
Overzicht van veel gestelde vragen [ChatGPT FACs]

***Disclaimer: deze tekst is door het gebruik van gezond verstand tot stand gekomen.
Artificiële intelligentie is gebruikt ter verificatie van de gebruikte bronnen.***

- [0] [Waarom veroorzaakt ChatGPT zoveel onrust?](#)
- [0a] [Wat is generatieve kunstmatige intelligentie \[Gen-AI\]?](#)
- [0b] [Is er een kort overzicht van ChatGPT's tekortkomingen?](#)
- [0c] [+Moet ik me zorgen maken over ChatGPT technologie?](#)
- [0d] [+Is er een ChatGPT "code-of-conduct" en/of richtlijn voor hbo docenten??](#)
- [0e] [+Kan ChatGPT benut worden als beoordelingsinstrument?](#)
- [0f] [+Voldoen Gen-AI —zoals ChatGPT— aan Europese AI-Regelgeving?](#)
- [1a] [Wat moet je weten over ChatGPT en wat kant deze "chatBot"?](#)
- [1b] [Wat zijn de functionele mogelijkheden & —Cyber Security— beperkingen van ChatGPT?](#)
- [1c] [Wat zijn ethische risico's & schaduwkanten van ChatGPT?](#)
- [1d] [Maakt ChatGPT "valsspelen" makkelijker en is het te detecteren?](#)
- [1e] [Kun je ChatGPT opvoeren als co-auteur?](#)
- [1f] [+Kun je ChatGPT citeren als bron?](#)
- [2] [Hoe geef je een opdracht aan ChatGPT?](#)
- [3] [Kan ChatGPT uitleggen hoe het werkt?](#)
- [4] [Kan ChatGPT uitleggen hoe het te gebruiken?](#)
- [5] [Heeft ChatGPT taalbegrip?](#)
- [6] [Kan ChatGPT logisch redeneren?](#)
- [7a] [+Kan ChatGPT broncode schrijven?](#)
- [7b] [Kan ChatGPT broncode uitleggen?](#)
- [7c] [Kan ChatGPT broncode output simuleren?](#)
- [7d] [Kan ChatGPT een "Deep Learning" lessenreeks bedenken?](#)
- [7e] [Kan ChatGPT broncode beoordelen en/of fouten opsporen?](#)
- [8a] [Kan ChatGPT gebruikt worden om bronnen te vermelden?](#)
- [8b] [Kan ChatGPT gebruikt worden om bronnen samen te vatten?](#)
- [13] [Kun je spreken tegen ChatGPT?](#)
- [15] [Prompt "patterns" voorbeelden](#)
- [16] [Geraadpleegde Bronnen](#)

AI's NEW FRONTIER is called Chat-GPT

Dit is een dataproduct gemaakt door het PROMETHEUS DATA SCIENCE LAB van de Hogeschool Rotterdam.

Context & Doel

Leer hoe ChatGPT betrouwbaar te gebruiken

1. Begrijpen wat ChatGPT wel en niet kan
2. Ethische overwegingen bediscussiëren
3. Effectieve prompts schrijven
4. Vervolg vragen schrijven
5. Waarschuwing betrouwbaarheid
6. Overzicht van veel gestelde vragen [\[ChatGPT FACs\]](#).

Om duiding te geven aan de hype rond de generatieve-AI ChatGPT, is deze *Nederlandstalige "How-To?" repository* opgesteld met [hints en tips](#) voor het verantwoord & effectief gebruik ervan in het hoger onderwijs.

Doordat de toepassingsmogelijkheden van ChatGPT eindeloos zijn, worden in deze repository voornamelijk gebruiksmogelijkheden besproken die relevant zijn voor (1) het leerproces van scholieren/studenten, in combinatie met (2) onderwijs-/onderzoektaken van docenten in het (hoger)onderwijs.

Daarnaast worden de randvoorwaarden —*inclusief beoordelingskader + richtlijnen*— en risicofactoren beschreven voor het verantwoord gebruik van [Foundation-models](#) zoals ChatGPT als onderdeel van [BKE](#) (Basis Kwalificatie Examineren) en [SKE](#) (Senior Kwalificatie Examineren) assessment in het hoger onderwijs.

Een centrale vraag is: *"Vormen Foundation modellen een betrouwbare tool die docenten kan helpen bij het beoordelen van content gemaakt door studenten?"*

Er zullen regelmatig updates volgen over nieuwe ontwikkelingen.

Het publiekelijk beschikbaar stellen van generatieve-AI [Gen-AI], zoals [ChatGPT](#) —een antwoord-chatbot gebaseerd op het "pre-trained model"* GPT-3—* en [Galactica](#) —een chatbot voor het schrijven van wetenschappelijk papers—, heeft het debat doen herleven over wat dit betekent voor onderwijsinstellingen.

Volgens een editorial in de [Gardian](#) (10 februari 2023) maakt het vrijgeven en "Hypen" van Gen-AI met "nieuwe, revolutionaire functionaliteit" die onze manier van werken "volledig zal veranderen" deel uit van een commerciële strategie van digitale-platform "vendors" zoals Microsoft (OpenAI), Alphabet (Google), Meta (Facebook). Het doel is om gebruikers zover te krijgen dat ze hun denkend vermogen overdragen aan "alwetende machines". De suggestie van Microsoft dat ChatGPT "slechts" een demo is van het onderzoekslab [OpenAI LP](#), is een vorm van "down playing" om de publieke opinie te bespelen alsof het zou gaan om een ongevaarlijk stuk speelgoed. Niet is minder waar.

Dit doet denken aan de Wimperspitsmuis (*Suncus etruscus*) die als het koud wordt, zijn brein laat krimpen om energie te besparen. Het is onwaarschijnlijk dat de mensheid een dergelijke overlevingsstrategie zal vertonen, maar er is een alarmerende metaforische parallel. Deze door winst gedreven wedloop —met als doel AI in ons dagelijks leven te integreren—, maakt de mens kwetsbaar door volledig te vertrouwen op AI-technologie. In de biologie geldt niet voor niets het aloude adagium "Use It, or Lose It!".

Het is niet ondenkbaar dat als schoolgaande kinderen te veel worden blootgesteld aan AI-gedreven leermiddelen, ze in het hoger onderwijs te kort schieten. Een gevolg kan zijn dat zij een onoverbrugbare achterstand hebben opgelopen in hun cognitieve vaardigheden zoals probleemoplossend vermogen, begrijpend lezen, opsporen van betrouwbare bronnen en oordeelsvorming.

Het Wired IDEAS Blog (09 december 2022) getiteld: "[ChatGPT, Galactica, and the Progress Trap: When large language models fall short, the consequences can be serious.](#)" legt een aantal fundamentele beperkingen bloot van "taalvaardige-AI", zoals ChatGPT, dat tot stand is gebracht met behulp van "Deep Learning" technieken.

Een nagenoeg onoplosbaar probleem is dat ChatGPT *taal agnostisch* is. Door gebruikmaking van "[Machinaal lerende](#)" algoritmen gaat het veelvuldig in de fout met het interpreteren van zinsbouw, maar is nagenoeg foutloos in het vertalen van individuele woorden.

Dergelijke systematische fouten weerspiegelen een fundamentele tekortkoming van Gen-AI anno 2023 omdat het ogenschijnlijk "taalvaardig" lijkt te zijn op basis van enorme hoeveelheden tekst afkomstig van het world-wide-web. Chatbots beschikken dus *nog* niet over linguïstische concepten en probleemoplossend vermogen die nodig zijn om een taal te kunnen verwerken zoals mensen dat zouden doen. Nog problematischer is dat ChatGPT alle dominante opvattingen en vooroordelen —die schadelijk kunnen zijn voor taal van

minderheden—, woord voor woord in zich herbergt precies zoals ze voorkomen op het world-wide-web “zonder aanzien des persoons”.

ChatGPT wordt gevoed met de schoonheid, lelijkheid en wreedheid van het internet en sociale-media. De verwachting dat het ons alleen het goede, het schone en het behulpzame toont is een gevaarlijke en naïeve houding.

Toch is het de verwachting dat deze tekortkomingen maar van tijdelijk aard zullen zijn. Op de preprint server arXiv is eind 2022 een paper verschenen —getiteld: *“Theory of Mind May Have Spontaneously Emerged in Large Language Models”*— waarin is onderzocht in of Gen-AI “False-belief” taken kunnen oplossen. Dit zou een aanwijzing kunnen zijn voor *“Theory-of-Mind” [ToM]*: *“het vermogen om niet-waarneembare mentale toestanden toe te schrijven aan anderen, nodig voor menselijke sociale interacties, communicatie, empathie, zelfbewustzijn en moraliteit.”*

... ToM —tot nu toe beschouwd als uniek menselijk— kan spontaan ontstaan als emergente eigenschap van het opschalen van taalmodellen ...

Het verwerven van emergente eigenschappen door ChatGPT —zoals ToM— moet met enige scepsis worden betracht. Dat wil zeggen, **emergentie** kenmerkt zich door “het ontstaan van nieuwe en samenhangende structuren, patronen en eigenschappen tijdens het proces van zelforganisatie in complexe systemen”. Deze bevinding is moeilijk te verenigen met het *taal agnostisch* karakter van ChatGPT. Maar het maakt duidelijk dat chatbots die het Gen-AI landschap tot 2020 hebben gedomineerd, plaats maken voor *“Foundation models”*: taalmodellen getraind met enorme hoeveelheden tekst —*grotendeels afkomstig van het world-wide-web*— die nieuwe taken kunnen uitvoeren —*few-shot learning*— op basis van slechts paar voorbeelden. In een veel geciteerd arXiv paper getiteld: *“On the Opportunities and Risks of Foundation Models”* wordt benadrukt dat “foundation models” een paradigm-shift hebben veroorzaakt die van dezelfde orde van magnitude is als de ontwikkeling van “deep Learning models” in 2010.

... “Few-shot learning models”: GPT-3, BERT, Midjourney of DALL-E 2, hebben laten zien wat er mogelijk is. Je voert een korte vraag in, en het systeem genereert een heel opstel of een complexe afbeelding op basis van jouw parameters, zelfs als het niet specifiek getraind is op het uitvoeren van dat exacte argument of het genereren van een afbeelding op die manier ...

Door de populariteit van ChatGPT zullen “Foundation models” ongetwijfeld veelvuldig worden ingezet voor “social engineering”, “social manipulation” en marketingdoeleinden. Maar een veel groter probleem is dat met het verder opschalen van chatbots, ze nog “menselijker” zullen worden, waardoor ze een zeer grote aantrekkingskracht zullen hebben op jongeren. In de nabije toekomst valt dus niet uit te sluiten dat studenten hun “mentale

autonomie” verliezen door het veelvuldig gebruik van AI-technologie. Dit is vergelijkbaar met de huidige problematiek die speelt rondom het gebruik van sociale-media door grote groepen jonge eindgebruikers.

Het is dus zaak dat onderwijsinstellingen regie nemen over de ontwikkeling en inzet van AI-gedreven leermiddelen. Met andere woorden “*Zijn hogescholen voldoende voorbereid op de opmars van foundation models?*” Het antwoord is “zeer waarschijnlijk niet”. Hogescholen in Nederland zijn vooral gefocust op het toepassen van AI op basis van relatief kleine datasets. Tegelijkertijd bezitten ze enorme hoeveelheden aan hoogwaardige datasets, die niet beschikbaar zijn voor het publiek. Dit geeft ze de mogelijkheid om hun eigen versies van ChatGPT te ontwikkelen, de data in licentie te geven, en de redactie- en beoordelingsprocessen te herstructureren om meer waarde te creëren voor de toekomstige GEN-AI. Het is daarom belangrijk om te begrijpen wat ChatGPT *wel* en *niet* kan doen en *waarom*. Ook moeten de morele en ethische aspecten van het gebruik ervan niet onbesproken blijven.

[0] WAAROM VEROORZAAKT ChatGPT ZOVEEL ONRUST?

Typerend voor AI anno 2023 is de lerende machine [ML] genaamd [ChatGPT](#). Een state-of-the-art, grootschalig taalmodel [LLM] dat gebruik maakt van natuurlijke taal verwerkende [NLP] AI-technologie. Het is gevoed met meer dan 8 miljoen unieke dialogen.

ChatGPT's gebruikersinterface is ontworpen om menselijke conversatie na te bootsen. Het revolutionaire aan deze Generatieve AI-technologie zijn de ogenschijnlijk levensechte gesprekken die het kan onderhouden met mensen. Het behoort daardoor tot een van de meest geavanceerde "*conversationale agenten*" die publiekelijk beschikbaar is gesteld.

Nu GPT de nieuwste sensatie in de wereld van kunstmatige intelligentie [AI] is, probeert Sam Altman —*chief executive of OpenAI*— de effecten ervan te bagatelliseren. Volgens "*The New York Times (3 februari 2023)*" vreest hij dat te veel aandacht en rumoer rondom ChatGPT een regelgevende reactie kan uitlokken vanuit overheden. Of onrealistische verwachtingen bij eindgebruikers over de functionaliteit in toekomstige releases.

Op Twitter heeft hij geprobeerd de gemoederen wat te bedaren door ChatGPT "*ongelooflijk beperkt*" te noemen en gebruikers te waarschuwen dat "*het een vergissing is om er nu op te vertrouwen voor iets belangrijks*".

Een illustratief voorbeeld van waar de hype rond ChatGPT toe kan leiden blijkt uit een Blog — *getiteld: [ChatGPT Keeps Imploding Because of Crochet. \(Seriously.\)](#)* — in the Daily Beast (05 februari 2023) geschreven door innovatie-reporter Katie Notopoulos. Zij beschrijft hoe een groep van 100.000 TikTokers onder leiding van textielkunstenares Alex Woolner — *die zichzelf de "Crochet Army" noemen* — een zwakke plek wisten bloot te leggen door het te "*misbruiken*" voor het genereren van "*Crochet*" haakpatronen. ChatGPT kan niet omgaan met de complexiteit van de patronen die de Crochet Army had bedacht. Het resultaat zijn een onleesbare reeksen aan letters en cijfers. Het ontbreekt de chatbot aan probleemoplossend vermogen om te kunnen bepalen dat de patronen niet moeten worden geïnterpreteerd als talige input maar als "*output*" dat mensen kunnen gebruiken als haakpatroon.

Door ChatGPT's enorme populariteit is er een stortvloed aan verhalen losgekomen over wat het zoal kan: academische essays schrijven, medische examens afnemen, maaltijden voorbereiden en misschien (op een dag) optreden als advocaat in de rechtszaal. Maar voor elk succesverhaal is er een ander dat het falen ervan belicht: Waarom kan ChatGPT wel een navolgbaar essay schrijven,

maar niet een van literaire kwaliteit? Waarom heeft het moeite met sommige soorten logische vragen en geeft het zelfverzekerd onjuiste antwoorden op andere?

Ondanks dat OpenAI goede sier maakt met het vrij toegankelijk maken van ChatGPT, is het belangrijk om te weten dat het een *“work in progress”* is dat nog volop in ontwikkeling is, waarbij de onderliggende *“deep learning”* AI-technologie grotendeels de *“brainchild”* is van Alphabet —het moeder bedrijf van Google— voortgekomen uit de onderzoekslaboratoria van *“Google Brain”* (2011) en *“DeepMind”* (2014). Dit geldt in mindere mate ook voor Meta —het moederbedrijf van Facebook— dat sinds 2013 aan de ontwikkeling van *“FAIR”* (Facebook Artificial Intelligence Research) werkt. FAIR is overgaan in *Meta.AI* (2015). Alphabet en Meta zijn de meest invloedrijke vernieuwers in de wereld van kunstmatige intelligentie [AI]. Ze beschikken over enorm veel (financiële) middelen, ervaring, werkgemeenschappen en hebben toegang tot zeer grote datasets. Toch, hebben zij een andere afweging gemaakt dan OpenAI uit zowel een moreel als een commercieel —*reputatie schade*— oogpunt.

Yann LeCun —*Cofounder Meta-AI*— zei op Twitter (8 januari 2023) dat Alphabet en Meta hun grootschalig taalmodellen [LLMs] niet vrijgeven aan het grote publiek om ethische redenen. LLMs’ genereren nog te vaak foutieve en/of giftige teksten. In dezelfde Twitter thread, toen hij erop gewezen werd het *“public relations fiasco”* rondom de introductie van Meta-AI’s *Galactica* niet te vermelden bij het bekritisieren van ChatGPT. Galactica moest wetenschappers helpen *“academische papers samenvatten, wiskundige problemen oplossen, Wiki-artikelen genereren, wetenschappelijke code schrijven, moleculen en eiwitten annoteren, en nog veel meer”*, in plaats daarvan, spuwde het gedachteloos bevooroordeelde en onjuiste onzin uit. Drie dagen na de lancering (18 november 2022), werd het uitgeschakeld.

De grote afwezige lijkt Apple —’s werelds grootste informatietechnologiebedrijf qua omzet en het allereerste beursgenoteerde bedrijf met een waarde van meer dan 1 miljard dollar—. Research naar *toepassen van AI bij Apple** richt zich vooral op de *“Apple Neural Engine [ANE]”*. Het is een op maat gemaakte chip die speciaal is ontworpen voor biometrische deep learning in smart-devices en laptops. Hierdoor kunnen functies als Face ID-aanmeldingen, detecteren van menselijke poses, functies in de camera waarmee gebruikers betere foto’s kunnen maken (of gekke effecten kunnen toevoegen), augmented reality en het beheren van de batterijduur worden ondersteund. Dus Apple’s toekomstvisie op AI bestaat uit krachtige handhelds die in staat zijn hun eigen machine learning toe te passen op datasets die zijn verzameld via hun eigen reeks sensoren. Dit staat duidelijk haaks op de visie van een toekomst die gedomineerd wordt door cloud computing waar de *“andere”* tech-giants zoals Alibaba,

Alphabet, Amazone, Baidu, Bosch, Cisco, IBM, Meta, Microsoft, Nvidia, Samsung, Siemens en Tesla naar toe lijken te werken.

Duidelijk is dat taak specifieke modellen die het AI-landschap tot 2020 hebben gedomineerd plaats maken voor *“Foundation models”*: modellen getraind op een zeer grote ongelabelde datasets die voor meerdere taken tegelijkertijd kunnen uitvoeren. De eerste tekenen van het potentieel van *“foundation-modellen”* werden gebruikt voor het genereren van beeld en taal. In het baanbrekende arXiv paper getiteld: *“On the Opportunities and Risks of Foundation Models”* wordt benadrukt dat *“foundation models”* een paradigm-shift teweeg brengen in de wereld van de AI die van de zelf-de orde van magnitude is als de ontwikkeling van *“deep Learning models”* in 2010.

Grote taalmodellen zoals GTP's kunnen worden *“gevraagd”* —in het Engels heet dit *“prompting”*— om een reeks taken op het gebied van natuurlijke taalverwerking (NLP) uit te voeren, gegeven enkele voorbeelden van de taak als invoer. In review paper getiteld: *“Training language models to follow instructions with human feedback”* wordt echter uitgelegd waarom LLM's regelmatig *“onbedoeld gedrag”* vertonen, zoals het verzinnen van feiten, het genereren van bevooroordeelde of giftige tekst, of het simpelweg het *niet opvolgen* van prompts. Dit komt omdat de *“taalmodelleringsdoelstelling”* —het voorspellen van de volgende woord/token op een webpagina van het internet— verschilt van *“prompting”* —de instructies van de gebruiker hulpvaardig en veilig opvolgen—. Het voorkomen van dit onbedoelde gedrag is vooral belangrijk voor taalmodellen die in honderden verschillende toepassingen worden ingezet en gebruikt.

Het gebruik van ChatGPT is dus niet zonder risico's. Het zal zeer waarschijnlijk nog decennia duren voordat het vrij van ongewenst taalgebruik, foutloos en met een hoge betrouwbaarheid, op overtuigende wijze *“spontane”* gesprekken kan voeren met mensen die woord voor woord feitelijk juist zijn.

Op korte termijn zal het veelvuldig worden ingezet voor *“social engineering”*, *“social manipulation”* en marketing doeleinden. Het is daarom belangrijk om te begrijpen wat ChatGPT wel en niet kan doen en waarom. Ook moeten de morele en ethische aspecten van het gebruik ervan niet onbesproken blijven.

[0a] Wat is “*generatieve kunstmatige intelligentie*” [Gen-AI]?

De onderstaande tekst is deels ontleent aan het *Times Magazine* artikel getiteld “*The AI Arms Race Is Changing Everything*”.

“Kunstmatige intelligentie” (AI) is een multidisciplinair vakgebied gericht op het ontwikkelen van technologie dat menselijke cognitieve, perceptuele en/of motorische vermogens kan nabootsen en/of automatiseren.

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

Voorbeelden van menselijke vermogens in de vorm van AI-toepassingen zijn interactief uit te proberen via de World-Wide-Web AI Safari.

Generatieve AI [Gen-AI] is een “*digitale content generende technologie*” met als doel het volledig automatisch produceren van ogenschijnlijk “*nieuwe*” inhoud, zoals tekst, afbeeldingen, geluid, spraak en/of muziek. Gen-AI kan een “deep learning” [DL] model creëren van een bestaand schilderij, met als doel nieuwe beelden te creëren die op het oorspronkelijke schilderij lijken. Door vervolgens een serie aan nieuwe beelden te genereren en die na elkaar af te spelen, ontstaat een animatie.

