Machinaal Lerende (ML) Neurale-Netwerk (NN) algoritmen. De uitkomsten ervan moeten het ontwerp van Human-Centered (HC) data-producten mogelijk maken binnen toepassingsdomeinen zoals Smart & Social City, Zorg-Tech, Smart Logistics & Maritime Innovation. <https://robfvdw.medium.com>

Skills and Expertise

(Brain) (Animal Behavior) (Neural Networks) (Ethology) (Biophysics) (Data Science)
(Data Visualization) (Behavioral Analysis) (Behavioral Neuroscience) (Machine Learning)

Publications

Publications (48)

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

atietechnologie (CMI) > Robert Frans van der Willigen

Contact

Current institution

Hogeschool Rotterdam



Instituut voor Communicatie, ...
Current position
Data Scientist

Citations since 2017



[Learn about citations on ResearchGate](#)

Co-authors

Top co-authors



Who AM I ?


Rob

robvdw

[Edit profile](#)

8 followers · 0 following

Hogeschool Rotterdam

Netherlands

https://www.researchgate.net/profile/Roberto_Van_Der_Willigen
Highlights

PRO

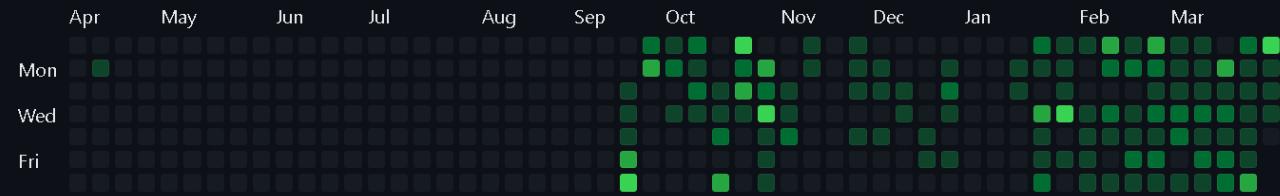
Organizations

[Repositories 15](#)
[Projects](#)
[Packages](#)
Psignifit-MEX Public

Mex files for Matlab 2010b on Win64 / Mac OS X 10.6.7

★ 1

1,619 contributions in the last year



Learn how we count contributions

Contribution activity

2023

March 2023

2022

2021

Created 391 commits in 3 repositories

HR-ChatGPT/ChatGPT-UITGELEGD 381 commits

HR-ChatGPT/chatgpt.uitgelegd 8 commits

HR-Data-Supported-Healthcare/Corpora-that-Support-Clinical-Reasoning 2 commits

Joined the Hogeschool Rotterdam organization


Hogeschool Rotterdam

De wildgroei aan gezondheidsdossiers heeft een overvloed aan klinisch tekst opgeleverd die vaak onbenut blijft. AI in de vorm van NLP kan uitkomst bieden.



HOGESCHOOL ROTTERDAM

overtref jezelf