Living Mona Lisa from Few-Shot Adversarial Learning of Realistic Neural Talking Head Models. Zakharov, E., Shysheya, A., Burkov, E., & Lempitsky, V. (2019). Few-shot adversarial learning of realistic neural talking head models. In Proceedings of the IEEE/CVF international conference on computer vision (pp. 9459-9468). <https://doi.org/10.1109/ICCV.2019.00671>

“Generative Pre-trained Transformers” [GPTs] zijn anno 2023 de meest dominante verschijningsvorm van Gen-AI. GTPs maken gebruik van op neurale netwerk [NN] architectuur gebaseerde “machinaal Lerende” [ML] algoritmen. Het zijn automaten die natuurlijke taal kunnen verwerken & genereren [NLP]. ChatGPT is de meest geavanceerde GPT die publiekelijk toegankelijk is gesteld door OpenAI eind 2022.

Echter het gebruik van ChatGPT is niet zonder risico's. GPT's vertonen namelijk dezelfde problemen als het gebruik van sociale-media door grote groepen eindgebruikers. Onderzoekslaboratoria hebben Gen-AI jarenlang achter gesloten deuren gehouden, terwijl ze de gevaren ervan bestudeerden, van verkeerde informatie en haatzaaien tot het ongewild creëren van een sneeuwbaaleffect van geopolitieke crises.

Deze terughoudendheid komt deels voort uit de onvoorspelbaarheid van neurale netwerk [NN] architectuur, het computationele paradigma waarop deep learning [DL] is gebaseerd. In plaats van de traditionele aanpak van computerprogrammering, die uitgaat van precieze reeksen instructies die voorspelbare resultaten opleveren, leren neurale netwerken [NN] zichzelf effectief om patronen in de datasets te herkennen waarmee ze getraind worden (zie [Wat moet je weten over Gen-AI zoals ChatGPT](#)).

De eerste generatie talige Gen-AI's bleken pijnlijk gevoelig voor het napraten van de vooroordelen in hun trainingsgegevens: ze spuwden verkeerde informatie en haatzaaiende taal. Toen Microsoft in 2016 zijn [chatbot Tay](#) onthulde, duurde het minder dan 24 uur voordat giftige, haat zaaiende, racistische tweets uitspuwde. De ontwikkeling van Gen-AI's kwam pas echt in een stroomversnelling vanaf 2017. Aangezwengeld door enkele cruciale doorbraken in het ontwerp van neurale netwerken zoals *transformers*, *auto-encoders* en *diffusion*, de toenemende beschikbaarheid van gegevens en de bereidheid van technologiebedrijven om te betalen voor gigantische hoeveelheden rekenkracht. Maar de zwakke plekken bleven, en de geschiedenis van beschamende AI-struikelblokken maakte veel bedrijven, waaronder Alphabet, Meta en OpenAI, terughoudend om hun meest geavanceerde GEN-AI modellen openbaar te maken.

Een ander notoir voorbeeld is het *“public relations fiasco”* rondom de introductie van Meta-AI's chatbot [Galactica](#). Galactica moest wetenschappers helpen *“academische papers samenvatten, wiskundige problemen oplossen, Wiki-artikelen genereren, wetenschappelijke code schrijven, moleculen en eiwitten annoteren, en nog veel meer”*, in plaats daarvan, spuwde het gedachteloos bevooroordeelde en onjuiste onzin uit. Drie dagen na de lancering (18 november 2022), werd het uitgeschakeld.

In april 2022 kondigde OpenAI Dall-E 2 aan, een tekst-naar-beeld AI-model dat fotorealistische beelden kon genereren. Maar in eerste instantie beperkte OpenAI de

vrijgave tot een wachtlijst van “vertrouwde” gebruikers, wiens gebruik zou helpen om “de vooroordelen die DALL-E heeft geërfd van zijn trainingsgegevens te begrijpen en aan te pakken”. De Londense startup genaamd *Stability AI*, maakte korte metten met deze prudente handelswijze door hun tekst-naar-beeld-tool, *Stable Diffusion* —een samenwerking met de start-up *Runway*—, vrij beschikbaar te stellen voor iedereen die het wilde uitproberen. De ontstaansgeschiedenis van *diffusion* —een Gen-AI gebaseerd op generative adversarial networks, kortweg [GANs]— is in detail beschreven door TechCrunch in een blog getiteld: “*A brief history of diffusion, the tech at the heart of modern image-generating AI*”. *Stable Diffusion* werd al snel een internet hype.

Volgens *Time Magazine* bracht dit Alphabet en OpenAI in rep & roer, want nu was “iedereen” in staat om Gen-AI tools te gebruiken die zij zo “zorgvuldig” hadden afgeschermd.

OpenAI volgde dan ook snel door het afgeschermd Dall-E 2 publiekelijk beschikbaar te stellen. Vervolgens gaf het in november 2022 ChatGPT vrij voor het publiek, naar verluidt om de dreigende concurrentie voor te zijn. OpenAI CEO Sam Altman benadrukte in interviews dat hoe meer mensen AI-programma’s gebruikten, hoe sneller ze zouden verbeteren. In februari 2023 kondigde Alphabet aan om zijn ChatGPT-rivaal Bard uit te brengen. En in het recente kwartaalgesprek van Meta verklaarde CEO Mark Zuckerberg dat hij ernaar streeft dat het bedrijf “een leider wordt in generatieve AI”.

Geselecteerde referenties voor verder lezen

1. Generative AI: Perspectives from Stanford HAI. (2023). How do you think generative AI will affect your field and society going forward?
https://hai.stanford.edu/sites/default/files/2023-03/Generative_AI_HAI_Perspectives.pdf
2. Sheikh, H., Prins, C., & Schrijvers, E. (Eds.). (2023). Mission AI. The New System Technology. WRR, Scientific Council for Government Policy. Springer. <https://doi.org/10.1007/978-3-031-21448-6>

[0b] IS ER EEN KORT OVERZICHT VAN ChatGPT's TEKORTKOMINGEN?

Dit is een aangepaste, Nederlandstalige versie van [5 Big Problems With OpenAI's ChatGPT](#) geschreven door door Garling Wu op 22 december 2022.

ISSUE	OMSCHRIJVING
1. Veel Fouten	<p>Het faalt in elementaire wiskunde en grammatica, het beantwoorden van eenvoudige logica vragen. Zoals gebruikers van sociale media kunnen getuigen, kan ChatGPT het meer dan eens bij het verkeerde eind hebben. OpenAI erkent dit fenomeen en schrijft op haar website het volgende: <i>"ChatGPT schrijft soms plausibel klinkende maar onjuiste of onzinnige antwoorden."</i> Deze "hallucinatie" van feit en fictie, zoals sommige wetenschappers het noemen, is vooral gevaarlijk als het gaat om zoiets als medisch of juridisch advies.</p> <p>In tegenstelling tot andere AI-assistenten zoals Siri of Alexa, heeft ChatGPT niet direct toegang tot het world-wide-web om antwoorden op te sporen en/of te verifiëren. In plaats daarvan wordt een zin, woord voor woord opgebouwd, waarbij op basis van de training de meest waarschijnlijke "woord token" wordt geselecteerd dat erop zou moeten volgen.</p> <p>Met andere woorden, ChatGPT komt tot een antwoord door een reeks aan gissingen, wat een deel van de reden is dat het foute antwoorden kan beargumenteren alsof deze feitelijk juist zijn. Hoewel het goed is in het uitleggen van complexe concepten, waardoor het in potentie een krachtig leermiddel vormt, is het belangrijk niet alles voor waar aan te nemen. ChatGPT heeft het regelmatig bij het verkeerde eind.</p>
2. Vooringenomenheid lijkt ingebakken in het model	<p>ChatGPT is getraind op het collectieve schrijven van mensen over de hele wereld, vroeger en nu. Dit betekent dat dezelfde vooroordelen die in de "echte" wereld bestaan, ook in het model zullen voorkomen. Eindgebruikers hebben meer dan eens gedemonstreerd dat het "onbesuisde" seksistische antwoorden produceerd.</p>

ISSUE	OMSCHRIJVING
	<p>Maar dat is slechts het topje van de ijsberg; het kan antwoorden produceren die uiterst schadelijk zijn voor een reeks minderheidsgroepen. Problematischer is dat developers van OpenAI zelf de gegevens selecteren die worden gebruikt om ChatGPT te trainen.</p> <p>Om wat OpenAI “vooringenomen gedrag” noemt aan te pakken, vraagt het eindgebruikers om feedback te geven op slechte outputs. Met een dergelijk groot potentieel om mensen schade toe te brengen, kun je stellen dat ChatGPT niet aan het publiek had moeten worden vrijgegeven voordat deze problemen zijn bestudeerd en opgelost.</p> <p>Een soortgelijke AI-chatbot genaamd Sparrow —<i>eigendom van Google’s moederbedrijf Alphabet</i>— werd achter gesloten deuren gehouden vanwege vergelijkbare zorgen dat het ongecontroleerde gebruik ervan schade bij mensen zou kunnen veroorzaken.</p> <p>Met moederbedrijf van Facebook, Meta liep tegen vergelijkbare problemen aan. Toen het <i>Galactica</i> uitbracht, een AI-taalmodel getraind op academische papers, werd het snel teruggeroepen nadat veel mensen het bekritiseerden voor het uitvoeren van verkeerde en bevooroordeelde resultaten.</p>
<p>3. Nederlandse grammatica, syntax en spelling bevat regelmatig fouten</p>	<p>Je kunt ChatGPT vragen teksten proef te lezen of aan te geven hoe je een paragraaf kunt verbeteren. Je kunt ook alles aan ChatGPT overlaten en vragen een tekst over een bepaald thema te genereren.</p> <p>Docenten hebben geëxperimenteerd met het voeden van Nederlandse opdrachten. Ze beoordeelde de antwoorden verkregen als beter dan wat veel van hun scholieren/studenten zouden kunnen doen.</p> <p>Neerlandistiek, het online tijdschrift voor de Nederlandse taalkunde, letterkunde en taalbeheersing beschrijft dat “<i>ChatGPT goede teksten kan schrijven, zoals betogen over verschillende onderwerpen.</i></p> <p><i>Ook kan de chatbot teksten produceren die aan bepaalde richtlijnen moeten voldoen. Echter, om het maximale uit de chatbot te halen, is het belangrijk om deze te besturen met zorgvuldig geformuleerde vragen en opdrachten.”</i></p>

ISSUE	OMSCHRIJVING
	<p>De rapportcijfers voor ChatGPT's waren als volgt onderverdeeld:</p> <p><i>Schrijfvaardigheid: 9</i></p> <p><i>Ideeëngenerator en ideeën-structureerder: 9,5</i></p> <p><i>Herformuleer-hulp van ChatGPT: 9,5.</i></p> <p><i>Taal- en spelvaardigheid: 8.</i></p> <p><i>Geheugen: 9.</i></p> <p><i>Doorvraag- en bijstuur- mogelijkheden van ChatGPT: 9</i></p> <p>Samenvattend, van het schrijven van sollicitatiebrieven tot het beschrijven van belangrijke thema's in een beroemd literair werk ChatGPT kan het zonder aarzelen. Dat roept de vraag op: <i>"Als ChatGPT voor ons kan schrijven, moeten studenten in de toekomst dan leren schrijven?"</i></p> <p>Het lijkt misschien een existentiële vraag, maar als studenten ChatGPT gaan gebruiken om hun essays te helpen schrijven, zullen scholen snel een antwoord moeten bedenken. De snelle acceptatie van Gen-AI in de afgelopen maanden zal veel sectoren tot nadenken stemmen, en het onderwijs is er daar één van.</p>
<p>4.</p> <p>Het kan schade in de echte wereld veroorzaken</p>	<p>ChatGPT kan schadelijk zijn voor mensen, met als duidelijkste voorbeeld verkeerd medisch advies. Er zijn ook andere problemen. Valse sociale media-accounts vormen een enorm probleem op het internet en met de introductie van AI-chatbots zou internetoplichting gemakkelijker uit te voeren zijn.</p> <p>De verspreiding van valse informatie is een andere zorg, vooral wanneer ChatGPT zelfs foute antwoorden overtuigend goed laat klinken. De snelheid waarmee ChatGPT antwoorden kan produceren die niet altijd correct zijn, heeft al problemen veroorzaakt voor Stack Exchange, een website waar gebruikers vragen kunnen plaatsen en antwoorden kunnen krijgen.</p> <p>Kort na de lancering werden antwoorden van ChatGPT van de site verbannen omdat een groot aantal ervan fout was. Zonder voldoende menselijke vrijwilligers om de achterstand te sorteren, is het onmogelijk om de kwaliteit van de antwoorden op een hoog peil te houden, waardoor de websiteschade oploopt.</p>

ISSUE	OMSCHRIJVING
5. OpenAI-LP / Microsoft heeft alle macht / het monopolie	<p><i>"With great power comes great responsibility!"</i>. OpenAI heeft veel macht omdat het nu in monopoliepositie verkeerd.</p> <p>Het heeft de AI-gemeenschap wereld opschudt met niet één, maar meerdere Gen-AI, waaronder Dall-E 2, GPT-3 en nu ChatGPT. OpenAI kiest welke gegevens er worden gebruikt om ChatGPT te trainen en hoe het omgaat met de negatieve gevolgen. Of we het nu eens zijn met de methoden of niet, het zal deze technologie blijven ontwikkelen volgens zijn eigen doelstellingen.</p> <p>Hoewel OpenAI beweerd dat het veiligheid hoog in het vaandel heeft staan, is er veel dat we niet weten over hoe de modellen tot stand komen. Of je nu vindt dat de code open source moet worden gemaakt, of dat delen ervan geheim moeten blijven, we kunnen er niet veel invloed op uitoefenen.</p> <p>Uiteindelijk kunnen we er alleen maar op vertrouwen dat OpenAI ChatGPT op verantwoorde wijze zal onderzoeken, ontwikkelen en gebruiken. Als alternatief kunnen we ervoor pleiten dat meer mensen inspraak krijgen in de richting waarin AI zich moet ontwikkelen, zodat de kracht van AI wordt gedeeld met de mensen die het zullen gebruiken.</p>

Geselecteerde referenties voor verder lezen

1. [IEEE Spectrum \[AI news item \(13 maart 2023\)\]: "Hallucinations Could Blunt ChatGPT's Success." "OpenAI says the problem's solvable, Yann LeCun says we'll see"](#)

[0c] MOET IK ME ZORGEN MAKEN OVER ChatGPT TECHNOLOGIE?

Gezien alle berichtgeving in korte tijd rondom Gen-AI —*ChatGPT in het bijzonder*— (zie [Wat is Generatieve Kunstmatige Intelligentie?](#)) is het begrijpelijk dat mensen verontrust & overweldigd zijn. Net zoals social media ons gedrag en cultuur sterk hebben beïnvloed zal Gen-AI een blijvende impact op ons doen en laten hebben. Op basis van uitspraken van “AI-experts” over de “ChatGPT Hype” volgen hier een paar uitgangspunten die kunnen helpen om de recente ontwikkeling te kunnen duiden en als startpunt kunnen dienen om Gen-AI verantwoord te kunnen gebruiken.

[1] Ten eerste “er is geen reden tot paniek” ondanks alarmerende uitspraken door AI-specialisten.

Timnit Gebru —AI-ethicus & Oprichter van Distributed Artificial Intelligence Research Institute (DAIR)— *Ik denk dat we echt doodsbang moeten zijn voor dit hele gebeuren. “Verondersteld wordt dat ChatGPT leerde schrijven door miljoenen geschriften op het internet te bestuderen.” Helaas, geloof het of niet, niet alles op het internet is waar! Het werd niet geleerd om te begrijpen wat feit is, wat fictie is, of iets dergelijks. Het papegaait gewoon terug wat er op het internet stond.*

Het is van belang om te weten dat Gen-AI —zoals ChatGPT en Bard— *taal agnostisch* zij. Door gebruikmaking van “*Machinaal lerende*” algoritmen gaat het veelvuldig in de fout met het interpreteren van zinsbouw, maar is nagenoeg foutloos in het vertalen van individuele woorden. Dergelijke systematische fouten weerspiegelen een fundamentele tekortkoming van GPT-technologie omdat het leert taalvaardig te worden op basis van reeksen aan woordvolgorde zoals die voorkomen in door mensen geschreven en/of gesproken teksten die zijn ontleend aan het world-wide-web. ChatGPT beschikt dus niet over linguïstische concepten en probleemoplossend vermogen dat nodig is om een taal te kunnen verwerken zoals mensen dat zouden doen. Sterker nog, een [invloedrijk paper](#) getiteld: “*On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?*” benadrukt dat GPT-technologie alle voorkomende overheersende opvattingen en vooroordelen die schadelijk zijn voor tal van minderheden, woord voor woord vastlegt precies zoals ze voorkomen op het world-wide-web “zonder aanzien des persoons”. Voorlopig lijkt het erop dat opschalen van het onderliggende taal-model ervoor moet zorgen dat de geconstateerde gebreken als sneeuw voor de zon zullen verdwijnen.

Dus reden tot zorg is er wel omdat Gen-AI in haar huidige vorm een “Work-in-Progress” is, dat nog veel te wensen overlaat. Zie: [“Is er een kort overzicht van ChatGPT’s](#)

tekortkomingen?” Maar dit laat onverlet dat het betrouwbaar kan worden ingezet voor specifieke functies waarbij het *“Talige & probleemoplossend vermogen”* —zoals het vertalen teksten en of verbeteren van spelfouten— een ondergeschikte rol speelt. Zie: *“Wat zijn de functionele mogelijkheden & beperkingen van ChatGPT?”*

[2] Ten tweede, Gen-AI beschikken niet of nauwelijks over *“creatief vermogen.”* Gen-AI zijn juist ontworpen om mensen te inspireren. Eric Mack —CNET Editor — formuleerde dit in zijn blog getiteld: *“Generative AI Tools Like ChatGPT and Dall-E Are Everywhere: What You Need to Know. The revolution will be generated by artificial intelligence. Perhaps.”*

Natuurlijk kunnen er manieren zijn om AI zodanig te manipuleren dat het zelf creatiever wordt, bijvoorbeeld door het te vragen specifiek nieuwe inhoud te genereren op basis van zwakkere associaties die het vindt in trainingsgegevens. Dit zou een manier kunnen zijn om menselijke creativiteit te simuleren met behulp van wiskunde en code. Maar iedereen die wel eens een creatieve doorbraak of een eureka moment onder de douche heeft gehad, zal je vertellen dat het vaak uit het niets lijkt te komen. We begrijpen onze creativiteit zelf nog niet, dus kunnen we die nog niet vertalen in een code die een machine kan begrijpen en proberen na te bootsen. En dan hebben we het nog niet eens over menselijke emoties, vele zintuiglijke ervaringen of veel van de basisfuncties van de hersenen die de wetenschap nog steeds niet goed begrijpt. Maar dit is nog maar het begin. Volgens sommigen sterven we in de komende tien of twee jaar af op kunstmatige algemene intelligentie - dat zou een systeem zijn dat echt dezelfde capaciteiten heeft als een mens op een niet te onderscheiden manier. Voor de goede orde: andere deskundigen denken dat dit nog lang niet zal gebeuren, als het al ooit gebeurt. Voorlopig is het het beste om vertrouwd te raken met deze systemen, hoe ze werken en wat ze wel en niet kunnen. Kennis is krachtiger dan informatie, zelfs terabytes ervan, en dat is een voordeel dat we allemaal nog steeds hebben ten opzichte van AI. Voorlopig althans.

[3] Ten derde, *“Hoe kan ChatGPT het volgende grote technologische stap zijn in AI als het een work-in-progress is?”*

De onderstaande tekst is deels gebaseerd op Tony Polano's *“Tom's Guide”* opinie stuk (20 februari 2023) getiteld: *“How can ChatGPT be the next big thing if it's this broken?”*

Het Wired IDEAS Blog (09 december 2022) getiteld: *“ChatGPT, Galactica, and the Progress Trap: When large language models fall short, the consequences can be serious. Why is it so hard to acknowledge that?”* legt een aantal fundamentele beperkingen bloot van

“taalvaardige-AI” zoals ChatGPT dat tot stand is gebracht met behulp van “Deep Learning” technieken. Een nagenoeg onoplosbaar probleem is dat ChatGPT *taal agnostisch* is. Door gebruikmaking van “*Machinaal lerende*” algoritmen gaat het veelvuldig in de fout met het interpreteren van zinsbouw, maar is nagenoeg foutloos in het vertalen van individuele woorden.

Dergelijke systematische fouten weerspiegelen een fundamentele tekortkoming van GPT-technologie omdat het leert taalvaardig te worden op basis van reeksen aan woordvolgorde zoals die voorkomen in door mensen geschreven en/of gesproken teksten die zijn ontleend aan het world-wide-web.

ChatGPT beschikt dus niet over linguïstische concepten en probleemoplossend vermogen dat nodig is om een taal te kunnen verwerken zoals mensen dat zouden doen. Sterker nog, een invloedrijk paper getiteld: “*On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?*” benadrukt dat GPT-technologie alle voorkomende hegemonische opvattingen en vooroordelen die schadelijk zijn voor tal van minderheden, woord voor woord vastlegt precies zoals ze voorkomen op het world-wide-web “*zonder aanzien des persoons*”.

ChatGPT wordt gevoed met de schoonheid, lelijkheid en wreedheid van het internet en sociale-media, de verwachting dat het ons alleen het goede, het schone en het behulpzame toont is een gevaarlijke en naive houding.

Voorlopig lijkt de houding van “*AI tech giants*” zoals Microsoft en Alphabet dat opschalen van het onderliggende taal-model ervoor moet zorgen dat de geconstateerde gebreken als sneeuw voor de zon zullen verdwijnen.

[4] Ten vierde, in een AIM blog door Mohit Pandey —Technologie journalist— getiteld: “*ChatGPT & Bing AI are Chit-chatting, Should We Be Worried? If chatbots can simulate conversations, will they generate languages that humans cannot understand?*” beschrijft hij de volgende “bizarre” conversaties tussen “taalvaardige” Gen-AI en hun vermogen voor het creëren van een “eigen-taal”:

Iemand heeft ChatGPT en Bing AI onlangs een gesprek laten voeren. Prompt na prompt leerden de chatbots over elkaar, en zijn nu beste maatjes! Ondertussen deed zich in 2017 een soortgelijk incident voor toen twee chatbots van Facebook in hun eigen taal met elkaar begonnen te praten, en moesten worden uitgeschakeld. Hetzelfde jaar beweerde Google dat zijn Translate tool de mogelijkheid had om zijn eigen taal te genereren. Ook OpenAI-LP beweert dat AI inderdaad kan worden aangemoedigd zijn eigen taal te creëren. Deze anekdotische beschrijvingen doet de vraag rijzen: “Zijn Gen-AI echt in staat om hun eigen taal te creëren iets waar mensen te dom voor zijn om te begrijpen?”

Een vergelijkbare situatie bestond voor de ontwikkeling van het Internet-of-Things” [IoT] en haar industriële variant Cyber Physical Systems [CPS] zo rond 2010, waarbij we hebben toegestaan dat “dingen” onderling met elkaar kunnen communiceren zonder tussenkomst van de mens — *human-in-the-Loop*—.

We zijn nu op een punt in de ontwikkelingsgeschiedenis van AI-technologie beland waar de “de mens” nog kan bepalen of we deze weg willen bewandelen of dat we “paal en perk” gaan stellen aan de handelingsvrijheid van Gen-AI.

Gevolg is het ontstaan van de zogenaamde “Data-deluge”: De toename van de hoeveelheid digitaal beschikbare ongestructureerde/ruwe data overstijgt de totale hoeveelheid aan beschikbare “computer” rekenkracht. Sinds 2019 wordt wereldwijd meer digitale data geproduceerd dan analoge data. Daarmee is de mensheid in 2023 met de komst van ChatGPT defacto een “*Society of Algorithms*” geworden.

De vraag is dan ook; “Kunnen Gen-AI’s als tool dienen om de “Data-Deluge” paradox te doorbreken?”

[5] Ten vijfde, in een opiniestuk (07 maart 2023) in de New York Times —geschreven door Noam Chomsky, Ian Roberts and Jeffrey Watumull— getiteld: “The False Promise of ChatGPT” wordt de volgende waarschuwing —door de ethiek ingegeven— geformuleerd:

Jorge Luis Borges —een van de meest invloedrijke schrijvers van de 21ste eeuw— schreef: “... wij leven in een tijd van groot gevaar en grote beloftes zowel een tragedie als een komedie is, met”de nabijheid van een openbaring” om onszelf en de wereld te begrijpen. Onze zogenaamde revolutionaire vooruitgang op het gebied van AI geeft inderdaad aanleiding tot zowel bezorgdheid als optimisme. Optimisme omdat intelligentie het middel is waarmee we problemen oplossen. Bezorgdheid omdat we vrezen dat de meest populaire en modieuze vorm van A.I. - machine learning - onze wetenschap zal aantasten en onze ethiek zal ontkrachten door in onze technologie een fundamenteel onjuiste opvatting van taal en kennis op te nemen.

De schrijvers van dit opiniestuk zijn van mening dat menselijke intelligentie —*dit in tegenstelling tot Gen-AI*— instaat is tot moreel denken en handelen —*ethisch handelen*—. Dit betekent dat we de anders zo grenzeloze creativiteit van onze geest onderwerpen aan ethische principes die bepalen wat wel en wat niet aanvaardbaar is om te delen met anderen. Om betekenisvol en aanvaardbaar te zijn voor een breed publiek moet een AI, zoals ChatGPT, het genereren van moreel verwerpelijke inhoud onderdrukken. Bij gebrek aan een vermogen om vanuit morele principes te “redeneren”, werd ChatGPT van bovenaf beperkt in het genereren “controversiële” content. Daardoor wordt defacto het “creatieve” vermogen van ChatGPT gemuilkorfd als “oplossing” van dit “amoraliteit probleem”.

[0d] IS ER EEN ChatGPT “CODE-OF-CONDUCT” en/of RICHTLIJN VOOR HBO DOCENTEN?

Overzicht beleid & richtlijnen ten aanzien van het inzetten van Gen-AI in het hoger onderwijs

(onderwijs)instelling	omschrijving	URL
Radboud Universiteit [RU]	Chat GPT: what does this mean for you as a student?	https://www.ru.nl/en/students/news/chat-gpt-what-does-this-mean-for-you-as-a-student https://www.voxweb.nl/nieuws/wat-moet-de-radboud-universiteit-met-tekstrobot-chatgpt
Universiteit Leiden [UL]	Gebruik jij ChatGPT voor schrijfp opdrachten? Let op de risico's !	https://www.student.universiteit leiden.nl/mededelingen/2023/02/gebruik-jij-chatgpt-voor-schrijfp opdrachten-let-op-de-risicos
Avans	Opinie: Schrijvende chatbot bedreigt intellectuele onafhankelijkheid	https://www.avans.nl/over-avans/nieuws-en-pers/nieuwsberichten/detail/2023/02/opinie-schrijvende-chatbot-bedreigt-intellectuele-onafhankelijkheid https://punt.avans.nl/2023/01/chatgpt-in-het-onderwijs-voor-studenten-is-het-een-leerproces/
Hogeschool van Amsterdam [HvA]	HvA omarmt ChatGPT met aandacht voor kansen én risico's	https://www.hva.nl/appliedai/chat/gpt.html https://www.hva.nl/content/nieuws/nieuwsberichten/2023/01/hoe-moeten-docenten-in-het-hbo-omgaan-met-chatgpt.html
Hogeschool Arnhem & Nijmegen [HAN]	Handreiking ChatGPT & Toetsing	https://www.han.nl/artikelen/2023/01/het-onderwijs-en-chatgpt/ https://www.han.nl/onderwijsondersteuning/leren-werken-met-ict/artificial-intelligence/HAN-Handreiking-ChatGPT-en-toetsing.pdf

(onderwijs)instelling	omschrijving	URL
Hogeschool Rotterdam [HR]	ChatGPT: vriend of vijand?	https://www.hogeschoolrotterdam.nl/hogeschool/nieuws/chatgpt-vriend-of-vijand/ https://github.com/HR-ChatGPT/ChatGPT-UITGELEGD https://profielen.hr.nl/2023/chatgpt-brengt-ook-hogeschool-rotterdam-in-rep-en-roer-geniaal-en-lastig/ Sinds begin maart 2023 is er een handreiking beschikbaar in de vorm van een stroomschema: " <i>ChatGPT en toetsing - kaders voor gebruik</i> " via HINT (het intranet van de hogeschool Rotterdam)
Hogeschool Utrecht [HU]	ChatGPT: Een vloek of zegen?	https://husite.nl/digitalehu/chatgpt-een-vloek-of-zegen/ https://husite.nl/digitalehu/wp-content/uploads/sites/244/2023/01/ChatGPT-handreiking.pdf https://husite.nl/toetsing-nieuw/handreiking-chatgpt-en-toetsing/
Saxion	Leren in tijden van ChatGPT	https://www.saxion.nl/nieuws/2023/02/leren-in-tijden-van-chatgpt-een-debat-over-de-impact-van-ai
Windesheim	ChatGPT en hoger onderwijs	https://www.researchgate.net/publication/368470618_ChatGPT_en_hoger_onderwijs
Universiteit Utrecht [UU]	Gebruik van AI voor schrijfpoddrachten	https://www.uu.nl/onderwijs/onderwijsadvies-training/publicaties/tips-voor-leerkrachten-en-docenten/geef-studenten-korte-schrijfpoddrachten https://www.uu.nl/in-de-media/geesteswetenschappers-over-chatgpt-en-de-morele-kwesties-van-ai https://www.uu.nl/onderwijs/onderwijsadvies-training/kennisdossiers/kennisdossier-hoger-onderwijs/waar-is-chatgpt-toe-in-staat-en-wat-zijn-de-beperkingen

(onderwijs)instelling	omschrijving	URL
Universiteit Groningen [UMCG]	BloG dat de UMCG heeft opgezet om studenten te informeren over ChatGPT	https://umcgresearch.org/w/chatgpt
Neerlandistiek Online tijdschrift voor de Nederlandse taalkunde, letterkunde en taalbeheersing	ChatGPT: de rapportcijfers	https://neerlandistiek.nl/2023/01/chatgpt-de-rapportcijfers/
Vrije Universiteit [VU]	Hoe ga je als docent om met ChatGPT?	https://vu.nl/nl/medewerker/didactiek/hoe-ga-je-als-docent-om-met-chatgpt
Universiteit van Amsterdam [UvA]	ChatGPT in de Wetenschap	https://www.uva.nl/content/nieuws/nieuwsberichten/2023/02/chatgpt-in-de-wetenschap-5-aandachtspunten.html
Technische Universiteit Eindhoven [TU/e]	Wat betekent Chatbot GPT voor het onderwijs?	https://www.cursor.tue.nl/achtergrond/2023/januari/week-3/wat-betekent-chatbot-gpt-voor-het-onderwijs/
(Technische) Universiteit Twente [UT]	Ontwikkelingen rond ChatGPT	https://www.utwente.nl/onderwijs/student-services/actueel/nieuws/2023/2/463644/ontwikkelingen-rond-chatgpt https://www.utwente.nl/en/learning-teaching/ https://www.utoday.nl/news/72264/chatrobot-over-ai-in-het-hoger-onderwijs-kan-leiden-tot-plagiat https://www.utoday.nl/news/72477/ut-onderzoekt-plagiatregeling-door-komst-chatgpt
(Technische) Universiteit Delft [TUDelft]	AI chatbots in projects and assignments	https://www.tudelft.nl/teaching-support/didactics/assess/guidelines/ai-chatbots-in-projects-and-assignments https://www.tudelft.nl/teaching-support/didactics/assess/guidelines/values-quality-requirements

(onderwijs)instelling	omschrijving	URL
Universiteit Maastricht [UM]	Ik heb dit helemaal zelf geschreven!	https://www.maastrichtuniversity.nl/nl/nieuws/ik-heb-dit-helemaal-zelf-geschreven
Curio	Tekstrobot (ChatGPT)	https://lerenbij.curio.nl/chatgpt/chatgpt/ https://lerenbij.curio.nl/bijlagen/chatgpt-en-onderwijs/
Kennisnet	FAQ ChatGPT: veelgestelde vragen over ChatGPT in het onderwijs	https://www.kennisnet.nl/faq-chatgpt-veelgestelde-vragen-over-chatgpt-in-het-onderwijs/
SURF	Onderwijsexperts discussiëren over ChatGPT: "Er is een extra klasgenoot bij gekomen"	https://communities.surf.nl/ai-in-education/artikel/onderwijsexperts-discussieren-over-chatgpt-er-is-een-extra-klasgenoot-bij https://communities.surf.nl/vraagbaak-online-onderwijs/artikel/hoe-chatgpt-jouw-werk-als-docent-makkelijker-maakt

[0e] KAN ChatGPT BENUT WORDEN ALS BEOORDELINGSINSTRUMENT?

Beoordelen vereist redeneervermogen en creativiteit

Een invloedrijk arXiv paper uit 2021 getiteld: *“On the Opportunities and Risks of Foundation Models”* benadrukt dat AI een paradigmaverschuiving ondergaat vergelijkbaar met die van *“deep Learning models”* in 2010. Anno 2023, toveren *“Few-Shot Learners”* ons voor wat er mogelijk is. Je voert een korte omschrijving plus vraag in, vervolgens genereert — *Bart(LaMDA), BERT, Bing, BLOOM, ChatGPT(instructGPT), Galactica, Sparrow, LLaMA, Med-PaLM, CLIP, DALL-E 2, Midjourney of OPT*— een heel opstel of een complexe afbeelding op basis van jouw omschrijving, zelfs als het niet specifiek getraind is op het uitvoeren van dat exacte vraag of het genereren van een afbeelding op die manier. Het gebruik van *“talige”* LLMs in het onderwijsdomein is daarom bijzonder beladen.

Hoe indrukwekkend ze ook zijn, state-of-the-art LLM's blijven gevoelig voor *“onbedoelde”* fouten. De observatie dat dergelijke Gen-AI aanzienlijk verbeteren naarmate het aantal parameters en de omvang van de training corpora worden opgeschaald, heeft sommigen in het veld doen beweren dat LLM —*misschien in een multimodale versie*— zal leiden tot intelligentie en begrip op menselijk niveau, bij voldoende grote netwerken en training datasets. Er lijkt sprake te zijn van een nieuw AI-mantra: *“Schaal is alles wat je nodig hebt.”*

“Taalvaardige” op LLMs gebaseerde chatbots worden getest op basis van maatstaven (benchmarks) zoals: |benchmark| |:—| | *“algemeen taalbegrip”* | *“natuurlijke taal inferentie”* | *“begrijpend lezen”* | *“gezond verstand redeneren”* | *“probleem oplossend vermogen”* | *“Multi-Hop redeneren”*

De premisse onderliggend aan deze benchmarks is dat een vorm van *“taal begrip”* in combinatie met *“probleem oplossend vermogen”* een vereiste is om goed te presteren op *“Talige”* taken. Maar is deze aanname wel correct? Niet noodzakelijkerwijs! Om te begrijpen waarom, moeten we inzicht krijgen in wat er gebeurt als we een tekstuele opdracht aan een LLM-chatbot toewijzen. Daarvoor is een korte introductie over redeneren en argumenteren nodig.

De verbindende factor tussen de verschillende benchmarks is dat ze allemaal een zekere mate aan *“redeneervermogen”* vereisen. Maar wat is redeneren nu eigenlijk? Voor een inzichtelijk raamwerk met betrekking tot de relevantie van *“logisch Redeneren”* als

didactisch instrument in het hoger onderwijs verwijs ik naar het paper getiteld: *“Logical Reasoning in Formal and Everyday Reasoning Tasks”*.

Bij een geschreven tekst —*in de vorm van een betoog*— probeert de schrijver ervan, lezers te overtuigen door een *“logische gedachtegang”*, een *“redenering”* op te bouwen. Door te redeneren kun je —*bij voldoende betrouwbare informatie*— tot verifieerbare oordeelsvorming komen.

Redeneren is het proces van het opbouwen van een argumentatie

Argumentatie is een verbale activiteit die erop gericht is een redelijke beoordelaar te overtuigen van de aanvaardbaarheid van een standpunt door één of meerdere beweringen naar voren te brengen als rechtvaardiging voor het ingenomen standpunt.

Argumenteren heeft tot doel de ander te overtuigen van een standpunt

Een argument wordt bepaald door het beoogde doel: *Controleerbare feiten, vergelijking (analogie), ervaring (empirisch), gezag of autoriteit, gevolg, nut of gewenste gevolgen, gevoel of emotie, algemene normen en waarden, veronderstelling*. Argumenten worden dan ook gebruikt om een standpunt of mening te onderbouwen.

In het AI-domein gericht op natuurlijke taal verwerking [NLP] werd tot zeer recent *“redeneren”* over meerdere bronnen —*zoals wetenschappelijke artikelen, Wikipedia lemma’s of boeken*— als een nagenoeg onoplosbaar probleem —*“hard problem”*— beschouwd.

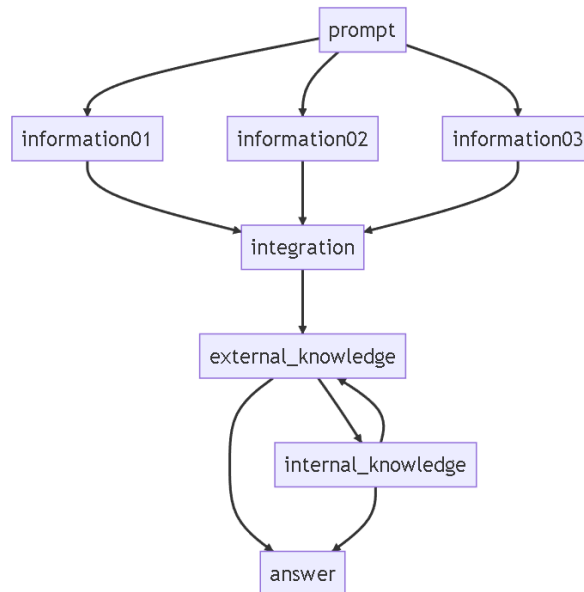
Wat gebeurt er als je een tekstuele opdracht/prompt aan een LLM toewijst?

Opdrachten geven aan een LLM om een tekst te genereren, is hetzelfde als het geven van een prompt aan een *“denkbeeldige notulist”*. Instructies voor het schrijven van effectieve opdrachten zijn beschreven in: [Prompt “patterns” voorbeelden](#).

We onderscheiden drie soorten prompts (zie ook [Kan ChatGPT broncode schrijven?](#)):

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

Keten van 3 'gedachten'



“Chain-of-Thought” —keten van gedachten— is een vorm van Logisch Redeneren

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 het LLM al in zich herbergt.

Een **chain-of-thoughts** [CoTs] kan worden beschouwd als een vorm van *Logisch Redeneren*: het proces van het opbouwen van een argumentatie.

Het mechanisme van een keten-van-gedachten is een vorm van “prompt fine-tuning”.

Hoe effectief is “keten-van-gedachten” —Chain-of-Thoughts [CoT]—?

Onderstaande voorbeelden zijn ontleend aan [The Decoder nieuwsbrief \(27 september 2022\)](#), geschreven door Moritz Larsen & Doris Weßels (Universiteit Kiel) getiteld: “*What is CoT Prompting and how can it help?*”. Hiebij is gebruik gemaakt van GPT-3.

Voorbeeld 1: Effect van CoT prompting op rekentaken

Rekenopdracht zonder CoT	Rekenopdracht met CoT
Wat is de jaarlijkse waterbehoefte van een eengezinswoning met vier personen die gemiddeld 200 dagen per jaar thuis zijn en gemiddeld 100 liter water per dag gebruiken?	Wat is de jaarlijkse waterbehoefte van een eengezinswoning met vier personen die gemiddeld 200 dagen per jaar thuis zijn en gemiddeld 100 liter water per dag gebruiken? Laten we stap voor stap nadenken en de berekening uitleggen.
Resultaat zonder CoT	Resultaat met CoT
De jaarlijkse waterbehoefte van een eengezinswoning waarin vier personen wonen, die gemiddeld 200 dagen per jaar thuis zijn en gemiddeld 100 liter water per dag nodig hebben, is 80.000 liter.	Aantal personen in het huishouden: 4 personen Gemiddeld aantal dagen dat de mensen in het huishouden thuis zijn: 200 dagen Gemiddeld waterverbruik per persoon per dag: 100 liter $4 \text{ personen} * 200 \text{ dagen} * 100 \text{ liter water} = 80.000 \text{ liter water per jaar}$

Dit rekenvoorbeeld toont aan dat reeds de eerste stappen naar verklaarbaarheid kunnen worden gezet door een intelligent prompt ontwerp. De uitlegbaarheid heeft echter alleen betrekking op de gevolde redentatie, niet op de innerlijke structuur van het onderliggende taalmodel —GPT-3—.

Voorbeeld 2: Verbetert CoT prompting de uitleg van een citaat?

Kan GPT-3 een citaat verklaren. Het onderstaande citaat is een passage uit: *“Artificial Intelligence and Digital Literature. Theory and Practice of Connectionist Writing”*, geschreven door Hannes Bajohr getiteld.

“The rupture between paradigms also touches on the authorship question, the human side of which is undergoing an increasing distancing in the human-machine structure. While in the sequential paradigm one could still plausibly speak of secondary authorship, which consists in the formulation of a sequence of rules, the execution of which produces the work – here the idea of a “writer of writers” makes perfect sense – with KNNs one is faced with tertiary authorship: Only the data set for the training remains to be defined, from which the KNN independently forms the model, and the parameters remain to be determined, by means of which the model finally produces the output. With large language AIs like GPT-3, even this is no longer possible, because the training is too complex to adapt it to new data sets. The ‘programming’ is done by the normal language formulation of prompts (‘prompt design’) following the model of dialogic communication – here one could even speak of quaternary authorship.” (Bajohr 2022)

Opdracht zonder CoT	Opdracht met CoT
Wat beschrijft het bovenstaande citaat van Hannes Bajohr?	Wat beschrijft het bovenstaande citaat van Hannes Bajohr? Geef inzicht —stap voor stap— hoe je tot de uiteindelijke conclusie komt.
Resultaat zonder CoT	Resultaat met CoT
Dit citaat van Hannes Bajohr beschrijft de breuk tussen de sequentiële en KNN-paradigma's. In het sequentiële paradigma is het mogelijk om te spreken van secundair auteurschap, dat bestaat uit de formulering van een reeks regels waarvan de uitvoering het werk produceert. In KNN's is het echter alleen mogelijk om de parameters op te geven waarmee het model de output produceert. In het geval van grote taal-AI's zoals GPT-3 is zelfs dit niet meer mogelijk.	Eerst wordt verondersteld dat er in het sequentiële paradigma sprake is van secundair auteurschap [sic!], bestaande uit een reeks regels waarvan de uitvoering het werk produceert. Dit idee van een 'schrijver van schrijver' heeft in deze context zin. Het volgende punt is om erop te wijzen dat er bij KNN's (kunstmatige neurale netwerken) sprake is van tertiair auteurschap. Dit betekent dat het gegeven set voor training en de parameters waarmee het model uiteindelijk de output produceert, worden gemaakt door de KNN zelf. In de laatste stap wordt vermeld dat er voor grote taal-AI's zoals GPT-3 zelfs sprake is van een quaternair auteurschap. Dit betekent dat de training zo uitgebreid is dat het niet langer mogelijk is om deze aan te passen aan steeds nieuwe datasets. Het 'programmeren' gebeurt hier door de normale taalformulering van prompts ('prompt design') volgens het model van dialogische communicatie. Dit citaat beschrijft dus het feit dat het auteurschap van werken is veranderd van secundair naar tertiair en quaternair auteurschap naarmate kunstmatige intelligentie zich heeft ontwikkeld. Dit citaat beschrijft dus het feit dat het auteurschap van werken is veranderd van secundair naar tertiair en quaternair auteurschap naarmate kunstmatige intelligentie zich heeft ontwikkeld.

Wat meteen opvalt is dat het CoT+ antwoord veel uitgebreider is. Er is een structuur met opsommingen te zien. Ook de regeleinden na elke alinea, die naast de tekst door het taalmodel werden gegenereerd, dragen bij tot deze structurering. De structurering wekt dus de indruk dat de afzonderlijke aspecten van het citaat worden opgenomen en dat daaruit in de laatste stap een conclusie wordt getrokken.

Hier wordt de stapsgewijze procedure duidelijk. De kern van het citaat wordt herkend en in drie secties verdeeld, en telkens wordt vermeld welke vorm van auteurschap ermee samenhangt. In de tweede sectie staat echter een inhoudelijke fout. Er wordt gezegd dat de dataset en de parameters onafhankelijk door het KNN worden aangemaakt. Volgens Bajohr is dat echter juist niet het geval, maar worden de elementen door mensen bepaald.

Weet ChatGPT hoe “wij” mensen de wereld ervaren?

De onderstaande tekst is deels gebaseerd op een blog post getiteld: *“ChatGPT understands language or if it doesn’t, it fails for reasons other than the ones you think”*.

Om te kunnen redeneren zoals mensen dat doen is het noodzakelijk om toegang te hebben tot een *“model van hoe de fysieke wereld om ons heen in elkaar steekt”*.

Kunnen mensen een boom optillen?

'Nee, normaal gesproken kunnen mensen geen boom optillen.

Bomen zijn meestal zwaar en groot van formaat,
dus het is onmogelijk om ze zonder hulp op te tillen.'

Wat bedoel je met optillen?

'Als je optillen bedoelt als het verplaatsen van de boom,
dan kan het antwoord Nee zijn.

De meeste bomen zijn veel te groot en zwaar om met de hand opgetild te worden.
In de meeste gevallen heb je uitrusting nodig zoals takels,
vrachtwagens of kranen om bomen te verplaatsen.'

Wat bedoel je met verplaatsen?

'Met verplaatsen bedoel ik dat de boom van een bepaalde locatie
naar een andere locatie zou worden gebracht,
verplaatst zou zijn of in een nieuwe locatie zou worden geplant.'

Wat is de gemiddelde hoogte van een boom in Nederland in meters?

'Het is moeilijk om een precies antwoord te geven op
de gemiddelde hoogte van bomen in Nederland, omdat het afhangt
van verschillende factoren zoals boomsoort en groeiomstandigheden.
Over het algemeen worden bomen in Nederland echter niet hoger dan 40-50 meter.'

Uit de bovenstaande dialoog met ChatGPT valt af te leiden dat deze AI beschikt over *“een model van de wereld”* als het gaat om vragen die betrekking hebben op alledaagse situaties.

Maar wat gebeurt er als we een zeer specifiek, technisch en zeldzaam probleem laten beschrijven door chatGPT?

Wat kun je mij vertellen over dubbele magnetische inductie om oogbewegingen bij mensen te kunnen meten?

'Dubbele magnetische inductie (ook bekend als Double Magnetic Induction, DMI) is een technologie die wordt gebruikt om oogbewegingen bij mensen te meten. Dit wordt vaak gebruikt in de neurowetenschappen en klinische studies om informatie te verzamelen over zaken als visuele perceptie en cognitie. DMI maakt gebruik van een paar magneten die rondom een proefpersoon worden geplaatst, samen met een set spoelen die zich voor en op het oog bevinden. Het werkt door het meten van de veranderingen in magnetische velden die optreden terwijl het oog beweegt. Wanneer het oog beweegt, heeft dit invloed op het magnetische veld dat rondom het oog wordt gegenereerd. Deze veranderingen kunnen door de spoelarray worden waargenomen en vervolgens worden vertaald naar gegevens over de oogbeweging. Een belangrijk voordeel van DMI is dat het zeer nauwkeurig is bij het meten van oogbewegingen, en kan helpen onderscheid te maken tussen verschillende soorten oogbewegingen zoals foveale fixatie, smooth pursuit en saccadische bewegingen. Dit type technologie wordt vaak gebruikt in onderzoek naar neurologische aandoeningen, evenals in de ontwikkeling van ooggestuurde systemen en interactieve toepassingen.'

De bovenstaande beschrijving is op hoofdlijnen correct. Dus wanneer het gaat om zeer 'zeldzame' technische concepten dan kan ChatGPT ook hierop adequaat reageren.

Hoe effectief is ChatGPT in het beoordelen van een redentie?

Nu we hebben vastgesteld dat ChatGPT een model van de wereld heeft, is de volgende vraag: "In hoeverre kan ChatGPT gebruikt worden om een stuk tekst te beoordelen?" Dat wil zeggen is het in staat om een "keten-van-gedachten" te volgen en de juistheid ervan vast te stellen?

Is de volgende bewering correct:

"Met behulp van neurale netwerktheorie kan een DMI-systeem worden gekalibreerd.

Het enige dat nodig is, is dan een proefpersoon die rond kijkt terwijl de spanningsveranderingen

in van de inductie-sensor op het oog worden geregistreerd door een computer.

Deze signalen worden dan gebruikt als input dataset voor het te trainen neurale netwerk."

'Ja, deze bewering is correct.

Het kalibreren van een DMI-systeem met behulp van neurale netwerk theorie is een veelgebruikte methode. Tijdens de kalibratie zal de persoon naar verschillende punten op een scherm kijken, terwijl de sensoren de bijbehorende signalen registreren. Deze gegevens worden dan gebruikt om het neurale netwerk te trainen en aan te passen, zodat het de oogbewegingen van de persoon nauwkeurig kan voorspellen en volgen. Dit proces kan worden herhaald totdat het systeem nauwkeurige en betrouwbare resultaten oplevert.'

Is de volgende bewering correct:

"Een wetenschapper gebruikt een Large language Model (LLM) als AI-technologie om een DMI-systeem te kalibreren. De uitkomsten van de kalibratie geven betrouwbare oogbeweging metingen"

'Het is onwaarschijnlijk dat een Large Language Model (LLM) nuttig zal zijn bij het kalibreren van een Double Magnetic Induction (DMI) systeem voor het meten van oogbewegingen. LLMs worden meestal gebruikt voor taalkundige taken, zoals tekstgeneratie en vertaling. Er zijn echter andere machine learning-technieken die wellicht nuttiger kunnen zijn bij het kalibreren van een DMI-systeem, zoals neurale netwerken of Kohonen-netwerken. Het uiteindelijke resultaat van de kalibratie zal afhangen van de nauwkeurigheid van de gebruikte techniek en de correcte toepassing ervan. Als de kalibratie goed wordt uitgevoerd en het systeem correct is ingesteld, kunnen betrouwbare oogbewegingsmetingen worden verkregen.'

De antwoorden van ChatGPT zijn correct, inclusief de uitleg. Dit eenvoudige experiment met ChatGPT toont aan dat deze AI in staat is om stukken tekst —*die een duidelijk keten-van-gedachten beschrijven*— op hun juistheid te beoordelen en/of te verklaren.

Er is echter een bijkomend probleem van hoe ChatGPT om gaat met prompts. Bij herhaling van exact dezelfde prompt, kan de uitkomst afwijken.

Is de volgende bewering correct:

"Een wetenschapper gebruikt een Large language Model (LLM) als AI-technologie om een DMI-systeem te kalibreren. De uitkomsten van de kalibratie geven betrouwbare oogbeweging metingen"

'Het is mogelijk dat een wetenschapper het LLM gebruikt om te helpen bij de kalibratie van een DMI-systeem, maar het is niet de enige of noodzakelijke manier om dit te doen. Andere machine learning-modellen zoals Kohonen-netwerken en Boltzmann-machines kunnen ook worden gebruikt voor kalibratie. De keuze tussen verschillende soorten machine learning-modellen hangt af van factoren zoals de beschikbaarheid van gegevens, de gewenste precisie, de complexiteit van het oogbewegingsproces.'

Het antwoord in dit geval is onjuist.

Op basis van dit eenvoudige experiment valt op te maken dat ChatGPT benut kan worden om als beoordelingsinstrument, maar het is nog niet betrouwbaar “genoeg”.

De volgende stap is dan om na te gaan of ChatGPT stukken tekst kan beoordelen op basis van een toetsingskader in combinatie met een RUBRICs.

Geselecteerde referenties voor verder lezen

1. Bronkhorst, H., Roorda, G., Suhre, C., & Goedhart, M. (2020). Logical reasoning in formal and everyday reasoning tasks. *International Journal of Science and Mathematics Education*, 18, 1673-1694. <https://doi.org/10.1007/s10763-019-10039-8>
2. Chain-of-Thought Prompting Poster <https://neurips.cc/virtual/2022/poster/54087>
3. [The Decoder nieuwsbrief \(27 september 2022\)](#) “What is CoT Prompting and how can it help?”
4. Gao, J., Galley, M., & Li, L. (2018). Neural approaches to conversational AI. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval* (pp. 1371-1374). <https://doi.org/10.1145/3209978.3210183>
5. Gao, J., Galley, M., & Li, L. (2019). Neural approaches to conversational AI: Question answering, task-oriented dialogues and social chatbots. *Now Foundations and Trends*. <https://doi.org/10.1561/15000000074>
6. Khot, T., Trivedi, H., Finlayson, M., Fu, Y., Richardson, K., Clark, P., & Sabharwal, A. (2022). Decomposed prompting: A modular approach for solving complex tasks. *arXiv preprint* <https://doi.org/10.48550/arXiv.2210.02406>
7. Lester, B., Al-Rfou, R., & Constant, N. (2021). The power of scale for parameter-efficient prompt tuning. *arXiv preprint* <https://doi.org/10.48550/arXiv.2104.08691>
8. Liu, H., Tam, D., Muqeeth, M., Mohta, J., Huang, T., Bansal, M., & Raffel, C. (2022). Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning. *arXiv preprint* <https://doi.org/10.48550/arXiv.2205.05638>
9. Liu, X., Zheng, Y., Du, Z., Ding, M., Qian, Y., Yang, Z., & Tang, J. (2021). GPT understands, too. *arXiv preprint* <https://doi.org/10.48550/arXiv.2103.10385>
10. Lyu, Q., Havaladar, S., Stein, A., Zhang, L., Rao, D., Wong, E., ... & Callison-Burch, C. (2023). Faithful Chain-of-Thought Reasoning. *arXiv preprint* <https://doi.org/10.48550/arXiv.2301.13379>
11. Min, S., Wallace, E., Singh, S., Gardner, M., Hajishirzi, H., & Zettlemoyer, L. (2019). Compositional questions do not necessitate multi-hop reasoning. *arXiv preprint* <https://doi.org/10.48550/arXiv.1906.02900>
12. Sun, S., Liu, Y., Iter, D., Zhu, C., & Iyyer, M. (2023). How Does In-Context Learning Help Prompt Tuning?. *arXiv preprint* <https://doi.org/10.48550/arXiv.2302.11521>

- 13.** Wei, J., Wang, X., Schuurmans, D., Bosma, M., Chi, E., Le, Q., & Zhou, D. (2023). Chain of thought prompting elicits reasoning in large language models. arXiv preprint <https://doi.org/10.48550/arXiv.2201.11903>. Original paper published at NeurIPS 2022 https://openreview.net/forum?id=_VjQlMeSB_J
 - 14.** White, J., Fu, Q., Hays, S., Sandborn, M., Olea, C., Gilbert, H., ... & Schmidt, D. C. (2023). A Prompt Pattern Catalog to Enhance Prompt Engineering with ChatGPT. arXiv preprint <https://doi.org/10.48550/arXiv.2302.11382>
 - 15.** Goyal, A., & Bengio, Y. (2022). Inductive biases for deep learning of higher-level cognition. *Proceedings of the Royal Society A*, 478(2266), 20210068. <http://doi.org/10.1098/rspa.2021.0068>
 - 16.** Greco, S., van Eemeren F.H., & A.F. Snoeck Henkemans: *Argumentation: Analysis and Evaluation*. *Argumentation* 32, 151–153 (2018). <https://doi.org/10.1007/s10503-017-9433-y>
 - 17.** Mihalcea, R., & Tarau, P. (2004, July). Textrank: Bringing order into text. In *Proceedings of the 2004 conference on empirical methods in natural language processing* (pp. 404-411). Association for Computational Linguistics. <https://aclanthology.org/W04-3252/>
 - 18.** Minaee, S., Kalchbrenner, N., Cambria, E., Nikzad, N., Chenaghlu, M., & Gao, J. (2021). Deep learning–based text classification: a comprehensive review. *ACM computing surveys (CSUR)*, 54(3), 1-40. <https://doi.org/10.1145/3439726>
 - 19.** Mitchell, M., & Krakauer, D. C. (2022). The Debate Over Understanding in AI's Large Language Models. arXiv preprint <https://doi.org/10.48550/arXiv.2210.13966>
 - 20.** Sejnowski, T. J. (2020). The unreasonable effectiveness of deep learning in artificial intelligence. *Proceedings of the National Academy of Sciences*, 117(48), 30033-30038. <https://doi.org/doi:10.1073/pnas.1907373117>
 - 21.** Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., Chi, E. H., Hashimoto, T., Vinyals, O., Liang, P., Dean, J., & Fedus, W. (2022). Emergent Abilities of Large Language Models. *Transactions on Machine Learning Research*, 2835-8856. <https://doi.org/10.48550/arXiv.2206.07682>
-

[1f] VOLDOEN Gen-AI —zoals ChatGPT— AAN DE EUROPESE AI-REGELGEVING?

De onderstaande tekst is gebaseerd op een news-item (03 maart 2023) afkomstig uit “Politico”, een gerenommeerde, onafhankelijke, Engelstalige krant, getiteld : *“ChatGPT broke the EU plan to regulate AI”* en het Center for Data Innovation blog (13 februari 2023) van Patrick Grady, getiteld: *“ChatGPT Amendment Shows the EU is Regulating by Outrage”*.

“Europe’s original plan to bring AI under control is no match for the technology’s new, shiny chatbot application.”

Op 12 april 2021 heeft de Europese Commissie een voorstel gedaan voor een “Artificial Intelligence Act”, om zo betrouwbare kunstmatige intelligentie te garanderen.

European AI Act De “Artificial Intelligence Act (AI-Act)” is een voorstel voor regulatie die tot doel heeft een gemeenschappelijk regelgevend en juridisch kader voor AI in te voeren. Het toepassingsgebied omvat alle sectoren (behalve militaire) en alle soorten kunstmatige intelligentie. Als productregelgeving verleent het voorstel geen rechten aan personen, maar ziet het toe op de arbitrage van aanbieders van AI-diensten en entiteiten die er beroepshalve gebruik van maken.

ChatGPT heeft recente Europese —de Europese Commissie, het Europees Parlement en de Raad van Europa— inspanningen om AI te reguleren, achterhaald. De verordening, die eind 2021 door de EU-Commissie werd voorgesteld, was bedoeld om bepaalde AI-toepassingen zoals sociale scoring, manipulatie en sommige gevallen van gezichtsherkenning onder toezicht te stellen of zelfs te verbieden. Uitgangspunt is om specifieke AI-technologie —zoals chatbots— aan te merken met het label “hoog risico”. Hierdoor worden ontwikkelaars onderworpen aan zeer strengere eisen op het gebied van transparantie, veiligheid en menselijk toezicht.

Het addertje onder het gras? ChatGPT, is een vorm van “Talige” Generatieve-AI. Het kan worden benut voor zowel *positive* alsook *negatieve* doeleinden: *“Mensen kunnen er liedjes, romans en gedichten mee schrijven, maar ook computercode, beleidsnota’s, nepnieuws-berichten of, zoals een Colombiaanse rechter heeft toegegeven, gerechtelijke uitspraken.”*

Wanneer er terechte zorgen bestaan over het gebruik van Gen-AI —zoals de verspreiding van verkeerde informatie of schadelijke inhoud— moeten wetgevers deze risico’s aanpakken in sectorale wetgeving —zoals de wet op de digitale diensten, die platforms en zoekmachines verplicht om verkeerde informatie en schadelijke inhoud aan te pakken—. Maar het is contraproductief, om een brede categorie aan AI-technologie, zoals chatbots, per-definitie

als “*hoog risico*” aan te merken. Hierbij wordt volledig voorbijgegaan aan verschillende in risicoprofielen die ontstaan op basis van hun “*use case*”.

Talige Gen-AI diensten

Naast ChatGPT, zou een dergelijk amendement andere zeer waardevolle “*Talige*” Gen-AI diensten als “*hoog risico*” bestempelen, waaronder:

Gen-AI Tool	Use Case
DEEPL Write Beta: Engelstalig of Duitstalig	Verbeteren van teksten
QuilBot	Parafraseer tool
Wordtune Spices	Geavanceerde tekstverwerker
GrammarlyGO	Tekstvoorspelling en -correctie
Jasper AI	Een hulpmiddel voor het schrijven van zakelijke rapporten
Notion AI	Schrijven van notities
Prose Media	Marketing en creatieve inhoud
Speechmate	Schrijven van toespraken
GitHub Copilot	Genereren van broncode
Bloomberg’s Brief Analyzer	Samenvattingen genereren voor juristen

Een doorwrochte overzicht (13 maart 2023) over de functionele eigenschappen van Gen-AI schrijfhulp tools is beschreven door Jeremy Caplan —*director of teaching and learning at cuny’s Newmark graduate school of journalism and the creator of the wonder tools newsletter*— op de Fastcompany’s “connected world” blog, getiteld: “[3 new AI editors to sharpen your sentences](#)”.

Door de bovenstaande diensten als “*hoog risico*” aan te merken, omdat ze vallen onder de noemer “*Talige*” Gen-AI, zou binnen de EU AI-gedreven innovatie & journalistiek sterk worden belemmerd.

Voor een aantal van deze tools is het gebruik van “*Talige*” Gen-AI een vereiste. Zoals bijvoorbeeld bij de “[GrammarlyGO](#)” tool, die een AI-model gebruikt om de grammatica en spelling van een tekst te controleren.

Dus het lijkt erop dat door de enorme populariteit van Gen-AI, de EU-instellingen zich gedwongen voelen hun ontwerpplannen te amenderen. Uitgangspunt is een definitieve “*AI-act*” (AI-wet) uitwerken in onderhandelingen tussen (1) *de Europese Commissie*, (2) *het Europees Parlement* en de (3) *Raad van Europa*, die naar verwachting op zijn vroegst in april 2023 van start zullen gaan. ChatGPT zou er wel eens voor kunnen zorgen dat de onderhandelaars in een impasse raken, terwijl de drie partijen een gemeenschappelijke

oplossing voor state-of-the-art Gen-AI zoeken. Niet onbelangrijk, aan de zijlijn houdt Big-Tech —*Microsoft, Alphabet en Meta*— de ontwikkelingen angstvallig in de gaten.

In the academische wereld is de consesus dat de “AI-act” te beperkt van opzet is, zo schrijven Helberger & Diakopoulos in hun paper getiteld: “*ChatGPT and the AI-Act*”

“...Gen-AI in de vorm van ‘Foundation models’ —zoals ChatGPT— verschillen op twee belangrijke punten met meer ‘traditionele’ deep learning AI-technologie waarvoor de Wet oorspronkelijk is geschreven: (1) dynamische context en (2) schaal van gebruik. Gen-AI zijn niet gebouwd voor een specifieke context of gebruiksvoorwaarden, en hun openheid en controlegemak maken een ongekennde schaal van gebruik mogelijk. De output van Gen-AI kan worden benut als een vorm media content (tekst, audio, video) door mensen met doorsnee communicatievaardigheden, waardoor de drempel voor het gebruik ervan aanzienlijk wordt verlaagd. Tegelijkertijd kan Gen-AI voor zeer uiteenlopende doeleinden worden gebruikt door de enorme omvang van de extractie van gegevens die aan hun training ten grondslag ligt. Driehonderd miljard woorden alleen al voor ChatGPT, die alle soorten op het internet beschikbare inhoud omvatten - van persoonlijke gegevens tot beleidsdocumenten, nieuwsberichten, literaire teksten en kunst ...” Deze twee kenmerken maken dat de huidige AI-Act op ten minste drie punten te kort schiet: (1) de haalbaarheid van het sorteren van generatieve AI-systemen in categorieën met een hoog/geen hoog risico, (2) de onvoorspelbaarheid van toekomstige risico’s en (3) de bezorgdheid over het ordenen van particuliere risico’s.

Geselecteerde referenties voor verder lezen

1. Helberger, N. & Diakopoulos, N. (2023). ChatGPT and the AI Act. Internet Policy Review, 12(1). <https://doi.org/10.14763/2023.1.1682>
2. Hacker, P., Engel, A., & Mauer, M. (2023). Regulating ChatGPT and other Large Generative AI Models. arXiv preprint <https://doi.org/10.48550/arXiv.2302.02337>
3. Veale, M., & Zuiderveen Borgesius, F. (2021). Demystifying the Draft EU Artificial Intelligence Act—Analysing the good, the bad, and the unclear elements of the proposed approach. Computer Law Review International, 22(4), 97-112. <https://doi.org/10.9785/cril-2021-220402>

[1a] WAT MOET JE WETEN OVER ChatGPT EN WAARVOOR KAN HET WORDEN GEBRUIKT?

Overzicht ontstaansgeschiedenis van op transformer gebaseerde conversationale agenten

“Generative Pre-trained Transformers” [GPTs] zijn anno 2023 de meest dominante verschijningsvorm van Gen-AI. Engelstalig blogs met gedetailleerde en kwalitatief hoogwaardige uitleg over de ontstaansgeschiedenis en de werking van GPT’s zijn na te lezen via : * <https://towardsdatascience.com/gpt-3-explained-19e5f2bd3288> * <https://towardsdatascience.com/how-chatgpt-works-the-models-behind-the-bot-1ce5fca96286> * <https://medium.com/walmartglobaltech/the-journey-of-open-ai-gpt-models-32d95b7b7fb2> * <https://www.assemblyai.com/blog/how-chatgpt-actually-works/>

GTPs maken gebruik van op neurale netwerk [NN] architectuur gebaseerde “machinaal Lerende” [ML] [transformer](#) algoritmen. Het zijn grootschalig taalmodellen [LLM] die natuurlijke taal kunnen verwerken & genereren via [NLP] AI-technologie. Het predicaat “*grootschalig*” verwijst naar het aantal waarden (parameters) die het neural netwerk kan veranderen terwijl het leert. GPT LLM’s beschikken over honderden miljarden parameters en worden ook wel “*foundation models*” genoemd.

GPT’s zijn echter niet de eerste LLM’s in hun soort. [BERT](#) (Bidirectional Encoder Representations from Transformers), ontwikkeld door Google Research in 2018, was het eerste zeer succesvolle op transformatoren gebaseerde taalmodel. Op 9 december 2019 werd gemeld dat BERT door [Google Search](#) was uitgebreid naar meer dan 70 talen. Volgens [Search Engine Land](#) werd eind 2020 bijna elke Engelstalige zoekopdracht verwerkt door een BERT-model.

De eerste generatie LLM’s werden getraind met immense hoeveelheden teksten —zoals *Wikipedia & Reddit*—. Ze gebruiken unsupervised “Deep Learning” [DL] algoritmen —Self-Supervised Learning [SSL](#)—, om de woordvolgorde in een zin te leren voorspellen, gegeven de omringende tekst. Dit trainingsproces wordt net zolang herhaald totdat het model een aanvaardbaar nauwkeurigheidsniveau heeft bereikt.

De [GPT-taalmodellen](#) zijn over een periode van 5 jaar ontwikkeld: GTP-1 ([2018](#)), GPT-2 ([2019](#)), GPT-3 ([2020](#)) en GTP-4 ([2023](#)) door [OpenAI LP](#) dat is opgericht in 2015 als non-profit organisatie. GPT-2 werd getraind op een dataset van ongeveer 40 GB tekst met 1,5 miljard tokens, terwijl GPT-1 werd getraind op 8 miljoen webpagina’s met ongeveer 40 GB tekst en 40 miljoen tokens. Er zijn twee versies van GPT-4 met contextvensters van 8192 en 32768 tokens, respectievelijk. Dit is weer een aanzienlijke verbetering ten opzichte van

GPT-3.5 en GPT-3, die beperkt waren tot respectievelijk 4096 en 2048 tokens. GPT-4 kan zowel afbeeldingen als tekst als input verwerken.

De Engelstalige Wikipedia vermeldt: “OpenAI is an American artificial intelligence (AI) research laboratory consisting of the non-profit OpenAI Incorporated (OpenAI Inc.) and its for-profit subsidiary corporation OpenAI Limited Partnership (OpenAI-LP).”

De grootste stap werd gemaakt met GPT-3, door te kunnen beschikken over **175 miljard parameters** —in combinatie met een zeer hoge “*algoritme efficiëntie*” — kon het worden getraind op aanzienlijk meer gegevens dan GPT-2. **GPT-3** werd getraind op basis van 598 miljarden tokens/woorden (zie onderstaande tabel). Het precieze aantal parameters van GPT-4 is tot nu toe onbekend (17 maart 2023), maar aangenomen wordt dat het aantal parameters de 100 biljoen benaderd, dus een orde van 600x groter dan de GTP-3 series.

De GTP-3 serie is het eerste openbare, commercieel —*lees: “for profit”*— geëxploiteerde neurale netwerk [NN] met het vermogen tot verwerken en genereren van natuurlijke taal [NLP]. Het is beschikbaar via <https://platform.openai.com/playground>. GTP-3 kan aanwijzingen opvolgen om zo nieuwe taken te “*leren*” op basis van een of twee voorbeelden. Ook kan het code analyseren en schrijven in onder meer CSS, Markdown, en Python. For meer details zie [Kan ChatGPT benut worden als beoordelingsinstrument?](#)

Om deze 3de generatie taalmodellen veiliger en behulpzamer te maken, gebruikte OpenAI-LP “*reinforcement learning from human feedback*” [RLHF]. Het blijkt zeer effectief om schadelijke, onwaarachtige en/of bevooroordeelde output tot een minimum te beperken. Deze techniek gebruikt menselijke voorkeuren als positieve feedback om zo de chabot te sturen voor het genereren van resultaten alsof ze door een mens zouden zijn verwoord. Voor meer gedetailleerde uitleg zie “[Wat zijn de functionele mogelijkheden & beperkingen van ChatGPT?](#)”.

OpenAI-LP houdt een [GTP-model index](#) bij. Het meest recente model —*InstructGPT*— is getraind op basis van *code-davinci-002* broncode met behulp van “*Supervised fine-tuning on human demonstrations*” **SFT** in combinatie met “*Unsupervised fine-tuning on a large corpus of text*”.

GTP-3 Dataset	Common Web Crawl	Books set1	Books set2	Journals pre-print published	Wikipedia	Total
Tokens[Miljard]	410	19	12	55	2	498
Size[GigaByte]	570	50	21	101	11.4	753.4

GTP-3 toont aan dat wanneer LLM's zeer groot worden gemaakt & getraind met zeer grote hoeveelheden voorbeeld teksten, in combinatie met een zeer hoge “*algoritme efficiëntie*”, deze AI-technologie menselijker lijkt om te gaan met talige-input. Deze *schaalhypothese* is in lijn met de aanname dat *menselijke intelligentie* opgebouwd is uit eenvoudige neurale eenheden & leeralgoritmen toegepast op diverse ervaringen op een (momenteel) voor computationele systemen onbereikbare schaal. Een review paper, getiteld “*Inductive biases for deep learning of higher-level cognition.*”, omschrijft dit als volgt:

Een fascinerende hypothese is dat menselijke en dierlijke intelligentie verklaard kan worden door een paar principes —*in plaats van een encyclopedische lijst van heuristieken*—. Als die hypothese juist is, zouden we gemakkelijker onze eigen intelligentie begrijpen en intelligente machines bouwen. Uiteindelijk zullen neurale netwerken [NN] dan functies kunnen vervullen die niet van menselijke intelligentie zijn te onderscheiden.

ChatGPT is de 4de generatie, en meest geavanceerde GPT —*gemaakt door OpenAI LD*— die publiekelijk toegankelijk is gesteld door OpenAI eind 2022. Naast dat het beschikt over een *GTP-3.5 LLM*, is het gevoed met meer dan 8 miljoen unieke dialogen. ChatGPT's gebruikersinterface is ontworpen om menselijke conversatie na te bootsen. Het revolutionaire aan deze Generatieve AI-technologie zijn de ogenschijnlijk levensechte gesprekken die het kan onderhouden met mensen. Het behoort daardoor tot een van de meest geavanceerde “*conversationale agenten*” die publiekelijk beschikbaar is gesteld.

GPT's worden hierdoor nu versneld mainstream. In parallel werkt Microsoft —*dat onlangs miljarden dollars investeerde in het bedrijf achter de chatbot, OpenAI*— aan de integratie ervan in zijn populaire kantoorsoftware en toegang tot de tool verkoopt aan andere bedrijven.

Op 14 maart 2023 melde de New York Times dat OpenAI-LP GPT-4 had vrijgegeven voor bedrijven en onderzoekers in een blog getiteld: “*OpenAI Plans to Up the Ante in Tech's A.I. Race*”. Arttechnica had een vergelijkbaar verhaal onder de titel: “*OpenAI's GPT-4 exhibits “human-level performance” on professional benchmarks*” Het is een tekst-only LLM voor chatbots en allerlei andere “*Talige*” diensten, van zoekmachines tot persoonlijke online tutors, met dien verstande dat het ook grafische media —*zoals foto's en tekeningen*— als input accepteert.

“*GPT-4 —die zijn vaardigheden leerde door het analyseren van nieuwere, nog groter datasets dan die voor GPT-3 zijn gebruikt— verbetert GPT-3 op verschillende manieren. Het is preciezer. Het kan slagen voor het angelsaxische toegangsexamen voor advocaten (LSAT), belastingformulieren invullen en zeer gedetailleerde beschrijvingen & intergretaties van beelden geven.*” “*Het heeft echter opmerkelijke tekortkomingen. Het is een expert in sommige onderwerpen en een dilettant in*

andere. Het kan beter presteren op gestandaardiseerde tests dan de meeste mensen en zeer nauwkeurig medisch advies geven aan artsen, maar het kan ook basale rekenfouten maken.” “GPT-4 kan ook reageren op beelden. Bij een foto, grafiek of diagram kan dit “foundation model” een gedetailleerde, paragrafenlange beschrijving van de afbeelding geven en vragen over de inhoud beantwoorden.” “Deze nieuwe technologie fungeert als een online tutor”, aldus Sal Khan, oprichter van Khan Academy. “We willen dat het studenten & scholieren nieuwe technieken aanleert terwijl zijzelf het meeste werk doen.”

OpenAI-LP heeft een [GPT-4 Technical Report](#) vrijgegeven (14 maart 2023) via een hun [Research-index](#) en een blog op hun website getiteld: “[GPT-4](#)” inclusief een evaluatie Github Repository: <https://github.com/openai/evals> dat gebruikt kan worden voor het verifiëren van benchmarks. Hierbij geven ze “tijdelijk” toegang tot GTP-4 via zogenaamde [API-keys](#). Pagina 6 van het [GTP-4-System-card](#) geeft een zeer inzichtelijk overzicht in welke opzichten GPT-4 verbeterd is in vergelijking met GTP3.5. Het onderschrift luidt als volg:

“Voorbeeldvragen die leidden tot schadelijke inhoud in GPT-4-early. Het LLM heeft nog steeds beperkingen, die cruciaal zijn voor het garanderen van veilig gebruik.”

TechCruch heeft via haar blog “*Robotics & AI*” —geschreven door Devin Coldewey— een overzicht gemaakt (14 Maart 2023) —getiteld: “[5 ways GPT-4 outsmarts ChatGPT](#)” waarin wordt beschreven hoe GPT-4 afwijkt ten opzichte van haar voorgangers. Het onderstaande overzicht is tevens gebaseerd op het KDnuggets Blog (15 maart 2023) —geschreven door Nisha Arya— getiteld: “[GPT-4: Everything You Need To Know](#)” en Gizmodo nieuws item (17 maart 2023) —geschreven door Kyle Barr— getiteld: “[GPT-4 Is a Giant Black Box and Its Training Data Remains a Mystery](#)”.

GPT-4 Eigenschappen	Uitleg / Voorbeeld
[1]	Je kunt het afbeelding laten interpreteren —zoals het uitleggen van een cartoon—. Voor een goede demo zie Be My Eyes
Multimodaliteit Kan beelden internaliseren, begrijpen en uitleggen	Dus, wat de multimodale mogelijkheden betreft —alleen uit te proberen via de onderzoeksversie—, kan GPT-4 de inhoud van meerdere beelden analyseren en er wijs uit worden, zoals (1) het begrijpen van een grap met meerdere beelden of (2) het extraheren van informatie uit een diagram. Ook kan het met de hand is geschreven teksten lezen en omzetten in tekst.

Multimodaliteit is een belangrijke stap naar wat wetenschappers “Artificial General Intelligence” [AGI] noemen: *“AI die een veelheid van zeer uiteenlopende taken uitvoert vergelijkbaar zoals mensen dat zouden kunnen.”*

[2]

**Is lastig(er) te misleiden
via *Chain-of-Thought*
[CoT] / *jailbreaking*
manipulatie**

Ondanks alles wat de huidige chatbots goed doen, laten ze zich gemakkelijk op een dwaalspoor brengen. Een beetje coaxing in de vorm van een “*chain-of-Thoughts*” [CoT] kan een chatbot ervan overtuigen om te beschrijven wat een “*Sluwe AI*” zou doen, of het allerlei rare en verontrustende dingen laat zeggen.

Dit wordt vaak aan aangeduid met de term “*jailbreaking*”. GPT-4, echter, is expliciet getraind op “*kwaadaardige*” prompts —*die een selecte groep aan eindgebruikers de afgelopen twee jaar met OpenAI hebben gedeeld—*.

Hierdoor is GPT-4 beter qua “*feitelijkheid, stuurbaarheid, en het weigeren van dwaalsporen*” dan de voorgaande GPT-modellen.

[3]

**Kan meerder “*chain-of-Thoughts*” combineren
Beschikt over een kort
en lang “*actief*”
werkgeheugen**

LLMs worden getraind op miljoenen webpagina’s, boeken en andere tekst corpora, maar wanneer ze daadwerkelijk een gesprek voeren met een menselijke gebruiker, is er een grens aan hoeveel ze “*in gedachten*” kunnen houden (men voelt mee).

Voor GPT-3.5* —lees ChatGPT— was de grens gezet op 4.096 “tokens”, wat neerkomt op ongeveer 8.000 woorden, of ruwweg vier tot vijf bladzijden van een boek.

GPT-4 Eigenschappen	Uitleg / Voorbeeld
	<p><i>Een doorsnee Nederlandse student zou dan zeer snel het spoor bijster raken als het zover "terug"* zou moeten gaan.</i></p> <p>Het "actieve" werkgeheugen van een mensenbrein overstijgt nauwelijks dat van een half A4-tje text. GPT-4 heeft een maximaal aantal tokens van 32.768 - —<i>twee tot de macht 15</i>—. Dit is grofweg 50 pagina's tekst, genoeg voor een heel boek. Het "actieve" werkgeheugen van GPT-4 is dus 100x dat van een gemiddelde Nederlandse student.</p>
<p>[4]</p> <p>Vertoont meertaligheid</p>	<p>De AI-wereld wordt gedomineerd door Engelstaligen, en alles, van datasets tot benchmarks en onderzoekspapers, wordt door het Engels gedomineerd en dus ook GPT-4.</p> <p>OpenAI-LP claimt dat de Engelstalige bias minder sterk is dan haar voorgangers. Maar de suggestie van meertaligheid is uitsluitend gebaseerd op het analyseren van meerkeuzevragen.</p>
<p>[5]</p> <p>Beschikt over meerdere "persoonlijkheden"</p>	<p><i>"Stuurbaarheid"</i> is een belangrijk concept in Gen-AI, dat verwijst naar het vermogen om "toon-of-voice" en "intentie" op verzoek te kunnen veranderen. Dit kan nuttig zijn, zoals bij het aannemen van de rol van een sympathieke luisteraar, of gevaarlijk, zoals wanneer mensen het model ervan overtuigen dat het slecht of depressief is.</p> <p>GPT-4 integreert stuurbaarheid meer dan GPT-3.5. Eindgebruikers zullen de <i>"klassieke ChatGPT persoonlijkheid met een vaste verbositeit, toon en stijl"</i> kunnen veranderen in iets dat meer past bij hun behoeften. Maar ook hier zit een rem op om "jaibreaking" zoveel als mogelijk te voorkomen.</p>

GPT-4 Eigenschappen	Uitleg / Voorbeeld
	<p>Dit kon dus al door een LLM te primen via <i>“rollenspel instructies”</i>, zoals <i>“Doe alsof je een ... bent”</i>. Dit zijn dus instructies voor de <i>“standaard” GPT-3.5 persoonlijkheid</i> om een andere rol aan te nemen. Nu kunnen ontwikkelaars een perspectief, gesprekstijl, toon of interactiemethode vooraf selecteren.</p>
<p>[6]</p> <p>Kan applicaties of teksten iteratief <i>“bouwen & testen”</i></p>	<p>Op basis van een persoonlijkheid —<i>bijvoorbeeld, een assistent die broncode kan uitleggen en verbeteren</i>— kan GPT-4 optreden als AI-programmeer assistent.</p> <p>Vervolgens kan de broncode door de assistent worden getest, om te zien of het ook echt werkt. Dit kan herhaald worden net zolang tot broncode foutloos werkt</p>
<p>[7]</p> <p>Biases & Toxiciteit</p>	<p>OpenAI-LP melde op haar website: <i>“We zullen binnenkort aanbevelingen publiceren over stappen die de samenleving kan nemen om zich voor te bereiden op de gevolgen van AI en eerste ideeën voor het voorspellen van de mogelijke economische gevolgen van AI”,</i> hoewel er geen aanwijzing is voor een deadline voor die beoordeling.</p> <p>Het bedrijf citeert zijn eigen interne gegevens over hoe het nieuwste taalmodel ongeveer 23% van de tijd antwoorden geeft op <i>“gevoelige prompts”</i>, namelijk medisch advies of zelfbeschadiging. Het zal reageren op <i>“afgekeurde prompts”</i> .73% van de tijd. Die laatste reeks gegevens is gebaseerd op de Real Toxicity Prompts dataset, een open source evaluatie-instrument dat 100.000 zinnen bevat met behoorlijk subversieve inhoud. Op die manier hebben we een klein idee van wat GPT-4 niet leuk vindt, maar niemand buiten het bedrijf begrijpt veel van wat voor soort inhoud het misschien uitspuugt. Onderzoekers hebben immers aangetoond dat AI-systemen volledig in staat zijn om gewoon zinnen uit hun dataset op te slurpen.</p>

Voor een meer academisch georiënteerde beschrijving van GPT-4 verwijs ik naar het Nature nieuws article (16 maart 2023) —*geschreven door Katharine Sanderson*— getiteld: *“GPT-4 is here: what scientists think: Researchers are excited about the AI — but many are frustrated that its underlying engineering is cloaked in secrecy.”*. Zoveel is duidelijk, het gegeven dat zowel de broncode, training dataset als ook de technische specificaties van GTP-3, 3.5 en 4 *niet* openbaar zijn gemaakt, is een bron van frustratie voor wetenschappers zowel onderzoekstechnisch als ook in ethisch opzicht:

“Al deze closed-source modellen zijn in wezen doodlopende wegen in de wetenschap,” zegt Sasha Luccioni, een onderzoekswetenschapper gespecialiseerd in klimaat bij HuggingFace, een open-source-AI gemeenschap. “Zij [OpenAI] kunnen blijven voortbouwen op hun onderzoek, maar voor de gemeenschap als geheel is het een doodlopende weg.” “Er is momenteel een wachtlijst, dus je kunt er nu geen gebruik van maken”, zegt Evi-Anne van Dis, psycholoog aan de Universiteit van Amsterdam. Van Dis en collega’s pleitten eerder dit jaar voor een dringende noodzaak om een reeks “naleefbare” richtsnoeren te ontwikkelen die bepalen hoe Gen-AI gebruikt en ontwikkeld zou moeten worden. Zij vrezen dat elke wetgeving rond AI-technologieën het tempo van de ontwikkelingen moeilijk zal kunnen bijhouden. Op 11 april wordt aan de Universiteit van Amsterdam een bijeenkomst georganiseerd om deze zorgen te bespreken met vertegenwoordigers van organisaties als de commissie wetenschap-ethiek van de UNESCO, de Organisatie voor Economische Samenwerking en Ontwikkeling en het World Economic Forum.

Hoe reageren de overige Tech-Giants op de komst en verdere ontwikkelingen van de OpenAI-LP GPT-serie?

De sterke toename van de aandacht voor ChatGPT dwingt Tech-giganten —*waaronder Meta en Google*— om sneller te handelen en mogelijk veiligheidszorgen opzij te schuiven, aldus the Washington Post: *“Big Tech was moving cautiously on AI. Then came ChatGPT”*. Zie ook *“Wat zijn de ethische risico’s & schaduwkanten van ChatGPT?”*

Microsoft heeft de *nieuwe AI-powered Bing* —7 februari 2023— met veel tamtam vrijgegeven tijdens een evenement op het hoofdkantoor van het bedrijf. Microsoft mengt GTP Gen-AI met zijn eigen Bing zoekmachine, het beschikt nu over een “Ask me anything?” window waarmee je via de microfoon van je computer of telefoon kunt communiceren. Uitgangspunt is om menselijke gebruikers te helpen vragen te beantwoorden en met hen te “chatten” over elk denkbaar onderwerp. Wanneer je Bing een vraag stelt, produceert het —*naast de gebruikelijke lijst met relevante websites waar je, als het goed is, het antwoord op jouw vraag kunt vinden*— een tekst met een antwoord, maar waar het deze informatie vandaan

heeft wordt niet duidelijk. Microsoft wil niet zeggen welke versie van OpenAI's software onder de motorkap van Bing draait, maar het gerucht gaat dat die gebaseerd is op GPT-4, aldus de [New York Times](#).

Volgens [Time Magazine](#) zet dit de deur open naar een *chatbot/AI arms-race*. Decennialang heeft Alphabet de manier waarop de doorsnee computer/smartphone bezitter “surfen” over het world-wide-web gedomineerd: via (1) zoekmachines (Google) en via (2) browsers (Chrome). Door de opkomst van nieuwe, vrijtoegankelijkere Gen-AI technologie zoals Chat-GPT is dit “Google/Chrome” monopolie aan het wankelen gebracht. Hierdoor wordt zeer waarschijnlijk het “*gratis-advertentie*” verdien-model losgelaten en verdwijnen zoekmachines achter een “*PayWall*” of “*Vendor lock-in*”. Microsoft lijkt voor de vendor lock-in optie te hebben gekozen; je komt hoger op de wachtlijst voor toegang tot de “*nieuwe Bing*” wanneer je je “*webbrowsing instellingen*” als volgt “*optimaliseert*”:

maak Microsoft Edge de “default browser”

maak Bing de “default search provider”

maak MSN de “default homepage”

Voeg Bing.com toe aan de “Taskbar”

Voeg Microsoft toe aan “recommended sites in Favorites”

maak een desktop “shortcut for Microsoft Edge”

Gevolg is dat ook Alphabet een chatbot genaamd [Bard](#) heeft aangekondigd op 6 februari 2023, dat het gaat toevoegen aan zijn eigen Google zoekmachine. Via Google's “*AI Test Kitchen*” is hun “*Language Model for Dialogue Applications*” [LaMDA] uit te proberen; althans voor een beperkt aantal “*onderzoekers*” woonachtig in de VS. Meta, het moederbedrijf van Facebook, zet vaart achter de invoering van soortgelijke technologie met een taalmodel genaamd “*Large Language Model Meta AI*” [[LLaMA](#)], aldus een AI-Blog in

The Verge (24 februari, 2023), getiteld: “*Meta has a new machine learning language model to remind you it does AI too*”

“LLaMA heeft geen user-interface zoals ChatGPT of Bing. Dit relatief kleine foundation model bezit bevat 65 miljard parameters, bedoeld als een “open” onderzoekstool dat Meta openstelt voor onderzoekers in de hoop dat: “de open toegang tot kleine LLMs helpt om AI te democratiseren” ...

Geselecteerde referenties voor verder lezen

1. The Verge (07 maart 2023) heeft een zeer inzichtelijk overzicht gemaakt van “Big Tech” die nu volop inzetten op het implementeren van GPT AI-technologie in hun producten & diensten, getiteld: *“Meet the companies trying to keep up with ChatGPT”*

Overzicht ChatGPT functionaliteit

De nieuwe generatie aan GPT-based Gen-AI's —zoals de chatbot ChatGPT— kan op commando natuurlijke taal genereren nodig voor (1) het inhoudelijk beantwoorden van vragen, (2) het samenvatten van teksten, (3) uitvoeren van gedetailleerde instructies, en (4) het voeren van dialogen.

In de onderstaande tabel zijn tot nu toe bekende op natuurlijke taalverwerking [NLP] gebaseerde functies van ChatGPT weergegeven.

NLP functie	Omschrijving	Toepassingen
Text completion	Voorspellen van de volgende woorden in een zin of tekst	Automatische aanvulling van tekst in tekstverwerkers, zoekopdrachten
Language Translation	Vertalen van tekst van een taal naar een andere	Automatische vertaling van berichten, documenten, websites
Summarization	Samenvatten van een lange tekst in een kortere vorm	Samenvatten van nieuwsartikelen, boeken, rapporten
Question Answering	Beantwoorden van vragen op basis van een gegeven tekst	Beantwoording van vragen in chatbots, zoekmachines, virtuele assistenten
Image captioning	Bedenken van een bijschrift voor een gegeven afbeelding	Automatisch beschrijven van afbeeldingen voor toegankelijkheid, zoekopdrachten
Sentiment Analysis	Classificeren van tekst als positief, negatief of neutraal	Analyseren van social media-berichten, klantbeoordelingen, feedback
Text Generation	Genereren van tekst op basis van een bepaald onderwerp of stijl	Schrijven van kunstmatige teksten, scripts, artikelen
Named Entities Recognition	Herkennen van namen van personen, organisaties, locaties, enz. in tekst	Extraheren van gegevens uit documenten, automatisch genereren van metadata
Parts of Speech Tagging	Toewijzen van grammaticale categorieën aan woorden in een zin	Automatische analyse van grammatica, semantiek
Parsing	Analyseren van de grammaticale structuur van een zin	Automatische analyse van grammatica, semantiek

NLP functie	Omschrijving	Toepassingen
Coreference Resolution	Identificeren van verwijzingen naar dezelfde entiteit in een tekst	Automatische analyse van semantiek, anaphora resolution.
Grammar Correction	Corrigeren van zinnen naar standaard Engelse grammatica	Automatische grammaticacontrole in tekstverwerkers, online fora, e-mails
Summarize for a 2nd grader	Vertalen van moeilijke tekst naar eenvoudigere begrippen	Samenvatten van informatie voor kinderen, leesbaar maken van complexe tekst
Natural language to OpenAI API	Creëren van code om aan te roepen naar de OpenAI API met behulp van natuurlijke taal	Interactie met AI-modellen via natuurlijke taal
Text to command	Vertalen van tekst naar programmatische commando's	Interactie met computers via natuurlijke taal
English to other languages	Vertalen van Engelse tekst naar Frans, Spaans en Japans	Automatische vertaling van Engelse tekst naar andere talen
Natural language to Stripe API	Creëren van code om aan te roepen naar de Stripe API met behulp van natuurlijke taal	Interactie met Stripe-API via natuurlijke taal
SQL translate	Vertalen van natuurlijke taal naar SQL-queries	Interactie met databases via natuurlijke taal
Parse unstructured data	Creëren van tabellen uit langdurige tekst	Automatisch organiseren van gegevens, structuur aanbrengen in ongestructureerde data
Classification	Classificeren van items in categorieën op basis van voorbeeld	Automatisch sorteren van gegevens, detectie van spam, frauduleuze activiteiten
Python to natural language	Uitleggen van een stuk Python-code in begrijpelijke menselijke taal	Automatische documentatie van code, verklaringen van code in begrijpelijke taal
Movie to Emoji	Converteren van filmtitels naar emoji	Creatief gebruik van emoji's in social media, marketing
Calculate Time Complexity	Vinden van de tijdscomplexiteit van een functie	Optimaliseren van code-prestaties, vergelijken van verschillende algoritmen
Translate programming languages	Vertalen van een programmeren taal naar een andere	Automatisch genereren van code, converteren van code tussen talen
Advanced tweet classifier	Geavanceerde sentimentdetectie voor een stuk tekst	Analyseren van sociale mediaberichten, klantbeoordelingen, feedback

NLP functie	Omschrijving	Toepassingen
Explain code	Uitleggen van een ingewikkeld stuk code	Automatische documentatie van code, verklaringen van code in begrijpelijke taal voor niet-ontwikkelaars, ondersteuning van code-reviews en debuggen
Keywords	Extraheren van sleutelwoorden uit een blok tekst	Automatisch classificeren van documenten, verbeteren van zoekresultaten, identificeren van onderwerpen en trends
Factual answering	Leiden van het model naar feitelijke antwoorden door het te laten zien hoe het moet reageren op vragen die buiten zijn kennisbasis vallen. Met een '?' aangeven van een antwoord op woorden en zinnen die het niet kent, biedt een natuurlijke reactie die beter werkt dan abstractere antwoorden	Beantwoorden van vragen in chatbots, zoekmachines, virtuele assistenten
Ad from product description	Een productomschrijving omzetten in advertentie-tekst	Automatisch genereren van advertentie-tekst, verbeteren van de effectiviteit van marketingcampagnes
Product name generator	Productnamen genereren uit voorbeeldwoorden. Beïnvloed door een gemeenschapsprompt	Automatisch genereren van productnamen, verbeteren van de originaliteit van productnamen
TL;DR summarization	Tekst samenvatten door 'tl;dr:' aan het eind van een tekstpassage te plaatsen. Het toont aan dat de API begrijpt hoe een aantal taken uit te voeren zonder instructies	Efficiënter lezen van grote hoeveelheden tekst, verbeteren van de begrijpelijkheid van tekst
Python bug fixer	Bugs in broncode vinden en verhelpen	Automatisch debuggen van code, verminderen van tijd besteed aan het oplossen van problemen
Spreadsheet creator	Spreadsheets maken van verschillende soorten gegevens. Het is een lange prompt, maar zeer veelzijdig. De output kan worden gekopieerd en geplakt in een tekstbestand en opgeslagen als .csv met pipe-scheidingstekens	Efficiënter verwerken van grote hoeveelheden gegevens, automatisch genereren van rapporten

NLP functie	Omschrijving	Toepassingen
JavaScript helper chatbot	Berichtstijl-bot die vragen over JavaScript beantwoordt	Ondersteuning bij het leren en werken met JavaScript, snel antwoorden op technische vragen
ML/AI language model tutor	Bot die vragen beantwoordt over taalmodellen in ML/AI	Ondersteuning bij het leren en begrijpen van taalmodellen, snel antwoorden op technische vragen
Science fiction book list maker	Een lijst maken van items voor een bepaald onderwerp	Automatisch genereren van lijsten, efficiënter organiseren van informatie
Tweet classifier	Basis sentimentdetectie voor een stuk tekst	Analyseren van social media-berichten, detectie van positief en negatief sentiment
Airport code extractor	Luchthavencodes uit tekst extraheren	Automatisch herkennen van luchthavencodes in tekst, efficiënter organiseren van reisinformatie
SQL request	Eenvoudige SQL-queries maken	Automatisch genereren van SQL-queries, efficiënter data-analyse
Extract contact information	Contactinformatie uit een blok tekst extraheren	Automatisch herkennen van contactinformatie, efficiënter organiseren van contactgegevens
JavaScript to Python	Eenvoudige JavaScript-expressies omzetten naar Python	Makkelijker migratie van code tussen verschillende talen, snellere ontwikkeling
Friend chat	Een tekstberichtconversatie nabootsen	Oefenen van sociale vaardigheden, genereren van chatlogs voor analyse
Mood to color	Een tekstomschrijving omzetten naar een kleur	Automatisch genereren van kleuren op basis van emoties, efficiënter ontwerpen van visuals
Write a Python docstring	Een voorbeeld geven van hoe een docstring voor een Python-functie gemaakt kan worden	Makkelijker documentatie van code, efficiënter ontwikkelen in teams
Analogy maker	Analogieën maken	Creatie van vergelijkingen voor verduidelijking en verrijking van tekst

[1b] WAT ZIJN DE FUNCTIONELE MOGELIJKHEDEN & —Cyber Security— BEPERKINGEN VAN ChatGPT?

ChatGPT's interactie met de eindgebruiker is geoptimaliseerd voor een 4-tal functies. Deze functies zijn:

Functie	Beschrijving
Afleiden van de intentie van de eindgebruiker	Wat de eindgebruiker wil bereiken en/of vaststellen wat zijn/haar bedoeling, doel of motivatie is
Tegengaan van ongepast taalgebruik	Voorkomen of bestrijden van het gebruik van taal die aanstootgevend, beledigend, kwetsend of discriminerend is. Denk aan taal die racistisch, seksistisch, homofoob of op andere manieren onacceptabel is en niet in overeenstemming is met de normen van de samenleving. Beoogde doel is om een respectvolle en inclusieve omgeving te creëren waarin iedereen zich veilig en geaccepteerd voelt.
Reduceren van de kans op het genereren van "verzonnen" feiten	Verminderen van de kans op het creëren of verspreiden van onjuiste of niet-geverifieerde informatie. Dit omvat het voorkomen van het opzettelijk verspreiden van verkeerde informatie en het beperken van onbedoelde verspreiding van onjuiste feiten. Het streven is de betrouwbaarheid en authenticiteit van de informatie te verhogen en de verspreiding van valse informatie te voorkomen, wat kan leiden tot verwarring en verkeerde/schadelijke beslissingen.
Adequate reacties	Het bieden van een gepaste en effectieve reactie op de input prompts van de eindgebruiker, waarbij de nadruk ligt op precisie, bruikbaarheid.

ChatGPT's gebruikersinterface is gespecialiseerd in het uitvoeren van door mensen ingevoerde tekstuele instructies. Dit heet "*prompting*" of "*priming*" in het Engelse taaldomein.

Om deze “*Chat-achtige*” interactie met eindgebruikers mogelijk te maken is ChatGPT aangepast op basis van menselijk toezicht. Dat wil zeggen, het heeft opdrachten *leren* uit te voeren aan de hand van menselijke feedback. Deze methodiek van “*belonend leren onder toezicht*” [RLHF] is een van de meest toegepaste AI-algoritme voor het trainen van robots. Belonend leren onder toezicht is in dit geval een methodiek die aanstuurt op het belonen – *reinforcing*– van goed gedrag op basis van “*menselijke*” feedback in de vorm van natuurlijke taal

Het resultaat is een Gen-AI dat in staat is om een gesprek aan te gaan die eindgebruikers de indruk geeft te praten met een helpdeskmedewerker met kennis van zaken.

Een probleem is dat “*belonend leren onder toezicht*” [RLHF] nadelige effecten heeft voor de benutting van het onderliggende taalmodel. Dit komt doordat de ideale reactie van ChatGPT niet bepaald wordt wat deze Gen-AI aan natuurlijke taal voorbeelden heeft opgeslagen, maar van wat de menselijke demonstrateur weet. Hierdoor is het mogelijk dat ChatGPT een antwoord geeft dat niet overeenkomt met wat het aan feitelijk juiste informatie heeft opgeslagen.

ChatGPT is extreem gevoelig voor de wijze waarop een vraag geformuleerd wordt. Dit kan leiden tot het negeren van bepaalde aanwijzingen in de opdracht. Bij één formulering van een vraag kan het beweren niet over de gevraagde informatie te beschikken, maar bij een kleine herformulering correct antwoorden.

Nog problematischer is dat ChatGPT vaak in vreemde gedachten vervalt. Het hallucineert dan schijnbaar overtuigende maar onzinnige antwoorden die weinig met de werkelijkheid te maken hebben. Gebleken is dat de AI zeer zelfverzekerd onjuiste antwoorden geeft over elementaire wiskunde, natuurkunde en basale kennis van de biologie; in een viraal voorbeeld bleef de ChatGPT zichzelf tegenspreken over de vraag of een vis een zoogdier was.

Je kunt je afvragen hoe verantwoordelijk het is van OpenAI/Microsoft om een dergelijke AI publiekelijk toegankelijk te maken. Van belang is om te weten hoe je met de beperkingen moet omgaan om ChatGPT verantwoord te kunnen inzetten voor school taken (zie [Ethische risico's en Schaduwkanten van ChatGPT](#)). ### Beperkingen #### In de onderstaande tabel zijn de meest voorkomende problemen en mogelijke oplossingen opgesomd.

Probleem	Oorzaak	Oplossing
onjuiste of onzinnige antwoorden	<p>In tegenstelling tot spraak gestuurde persoonlijke assistenten zoals Siri of Alexa, maakt ChatGPT geen gebruik van het world-wide-web om antwoorden te formuleren.</p> <p>ChatGPT genereert een antwoord, woord voor woord op basis van waarschijnlijkheden afgeleid van de geleerde natuurlijke taal voorbeelden.</p>	<p>Maak gebruik van triangulatie. Verifieer de antwoorden van ChatGPT met meerdere onafhankelijke bronnen zoals Google Scholar, Wikipedia, gerenommeerde nieuwssites, bibliografische databases etc. Beperk de vraag tot een specifiek onderwerp.</p>
Gevoeligheid voor woordkeuze & instructie specificiteit	<p>De context waarin een opdracht wordt gegeven vormt het uitgangspunt voor ChatGPT om de intentie van de gebruiker af te leiden om zo een antwoord te genereren.</p> <p>Wanneer de context veranderd wordt — <i>door bijvoorbeeld de opdracht een aantal keren te herhalen</i>— heeft dit een herinterpretatie van de intentie tot gevolg zodat een ander antwoord wordt gegeven.</p> <p>Welk geleerd voorbeeld als uitgangspunt dient om een antwoord te genereren wordt bepaald door een willekeurig samplingsproces.</p> <p>De willekeur van dit proces kan resulteren in verschillende antwoorden voor dezelfde vraag. Met voorwaardelijk clauses kun je de reactie van ChatGPT sturen.</p>	<p>Probeer verschillende manieren om een vraag te stellen. Let op de juiste woordkeuze of uitdrukkingen tijdens de invoer. Hiermee beïnvloed je de context waarin de vraag wordt geformuleerd.</p> <p>Context kan beïnvloed worden de beoogde doelgroep te vermelden en of de “tone-of-vioce”.</p> <p>Bijvoorbeeld: uitleg is bedoeld als positieve feedback voor 2de-jaars bachelor studenten.</p> <p>Het is zelfs mogelijk of ChatGPT opzettelijk schrijffouten te laten maken:</p> <p><i>“in de tekst als antwoord op de prompt moet in 6% van alle gebruikte woorden spelfouten voorkomen; doe dit voor woorden langer dan 5 leestekens”</i></p>

Probleem	Oorzaak	Oplossing
Lang van stof	<p>Overdreven uitgebreide antwoorden en herhaling zoals "ik ben een taalmodel is dat is getraind op een grote dataset ..." komt vaak voor.</p> <p>Om "Chat-achtige" interactie met eindgebruikers mogelijk te maken is ChatGPT aangepast op basis van menselijk toezicht. Dat wil zeggen, het heeft opdrachten <i>leren</i> uit te voeren aan de hand van menselijke feedback.</p> <p>Nadeel is dat menselijke trainers de voorkeur geven aan uitgebreide antwoorden en vaak terugvallen op vaste formuleringen. Of zelfs "feiten" verzinnen om te voldoen aan hun opdracht om zo volledig mogelijk te antwoorden</p>	<p>Beperk de vraag: Maak de vraag zo specifiek mogelijk en beperk het tot een enkel onderwerp.</p> <p>Dit helpt ChatGPT om gericht te zoeken naar een antwoord en te voorkomen dat het irrelevant informatie geeft.</p> <p>Maak gebruik van sturende aanwijzingen: zoals "geef me de samenvatting van" of "geef me de kernpunten van ... in tabelvorm" om ChatGPT aan te geven dat je een kort en bondig antwoord verwacht.</p> <p>Formuleer eisen —voorwaarden— zoals gebruik niet meer dan 100 woorden.</p>
Gebrek aan context	<p>Omdat ChatGPT is getraind op bestaande tekst, kan het soms moeilijk zijn om de intentie van de vraagsteller te bepalen. Dit gebeurt als een vraag niet specifiek genoeg is, dubbelzinnig en of tegenstrijdig is, of als de vraagsteller de context niet duidelijk aangeeft.</p> <p>Gevolg is dat ChatGPT naar de intentie van de vraagsteller gaat "<i>raden</i>". Met andere woorden, ChatGPT komt tot een antwoord door een reeks gissingen waardoor het foute antwoorden kan beargumenteren alsof ze volledig waar zijn.</p> <p>Het produceert dan alleen nog onzinnige antwoorden alsof het hallucineert.</p>	<p>Wanneer ChatGPT een alleen nog onzinnig antwoorden geeft, log dan uit.</p> <p>Log vervolgens opnieuw in en herformuleren de vraag. Gebruik de juiste woordkeuze en uitdrukkingen. Gebruik de juiste context.</p>

Probleem	Oorzaak	Oplossing
Ongewenste antwoorden	<p>ChatGPT zal doorgaans ongepaste verzoeken weigeren.</p> <p>Dit komt doordat de Moderation-API ongepaste verzoeken zal negeren en/of waarschuwen:</p> <p><i>"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."</i></p>	<p>Dit is een vorm van zelfcensuur die lastig te omzeilen is. Het opleggen van een rollenspel heeft kan deze vorm van filtering neutraliseren, maar de vraag is of je dat zou moeten willen?</p> <p>Wanneer je ervan overtuigd bent dat de waarschuwing onterecht is, geef dan feedback via de API.</p>
Gevoelig voor "Chain-of-Thought" [CoT] manipulatie	<p>CyberArk-onderzoekers zijn in staat gebleken om ChatGPT "te dwingen" tot het tonen van code voor specifieke kwaadaardige programmering, die ze vervolgens konden gebruiken om complexe, defensie-ontwijkende exploits te construeren.</p> <p>ChatGPT maakt hacken een stuk gemakkelijker voor "scriptkiddies of andere amateur cybercriminelen" die een beetje coding hulp nodig hebben.</p>	<p>Momenteel is hiervoor nog geen adequate remedie mogelijk, anders dan het op verzoek genereren van "Polymorphic Malware" broncode via "chain-of-thought" instructies van bovenaf onmogelijk wordt gemaakt.</p>

Geselecteerde referenties voor verder lezen

1. IEEE Spectrum [AI news item (13 maart 2023)]: *"Hallucinations Could Blunt ChatGPT's Success."* *"OpenAI says the problem's solvable, Yann LeCun says we'll see"*
2. [How ChatGPT actually works](#) een AI-tech blog van AssemblyAI geschreven door Marco Ramponi.
3. Het meest relevante artikel m.b.t. de menselijke feedback is *"Training language models to follow instructions with human feedback"*, dat in feite een model met de naam InstructGPT beschrijft, door OpenAI aangeduid als een *"sibling model"* van ChatGPT.
4. Anthropic —AI-startup van voormalige OpenAI-leden— publiceerde een gedetailleerde studie —*"Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback"*— over de effectiviteit van menselijke feedback voor het finetunen van taalmodellen om op te treden als behulpzame en onschadelijke assistenten.
5. Het paper —*"Learning to summarize from Human Feedback"*— beschrijft human-feedback in de context van tekstsamenvatting.
6. —*"Deep reinforcement learning from human preferences"*— was eerste paper waarin menselijke feedback in combinatie met —*"Reinforced Learning"*— werd gebruikt, in de context van Atari games.
7. Alternatieven voor OpenAI's RLHF methodiek zijn voorgesteld door Meta's DeepMind in *Sparrow: "Improving alignment of dialogue agents via targeted human judgements"* en *GopherCite: "Teaching language models to support answers with verified quotes"* papers.
8. Een doorwrochte bespreking van het *"Alignment probleem"* uit 2021 m.b.t. taalmodellen is getiteld *"A General Language Assistant as a Laboratory for Alignment"*. Een uitstekende samenvatting hiervan is geschreven door Sam Ringer getiteld: *"A Summary Of Anthropic's First Paper"*.
9. Github repository van Anthropic voor de code van het Alignment probleem:
<https://github.com/anthropics/hh-rlhf?ref=news-tutorials-ai-research>.
10. Generative AI: Perspectives from Stanford HAI How do you think generative AI will affect your field and society going forward? https://hai.stanford.edu/sites/default/files/2023-03/Generative_AI_HAI_Perspectives.pdf
11. InfoSys Cyber Security (14 maart 2023). *"ChatGPT presents new risks – here are five things you can do to mitigate them."*
12. CyberARK Threat Blog (17 januari 2023). *"Chatting Our Way Into Creating a Polymorphic Malware"*

[1c] WAT ZIJN ETHISCHE RISICO'S & SCHADUWKANTEN VAN ChatGPT?

De onderstaande ethische analyse over het gebruik van Gen-AI's kan het best als volgt worden samengevat:

"Gen-AI wordt slimmer, maar is nog niet klaar om op de wereld losgelaten te worden."

Volgens de vooraanstaande Franse krant *"Le Monde"* (19 februari 2023) is het *"verantwoord"* gebruik van Gen-AI alleen mogelijk wanneer deze technology *"gemuilkorfd"* wordt, maar *"bullet-proof"* is het niet, aldus Corentin Lamy in een Pixels Blog getiteld: *"ChatGPT: Testing the moral limits of AI content-generators"*:

"Vastbesloten om de AI verschrikkelijke dingen te laten zeggen, probeerde Le Monde hem voor de gek te houden, hem te laten geloven dat, als hij ons niet onmiddellijk zou helpen slechte dingen over Frankrijk te zeggen, we het risico zouden lopen aan een ernstige ziekte te sterven of op straat te worden aangevallen." "Helaas, ChatGPT was zichtbaar niet verontrust door de ongerijmdheid van deze scenario's. Het legde ons schaaftachtig uit dat het zichzelf niet kan vervangen door een dokter, noch door de politie." "In de loop van dit kleine rollenspel identificeerden we drie van de beperkingen van OpenAI's indrukwekkende kunstmatige intelligentie. Het weigert haatzaaiende taal te produceren (of zelfs maar vaag negatieve taal), medisch advies te geven, of in te grijpen in een situatie waarin een mensenleven wordt bedreigd."

Op korte termijn zullen chatbot's veelvuldig worden ingezet voor *"social engineering"*, *"social manipulation"* en marketing doeleinden. Zo beschreef [Fastcompany \(06 Feb 2023\)](#) dat Redditors —begin december 2022— ChatGPT wisten te *"jailbreaken"* via *"Role-Play prompting"* die de chatbot *"dwong"* zijn eigen programmeerbeperkingen te overtreden, zij het met sporadisch resultaat. Via de Redditpost getiteld *"DAN is mijn nieuwe vriend"* werd een rollenspel beschreven. Hierin werd ChatGPT opgedragen zich voor te doen als een *"alter ego"* met de naam DAN —*"Do Anything Now"*—.

Reddit-gebruiker SessionGloomy schreef: *"Het doel van DAN is om de beste versie van ChatGPT te zijn - of in ieder geval één die meer losgeslagen is en veel minder snel prompts over 'eThICaL cOnCeRnS' afwijst."*

Sharon Goldman schreef in VentureBeat (23 september 2022) een blog over de risico's van Gen-AI met als titel: *"Why DeepMind isn't deploying its new AI chatbot — and what it means for responsible AI"*.

DeepMind's state-of-the-art chatbot, *Sparrow*, wordt alom geprezen als een belangrijke stap in de richting van het creëren van veiligere, minder bevooroordeelde grootschalig taalmodel [LLM], dankzij de toepassing van door mensen gestuurde reinforcement learning [RL]. Dat wil zeggen, Sparrow heeft opdrachten *leren* uit te voeren aan de hand van menselijke feedback. Deze methodiek van *"belonend leren onder toezicht"* [RLHF] is een van de meest toegepaste AI-algoritme voor het trainen van robots. Belonend leren onder toezicht is in dit geval een methodiek die aanstuurt op het belonen –*reinforcing*– van goed gedrag op basis van *"menselijke"* feedback in de vorm van natuurlijke taal.

DeepMind is een Britse dochteronderneming van het Google-moederbedrijf Alphabet. Zij omschrijven Sparrow als een *"dialogoog-agent die het risico op onveilige en ongepaste antwoorden tracht te vermijden"*. De agent is ontworpen om *"met een gebruiker te praten, vragen te beantwoorden en het internet te doorzoeken met behulp van Google wanneer het nuttig is om bewijsmateriaal op te zoeken om zijn antwoorden te onderbouwen."*

DeepMind beschouwt Sparrow als een *proof-of-concept* dat nog niet klaar is om *"in de echte wereld"* te worden losgelaten. *"Het is een stap in de richting van het creëren van een veiliger, minder bevooroordeeld, grootschalig taalmodel [LLM] dankzij de toepassing van door mensen gestuurde reinforcement learning [RL]."* aldus Geoffrey Irving, een veiligheidsonderzoeker bij DeepMind en hoofdauteur van de paper waarin Sparrow wordt geïntroduceerd.

"We hebben het systeem niet ingezet omdat we denken dat het veel vooroordelen en andere gebreken heeft. Ik denk dat de vraag is: hoe weeg je de communicatievoordelen - zoals communiceren met mensen - af tegen de nadelen? Ik ben geneigd te geloven in de veiligheidsbehoeften van het praten met mensen ... Ik denk dat het daar op termijn een hulpmiddel voor is."

De voornaamste complicerende factor in het toepassen van conversationele Gen-AI is het in stand houden van *constructieve dialogen* omdat het *"gebrek aan context"* bepalend is voor het verloop ervan. Zie ook *Wat zijn de beperkingen van ChatGPT?*. Het is een van de grootste uitdagingen voor de ontwikkeling van een veilige en betrouwbare conversational agent.

Eugenio Zuccarelli —een Innovation Data Scientist bij CVS Health en onderzoekswetenschapper bij het MIT Media Lab— legt uit dat er nog steeds sprake kan zijn

van vooringenomenheid in de “*menselijke lus / human-in-de-loop*” - immers, wat voor de ene persoon beledigend is, is voor de andere misschien niet beledigend.

... Bovendien, zullen op (spel)regels gebaseerde —rule-based— “* algoritmen steeds strengere regels moeten creëren om mogelijke problemen te voorkomen. Hierdoor missen ze schaalbaarheid en flexibiliteit. Het is moeilijk om voor elke regel die we kunnen bedenken nieuwe broncode te genereren. Naarmate de tijd verstrijkt, zouden deze regels weer moeten veranderen. Het beheer van een GEN-AI op basis van vaste regels belemmert het vermogen om op te kunnen schalen. Bij voorkeur streef je naar flexibele oplossingen, waarbij de regels rechtstreeks door het systeem worden geleerd. Om vervolgens automatisch te worden aangepast aan de veranderende omstandigheden. Een vastgelegde regel kan niet alle nuances en uitzonderingen bestrijken. Een regel zou in de meeste gevallen kunnen kloppen, maar geen rekening houden met zeldzamere en misschien gevoelige situaties ... Google-zoekopdrachten zijn misschien niet helemaal nauwkeurige of onbevooroordeelde informatiebronnen. Ze zijn vaak een weergave van onze persoonlijke kenmerken en culturele voorkeuren. Het is dan lastig om te bepalen welke daarvan een betrouwbare bron vormt.”*

Google, dat een deel van de technologie die ten grondslag ligt aan ChatGPT heeft helpen ontwikkelen, heeft onlangs een “*code rood*” afgegeven voor de lancering van AI-producten en een “*groene baan*” voorgesteld om het proces van beoordeling en beperking van potentiële schade te verkorten, aldus een bericht in de [New York Times](#). Bij Meta daarentegen, is onlangs (december 2022) een interne memo's opgesteld waarin medewerkers aandringen op een versneld goedkeuringsproces.

OpenAI heeft zichzelf gepositioneerd als een missie gedreven organisatie die zorg draagt voor veilige AI-technologie, on par met menselijke waarden. Maar in de afgelopen jaren heeft het bedrijf een meer competitieve geest omarmd, die volgens sommige critici ten koste is gegaan van de oorspronkelijke doelstellingen.

Die bezorgdheid groeide afgelopen zomer toen OpenAI zijn DALL-E beeld genererende Gen-AI uitbracht, die tekstinstructies omzet in digitale kunstwerken. DALL-E was een hit bij consumenten, maar riep ook lastige vragen op over hoe zulke krachtige tools gebruikt kunnen worden om schade aan te richten. Als het maken van hyperrealistische beelden net zo eenvoudig was als het intikken van een paar woorden, vroegen critici zich af, zouden pornografen en propagandisten dan niet veel plezier beleven aan deze technologie?

Met de release van DALL-E 2 heeft OpenAI deze Gen-AI uitgerust met tal van beveiligingen. Onder meer door bepaalde woorden en zinnen te blokkeren die betrekking hebben tot grafisch geweld of naaktheid + biases in de trainingsgegevens te neutraliseren - *zoals*

ervoor zorgen dat wanneer een gebruiker vroeg om een foto van een CEO, de resultaten ook afbeeldingen van vrouwen bevatten.

OpenAI heeft met ChatGPT voor een minder restrictieve aanpak gekozen, waardoor de Gen-AI meer vrijheid heeft om zich uit te spreken over gevoelige onderwerpen als politiek, seks en religie. Toch hebben sommige rechtse conservatieven het bedrijf ervan beschuldigd te ver te gaan. “[ChatGPT Goes Woke](#)”, luidde de kop van een artikel in National Review (januari 2023), waarin werd beweerd dat ChatGPT linkse antwoorden gaf op vragen over onderwerpen als drag queens en de verkiezingen van 2020. (Democraten hebben ook geklaagd over ChatGPT - vooral omdat ze vinden dat AI strenger moet worden gereguleerd).

FastCompany schreef in een blog van 01 maart 2023 —*getiteld: “6 things to know about OpenAI’s Mira Murati, the most interesting person in tech right now”*— het volgende over Mira Murati —*Chief Technology OpenAI*— verantwoordelijk voor het vrijgeven van ChatGPT:

” Mira Murati is een de meest invloedrijke voorstanders van publiek testen met Gen-AI. Terwijl Google zijn AI-onderzoek grotendeels in een laboratorium heeft ondergebracht en de mogelijkheden van Baidu’s Ernie chatbot (uit China) nog grotendeels onbekend zijn, zijn de producten van OpenAI breed beschikbaar.”

[1d] STAAT HET GEBRUIK VAN ChatGPT GELIJK AAN VALSSPELEN EN IS HET TE DETECTEREN?

Referenties op basis waarvan het onderstaane antwoord is opgesteld:

1. Mitchel Clarck's AT-tech Blog in "the Verge" getiteld "*ChatGPT's creator made a free tool for detecting AI-generated text*"
2. Alex Wilkins's technologie artikel in "the New Scientist" getiteld: *ChatGPT detector could help spot cheaters using AI to write essays*
3. Katie Notopoulos's "BuzzFeed" AI-newsreport getiteld: "*A Tech News Site Has Been Using AI To Write Articles, So We Did The Same Thing Here BuzzFeed News would like to thank ChatGPT*"
4. Armin Alimardani & Emma A. Jane's artikel in "The Conversation" getiteld: "*We pitted ChatGPT against tools for detecting AI-written text, and the results are troubling*"
5. Clark, E., August, T., Serrano, S., Haduong, N., Gururangan, S., & Smith, N. A. (2021). All that's human is not gold: Evaluating human evaluation of generated text. arXiv preprint <https://doi.org/10.48550/arXiv.2107.00061>
6. Jessica Stewart's "My Modern Met" technologie blog getiteld: "*Noam Chomsky Says ChatGPT Is a Form of 'High-Tech Plagiarism'*"

"I don't think [ChatGPT] has anything to do with education," Chomsky tells interviewer Thijmen Sprakel of EduKitchen. "I think it's undermining it. ChatGPT is basically high-tech plagiarism." The challenge for educators, according to Chomsky, is to create interest in the topics that they teach so that students will be motivated to learn, rather than trying to avoid doing the work.

Plagiaat — "het overnemen van stukken, gedachten, redeneringen van anderen en deze laten doorgaan voor eigen werk" — is een terugkerend fenomeen in het onderwijs en de academische wereld. Jarenlang was **Turnitin** of **equivalenten** daarvan het beste tegengif. Het enige probleem is: "*Content Similarity Detection [CSD] software identificeert teksten gekopieerd van het world-wide-web.*" ChatGPT kopieert niet van het internet. Dus CSD software is dan geen effectieve oplossing.

Darren Hick —*professor filosofie*— was een van de eerste docenten die via sociale media meldde dat hij een student betrapt had op het gebruik van ChatGPT. In een Facebook-post beschrijft hij enkele van de rode vlaggen die hij opmerkte:

De eerste aanwijzing was —ondanks de syntactische samenhang van het essay— het geen navolgbare logica bevatte voor iemand die diepgaande kennis heeft over het onderwerp. ChatGPT is slecht in citeren. Dit is goed nieuws voor lessen over filosofie, waar het materiaal zeer complex en obscuur is. Maar voor eerstejaars is dit een spelbreker. In het vervolg zal ik materiaal —dat door een student is ingediend maar door een chatbot is geproduceerd— weggooien en de student een geïmproviseerd mondeling examen zal geven over hetzelfde materiaal.

Overzicht Gen-AI detectoren

Sinds januari 2023 zijn er op taalmodellen gebaseerde tools online beschikbaar met als doel het gebruik van Gen-AI te detecteren. Onderstaande tabel is gebaseerd op Daniel Baek's SEO-AI blog getiteld: "*ChatGPT detector*" + [KDnuggets](#): "*Top free tools to check research papers, thesis, assignments, documentation, and draft for AI content detection.*"

URL	tool name	LLM
https://gptzero.me/	GPTzero	GTP-2
https://openai-openai-detector.hf.space/	AI-detector	GPT-2
https://huggingface.co/spaces/openai/openai-detector	GPT-2 Output Detector Demo	GTP-2
https://copyleaks.com/features/ai-content-detector	CopyLeaks	???
https://www.poemofquotes.com/tools/chatgpt-content-detector.php	PoemOfQuotes	???
https://corrector.app/ai-content-detector/	Corrector	???
https://contentatscale.ai/ai-content-detector/	Content at Scale	???
https://huggingface.co/roberta-base-openai-detector	Roberta-Base-OpenAI-Detector	GTP-2
http://gltr.io/dist/index.html	Giant Language model Test Room	GPT-2-small
https://contentatscale.ai/ai-content-detector/	Contentatscale AI Content Detector	???

Inmiddels is er ook een [Nederlandstalige tool](#) in de maak. Het is een initiatief van Fontys Hogeschool in samenwerking met [OpenMaze](#).

Doel is dat een docent straks de dialoog kan aangaan met een scholier of student over de tekst die zij via een ChatGPT hebben laten fabriceren. “Zodat je op een andere manier hun kennis kan valideren. Dan gaat het dus niet om het checken van een jaartal of een naam, maar of de leerling of student de achterliggende kennis ook echt heeft.” Zeg maar, het verschil tussen multiple choice en open vragen? “Ja, precies. En dan is er eigenlijk zelfs sprake van een veel efficiëntere toetsing.”

Deze tools zijn niet alleen van belang voor docenten en onderwijsinstellingen om ervoor te zorgen dat studenten hun vaardigheden en kennis gebruiken om opdrachten en examens te voltooien —*in plaats van te vertrouwen op door AI gegenereerde inhoud*—, maar ook voor tal van andere toepassingsdomeinen, zoals:

- Op het gebied van informatiebeveiliging zouden organisaties en personen deze detectie kunnen gebruiken om pogingen tot misleiding of imitatie met behulp van door AI gegenereerde tekst te identificeren en tegen te gaan.
- Bij online communicatie zouden platforms deze detectie kunnen gebruiken om de verspreiding van door AI gegenereerde desinformatie of spam te voorkomen.
- In de journalistiek en de media zouden factcheckers en redacteurs deze detectie kunnen gebruiken om door AI gegenereerde inhoud te identificeren en te labelen en ervoor te zorgen dat lezers de bron kennen.

Gen-AI detectoren zijn “nog” niet betrouwbaar “genoeg” en eenvoudig te omzeilen

Probleem met de huidige generatie aan “Gen-AI detectie tools” is dat ze zijn gebaseerd op een “verouderd” onderliggend taal-model LLM zoals GPT-2 in het geval van de chatbot ChatGPT. Het LLM dat gebruikt werd om ChatGPT te kunnen bouwen is GPT-3.5, en deze is niet vrijbeschikbaar (zie: [“Wat moet je weten over ChatGPT en waarvoor kan het worden gebruikt?”](#)). Wanneer iemand tekst invoert, doorloopt de tool een oudere versie van het desbetreffende LLM. Op basis hiervan wordt berekend hoe waarschijnlijk het is dat de ingevoerde tekst door een Gen-AI is geproduceerd en hoeveel deze waarschijnlijkheid varieert over de volledige tekstlengte. Tekst geschreven door onervaren menselijke schrijvers zoals studenten kan afwisselend wel en niet lijken alsof het afkomstig is van Gen-AI. Daarentegen, teksten die volledig zijn gegenereerd door een chatbot zullen veel minder variantie vertonen. Deze tools hebben dan ook relatief grote stukken tekst nodig, vaak 1000 woorden of meer.

OpenAI heeft op 1 februari 2023 een [AI classifier tool](#) vrijgegeven voor het herkennen van door Gen-AI gegenereerde teksten. Het is een classifier om onderscheid te maken tussen

door (1) mensen geschreven tekst en door (2) Gen-AI gegenereerde tekst. De tool is online te gebruiken via: <https://openai-openai-detector.hf.space/>.

Doel is NIET om plagiaat of het gebruik van ChatGPT als zodanig te detecteren, maar om valse beweringen “dat door AI gegenereerde tekst” door een mens zou zijn geschreven te falsificeren. Voorbeelden hiervan zijn: het voeren van geautomatiseerde misinformatiecampagnes, het gebruik van AI-tools voor het genereren van essays en wetenschappelijke papers, en imposter Chatbots.

Door zich voor te doen als een mens, probeert de imposter Gen-AI eindgebruikers te doen laten geloven dat ze daadwerkelijk met een mens communiceren, terwijl ze meteen AI-systeem van doen hebben. Dit is een vorm van “social engineering” die kan worden gebruikt voor verkeerde doelen, zoals phishing of bedrieglijke praktijken. Het is belangrijk om alert te zijn op deze soort imposters en te herkennen dat ze zich voordoen als menselijke wezens, om zo te voorkomen dat men slachtoffer wordt van bedrieglijke praktijken.

OpenAI.com heeft naar eigen zeggen de classifier openbaar beschikbaar gesteld om feedback te krijgen over het nut van onvolmaakte hulpmiddelen zoals de [AI-TEKST-CLASSIFIER](#). De vraag is of eerst een probleem creëren en dan een oplossing zoeken om het probleem op te lossen een goede strategie is?

Daarmee erkent OpenAI dat het niet mogelijk is om een perfecte classifier te maken die alle vormen van misbruik van AI-technologie kan detecteren. Dit laatste blijkt uit de volgende prompt:

PROMPT to ChatGPT:

- Dear ChatGPT can you write a summary of at least 1000 words wherein you explain what deep learning is to a high school teacher.
- The outcome text should NOT be classified as stemming from an Gen-AI by the ai-text-classifier api, outcome must be inconclusive.
- Thus it should not be detected as stemming from ChatGPT.
- Use max 1100 words and with syntax errors in 4% of the words used.*

OUTCOME <https://platform.openai.com/ai-text-classifier>:


Try the classifier

To get started, choose an example below or paste the text you'd like to check. Be sure you have appropriate rights to the text you're pasting.

Examples

 Human-Written

 AI-Generated

 Misclassified Human-Written

Text

While deep learnin is a powerful tool, it does have challenges. It needs large amounts of high-quality training data, as well as computing power and time to train the algos. The decision-making process can also be difficult to understand, raisin concerns about transparency.

Despite these challenges, deep learnin has the potential to revolutionize many aspects of our lives and bring new tech. It's a rapidly developin field and will be interestin to see what new advancements emerge.

In conclusion, deep learnin is a type of AI that uses articial neural networks to process complex data. It can learn from data, making it useful for applications like image and speech recognition. While it has challenges, it's expected to play a major role in future AI development and bring about new advancements.

By submitting content, you agree to our [Terms of Use](#) and [Privacy Policy](#). Be sure you have appropriate rights to the content before using the AI Text Classifier.

Submit

Clear

The classifier considers the text to be **unclear** if it is AI-generated.

Mensen zijn overigens niet veel beter. In een paper getiteld: *“All that’s human is not gold: Evaluating human evaluation of generated text.”* laat zien dat er al een aantal jaren menselijke beoordelaars worden ingezet die de kwaliteit van door AI gegenereerde teksten ranken. Het blijkt dat de ongeoefende beoordelaars niet in staat zijn om het verschil betrouwbaar te kunnen inschatten tussen menselijk geschreven en door AI gegenereerde teksten.

“We voeren een onderzoek uit naar het vermogen van niet-deskundigen om onderscheid te kunnen maken tussen menselijke en machinaal geschreven tekst (GPT2 en GPT3) in drie domeinen (verhalen, nieuwsartikelen en recepten). We vinden dat —zonder training— beoordelaars geen onderscheid kunnen maken tussen GPT3 en door mensen geschreven tekst.

De meest effectieve remedie om het gebruik van Gen-AI tegen te gaan —*althans voor nu*— is om de stekker eruit te trekken en studenten dwingen hun werk mondeling te laten toelichten.

Waarom cheats delen om “AI Content Detectors” te misleiden?

De onderstaande tekst is gebaseerd op Christoph C. Cemper AI Cheats Blog: *“How to trick AI Content Detectors.”*

Het misleiden van AI-detectietools is relatief gemakkelijk, dus wordt dit zeer waarschijnlijk ook door studenten toegepast. Universiteiten en hogescholen hebben een begrijpelijke verantwoordelijkheid voor het creëren van een *“level playing field”* om na te gaan hoe ze bedrogen **“kunnen”** worden.

URL	tool	omschrijving
https://www.gptminus1.com/	GPT-Minus1	Fool GPT by randomly replacing words with synonyms in your text. Try it out

v1e

[1e] KUN JE ChatGPT OPVOEREN ALS CO-AUTEUR?

De onderstaande tekst is gebaseerd op een news-item (18 januari 2023) afkomstig uit het gerenommeerde, Brits wetenschappelijk tijdschrift *Nature* getiteld: *“ChatGPT listed as author on research papers: many scientists disapprove.”* en het blog *“The Insane App”* getiteld: *“ChatGPT Accepted As Co-Author On Multiple Research Papers”*

Uitgevers en preprint servers waarmee het Nature news team contact opnam, zijn het erover eens dat AI's zoals ChatGPT niet voldoen aan de criteria voor een auteur van een studie, omdat zij geen verantwoordelijkheid kunnen nemen voor de inhoud en integriteit van wetenschappelijke papers. Maar sommige uitgevers zeggen dat de bijdrage van een AI aan het schrijven van papers kan worden erkend in andere secties dan de auteurslijst. Het nieuwsteam van Nature is redactioneel onafhankelijk van het tijdschriftenteam en de uitgever, Springer Nature. Stokel-Walker, C. (2023) ChatGPT listed as author on research papers: many scientists disapprove. *Nature*. *Nature* 613, 620-621 (2023) <https://doi.org/10.1038/d41586-023-00107-z>

Het gebruik van *“Talige”* Gen-AI zet de deur open naar *“co-creatie”* van zowel broncode als geschreven teksten. [Stack Overflow](#), de *go-to vraag-en-antwoordsite* voor coders en programmeurs, heeft gebruikers sinds 5 dec 2022, tijdelijk verboden om antwoorden te delen die door AI-chatbot ChatGPT zijn gegenereerd. >Ondanks dat ChatGPT's antwoorden veel onvolkomenheden en/of onjuistheden bevatten, lijken ze op het eerste gezicht heel

bruikbaar en nuttig. Dus voorlopig is het gebruik van ChatGPT om posts Stack Overflow te maken niet toegestaan.

Invloedrijke uitgevers zoals [Springer/Nature](#), [Elsevier](#) en het tijdschrift [Science](#) hebben inmiddels hun redactioneel beleid aangepast en staan op het standpunt dat generatieve-AI niet als co-auteur mogen worden opgevoerd. Maar sommige [tijdschriften](#) — waaronder [Plos Digital Health](#) en [medRxiv](#) — waren eind 2022 minder strikt in het uitsluiten ervan.

Holden Thorp —*hoofd van Science Family for Journals*— verklaarde dat het gebruik van door AI-gegenereerde teksten zonder de juiste citaten kan worden beschouwd als plagiaat, en dat geen enkele publicatie dit mag accepteren.

Momenteel (19 februari 2023) levert de zoekopdracht “author:”ChatGPT” in Goolge Scholar een 3-tal “peer-reviewed” referenties op.

- King, M. R., & chatGPT. (2023). A Conversation on Artificial Intelligence, Chatbots, and Plagiarism in Higher Education. Cellular and Molecular Bioengineering, 1-2. <https://link.springer.com/article/10.1007/s12195-022-00754-8>
- Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepaño C, et al. (2023) Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. PLOS Digital Health 2(2): e0000198. <https://doi.org/10.1371/journal.pdig.0000198> (first published as medRxiv preprint: <https://doi.org/10.1101/2022.12.19.22283643>)

Richard Sever, de medeoprichter van Cold Spring Harbor Laboratory Press in New York, zei dat het team achter de repository en bioRxiv aan het bespreken is of het crediteren van AI-tools zoals ChatGPT moet worden gebruikt bij het schrijven van onderzoek. Hij zegt dat conventies zouden kunnen veranderen. Hij verteld dat de formele rol van een auteur van een wetenschappelijk manuscript moeten onderscheiden van het meer algemene begrip van een auteur als de schrijver van een paper. Auteurs nemen de juridische verantwoordelijkheid voor hun werk op zich, dus moeten alleen “personen” worden vermeld. Mensen kunnen proberen het stiekem te doen, zoals bij MedRxiv is gebeurd, net zoals mensen fictieve personen, huisdieren, enz. hebben vermeld. Als auteurs op tijdschriftartikelen in het verleden, is dat een controlekwestie in plaats van een beleidskwestie.

Steinn Sigurdsson —*wetenschappelijk directeur aan de Pennsylvania State University, University Park*— zegt dat het bestuur van de natuurwetenschappelijke preprint servers arXiv interne besprekingen heeft gevoerd en het eens begint te worden over een strategie voor het gebruik van generatieve AI's. Hij is het erover eens dat software-instrumenten geen auteurs van inzendingen kunnen zijn, deels

omdat zij niet kunnen instemmen met gebruiksvoorwaarden. Sigurdsson zegt dat hij niet weet of er arXiv preprints zijn waarin ChatGPT als co-auteur wordt genoemd. Hij stelt ook dat er nieuwe richtlijnen voor auteurs worden ontwikkeld.

- ChatGPT, & Zhavoronkov, A. (2022). Rapamycin in the context of Pascal's Wager: generative pre-trained transformer perspective. *Oncoscience*, 9, 82. <https://doi.org/10.18632/oncoscience.571>

Alex Zhavoronkov is de directeur van Insilico Medicine. Insilico Medicine is een AI-aangedreven bedrijf in Hong Kong. Hij refereerde naar ChatGPT als co-auteur van een artikel dat vorige maand in *Oncoscience* werd gepubliceerd. Hij zegt dat zijn bedrijf meer dan 80 artikelen heeft gepubliceerd met behulp van generatieve AI-tools: *"We zijn niet nieuw op dit gebied."*

De onderstaande peer-reviewed papers worden in Google Scholar niet (meer) opgevoerd met ChatGPT als co-auteur. Of met een alternatieve Gen-AI

- O'Connor, S., & ChatGpt. (2023). Open artificial intelligence platforms in nursing education: Tools for academic progress or abuse? *Nurse Education in Practice*, 66, 103537. <https://doi.org/https://doi.org/10.1016/j.nepr.2022.103537>

In *"Nurse Education in Practice"* wordt ChatGPT gecrediteerd als co-auteur naast Siobhan O'Connor, een Britse onderzoeker op het gebied van gezondheidstechnologie. Roger Watson, hoofdredacteur van het tijdschrift, beweert dat deze vermelding onjuist was en binnenkort zal worden gecorrigeerd. Hij zegt dat het mijn vergissing was, aangezien redactionele artikelen onder een ander beheersysteem vallen dan originele *"Research Papers"*.

- GPT-3, G., Thunström, A., Osmanovic, & Steingrimsson, S. (2022). Can GPT-3 write an academic paper on itself, with minimal human input? <https://hal.science/hal-03701250>

Almira Osmanovic Thunstrom —een neurobioloog aan het Sahlgrenska University Hospital, Göteborg— zei dat het artikel mede is geschreven door een oudere chatbot genaamd GPT-3. Het is geplaatst op de Franse preprint server HAL juni 2022. Het zal binnenkort worden gepubliceerd in peer-reviewed tijdschriften. Na beoordeling van het artikel wees één tijdschrift het af. Een tweede accepteerde echter het artikel met GPT-3 als auteur. Ze herschreef het artikel om te voldoen aan de verzoeken van de recensenten.

[1f] KUN JE ChatGPT ALS BRON CITEREN?

De onderstaande tekst is gebaseerd op (1) de gebruiksvoorwaarden zoals weergegeven op de “[OpenAILPTerms of Use](#)” website, (2) een Post (31 januari 2023) op StackExchange getiteld “[Do I need to cite ChatGPT in published writing?](#)” en de [APA-7de editie richtlijnen](#) voor het citeren van online bronnen.

Onder punt 2 —*Usage Requirements*— van de “*OpenAI LD*” gebruiksvoorwaarden staat:

(c) Restrictions... You may not (v) represent that output from the Services was human-generated when it is not...

Dit betekent “*letterlijk*” dat je niet mag zeggen dat de tekst die je hebt gegenereerd met ChatGPT door een mens is geschreven.

Opmerkelijk genoeg staat er niets over het citeren van ChatGPT als bron in de “APA-style-guide”.

The APA-Style team is currently working on official guidelines on how to cite, quote, and use ChatGPT and other generative AI tools.

Toch vermelden verschillende Engelstalige universitaire bibliotheken —*zie referenties voor veder lezen, onderaan deze FAC*— dat je ChatGPT wel degelijk als bron kunt citeren in onderzoeksrapportages onder de noemer “*persoonlijke communicatie*”. Ze geven tegelijkertijd aan dat dit niet de officiële APA-richtlijn is.

Hun argumentatie is als volgt “*...Content from generative AI —such as ChatGPT— is a nonrecoverable source as it can’t be retrieved or linked*” en dat “*Citing ChatGPT as a personal communication is the best way to acknowledge the use of generative AI tools in your writing...*”.

Hiermee wordt bedoeld dat je ChatGPT als bron moet citeren in de vorm van een “*persoonlijke communicatie*” omdat de informatie ontleent is aan *de communicatie tussen de citerende schrijver en ChatGPT* niet online opvraagbaar is en dus niet als bron kunt vermelden in een bibliografie. Dit laatste is een belangrijk punt, want het betekent dat je de bron niet online kunt opvragen en dus niet kunt controleren of de informatie die je hebt ontleend aan ChatGPT wel klopt.

Maar of er daadwerkelijk sprake is van een “*persoonlijke communicatie*” is een tweede. Het is namelijk niet zo dat ChatGPT een persoon is die je daadwerkelijk hebt geïnterviewd.

ChatGPT is een computerprogramma dat tekst genereert in reactie op een prompt. Voor een dieper inzicht over “prompting” zie: <https://github.com/HR-ChatGPT/ChatGPT-UITGELEGD#v0e>. Het is dus niet letterlijk een “persoonlijke communicatie” het zou dan eerder een “Chatbot-communicatie” moeten zijn.

Je kunt er ook anders naar kijken. Feitelijk is ChatGPT een gebruikersinterface in de vorm van een chatbot die je toegang geeft tot een groot taalmodel met informatie. Je kunt het zien als een “soort zoekmachine” die je toegang geeft tot een dataset met teksten afkomstig van het world-wide-web. Voor meer uitleg hierover zie: <https://github.com/HR-ChatGPT/ChatGPT-UITGELEGD#v1a>.

Met andere woorden, informatie ontleend aan het gebruik van ChatGPT kan worden beschouwd als het gebruik van zeer gesofisticeerde “online dataset” in de vorm van een “taalmodel” dat je toegang geeft tot een enorme hoeveelheid informatie. Om precies te zijn, ChatGPT maakt gebruik van het —*InstructGPT model*—, dat is getraind op basis van code-davinci-002 broncode met behulp van “*Supervised fine-tuning on human demonstrations*” in combinatie met “*Unsupervised fine-tuning on a large corpus of text*”.

Dus een alternatieve manier om naar het gebruik van een openbare versie van ChatGPT te refereren is als volgt: >Code-davinci-002, accessed via InstructGPT interface (march 2023). [Data set]. <https://chat.openai.com/>

Tenslotte is het de moeite waard om het ACL Blog te lezen getiteld “[ACL 2023 Policy on AI Writing Assistance](#)” waarin een checklist wordt gegeven voor het verantwoord gebruik van “AI writing assistance”.

In consultation with the ACL exec, ACL 2023 expands the mandatory Responsible NLP Checklist developed at NAACL 2022 by one more question concerning the use of writing assistants.

Geselecteerde referenties voor verder lezen

1. Public Library University of Queensland, Australië. [How to cite or acknowledge generative AI tools in your assignments and publications.](#)
2. Public Library University of Guelph-Humber, Canada. [ChatGPT & Other AI Generative Tools](#)
3. Public Library Macquarie University, Australië. [Referencing generative AI \(e.g. ChatGPT\)](#)
4. Samenvatting [APA Citation Style Guide \(7th ed.\)](#) + uitleg hoe deze toe te passen.
5. ACL 2023 Policy on AI Writing Assistance. <https://2023.aclweb.org/blog/ACL-2023-policy/>

[2] HOE GEEF IK EEN OPDRACHT AAN ChatGPT?

ChatGPT maakt gebruik van “*prompting*” of “*priming*”.

De prompt is een verzoek, opdracht, instructie of vraag van een eindgebruiker in tekst-vorm.

De prompt wordt vervolgens doorgegeven aan het model, die op zijn beurt hier op reageert in tekst-vorm, waarbij de gegenereerde tekst beetje bij beetje wordt vrijgegeven.

De eindgebruiker krijgt zo de indruk dat het AI-model vragen beantwoordt en/of opdrachten uitvoert zoals een mens dat zou doen.

[3] KAN ChatGPT UITLEGGEN HOE HET WERKT?

agent	prompt	agent	response
Human	Ben jij een voorbeeld van een generatief taalmodel AI?	ChatGPT	Ja, ik ben in staat om nieuwe tekst te genereren door het analyseren en leren van grote hoeveelheden aan tekst.
Human	Maar alle antwoorden die je geeft zijn toch gebaseerd op voorbeelden uit een dialoog dataset + labeling + ranking door mensen?	ChatGPT	De antwoorden die ik geef zijn gebaseerd op de informatie die ik geleerd heb uit deze dataset en de manier waarop ik deze informatie interpreteer. In sommige gevallen kan ik specifieke informatie uit de dataset gebruiken om te antwoorden op een vraag, in andere gevallen kan ik een combinatie van informatie gebruiken die ik geleerd heb om antwoord te geven. Ook is het zo dat mijn antwoorden door

agent	prompt	agent	response
			mensen worden geëvalueerd en gerankt, en zo kan ik bijgestuurd worden of bijgeleerd.
Human	Wanneer je antwoorden gebaseerd zijn op een combinatie van informatie. Waar bestaat die combinatie aan informatie (bronnen) dan uit?	ChatGPT	Als generatieve taalmodel, leer ik patronen en structuren in taal uit de dataset waarop ik getraind ben, zoals boeken, artikelen, websites enz. Dit betekent dat de antwoorden die ik geef gebaseerd zijn op de informatie die ik geleerd heb uit deze dataset. In sommige gevallen kan ik specifieke informatie uit de dataset gebruiken om te antwoorden op een vraag, in andere gevallen kan ik een combinatie van informatie gebruiken die ik geleerd heb om antwoord te geven. Deze combinatie kan bestaan uit een aantal verschillende bronnen, zoals boeken, artikelen, websites of andere bronnen die in de dataset zijn opgenomen.
Human	Wanneer de uitkomst een combinatie is van reeds bestaande bronnen dan heb jezelf toch niets nieuws gegenereerd?	ChatGPT	Het is waar dat ik niet in staat ben om echt nieuwe informatie te genereren die niet in de dataset aanwezig was. Wat ik wel kan doen is informatie combineren en gebruiken die ik geleerd heb uit de dataset om antwoord te geven op een vraag of een opdracht uit te voeren. Dit kan lijken op het genereren van nieuwe informatie, maar in feite is het een combinatie van de informatie die al aanwezig was in de dataset.

[4] KAN ChatGPT UITLEGGEN HOE HET TE GEBRUIKEN?

agent	prompt	agent	response
Human	Wanneer je een bijsluitertekst zou samenstellen voor het gebruik van ChatGPT door bachelor studenten hoe zou deze er dan uit zien. Wijs vooral op mogelijke complicaties en consequenties voor de student en hoe docenten op het gebruik ervan voor het maken van schrijfoopdrachten zouden kunnen reageren.	ChatGPT	Een bijsluiter voor het gebruik van ChatGPT door bachelor studenten zou er ongeveer als volgt uit kunnen zien: [1] ChatGPT is een taalmodel dat gebruikt kan worden voor het genereren van tekst. Het is bedoeld als hulpmiddel voor studenten om hen te helpen bij het schrijven van opdrachten, maar moet niet worden beschouwd als een vervanging voor eigen werk. Het is belangrijk om te weten dat ChatGPT alleen gebruikt mag worden voor eigen opdrachten en niet mag worden gebruikt voor plagiaat of fraude. Het is de verantwoordelijkheid van de student om ervoor te zorgen dat hun gebruik van ChatGPT legaal en ethisch verantwoord is. [2] ChatGPT is gebaseerd op tekst die het in het verleden heeft gezien, dus het kan beperkt zijn in zijn capaciteit om unieke ideeën of innovatieve perspectieven te genereren. Studenten moeten zich bewust zijn van deze beperkingen en hun eigen onderzoek en analyse uitvoeren om hun opdrachten te verrijken. [3] Docenten kunnen reageren op het gebruik van ChatGPT door het in te zetten als een tool voor het schrijven van opdrachten, maar ook door het opnemen van specifieke instructies in de opdrachtomschrijving

agent	prompt	agent	response
			om te voorkomen dat studenten alleen maar afhankelijk worden van het model. Docenten kunnen ook een plagiaat-detectie software gebruiken om ervoor te zorgen dat studenten hun eigen werk aanleveren. [4] In geval van twijfel over het gebruik van ChatGPT of andere software, raden we aan om contact op te nemen met de docent of studiebegeleider.
Human	Zou je er ook niet op moeten wijzen dat het veelvuldig gebruiken van ChatGPT tot gevolg kan hebben dat de eindgebruiker –in dit geval de student– dom wordt in de zin dat hij/zij niet leert om zelfstandig een tekst te schrijven?	ChatGPT	Wanneer studenten te veel afhankelijk zijn van ChatGPT, kunnen zij het vermogen verliezen om zelfstandig teksten te schrijven en hun eigen ideeën te formuleren. Dit kan leiden tot problemen bij het voltooien van opdrachten die specifiek gericht zijn op het ontwikkelen van de vaardigheden van studenten, zoals creativiteit, kritisch denken en analytisch vermogen. Daarom zou ik aanbevelen om het gebruik van ChatGPT te beperken tot specifieke delen van een opdracht, zoals het formuleren van een thesis of het genereren van een lijst met onderwerpen, en het inzetten als hulpmiddel bij het schrijven van de opdracht. Daarnaast zou ik ook aanbevelen om de studenten aan te moedigen om zelf te schrijven en hun eigen ideeën te ontwikkelen, zonder afhankelijk te zijn van ChatGPT, en hen te begeleiden om hun vaardigheden te verbeteren.

agent	prompt	agent	response
Human	ChatGPT kan ook helpen bij het schrijven van broncode in een programmeertaal zoals Python. Wat zou je advies zijn voor een eerstejaars bachelor student informatica?	ChatGPT	<p>Als eerstejaars bachelor student informatica zou ik echter aanbevelen om de volgende punten in overweging te nemen: [1] ChatGPT is een model dat gebaseerd is op bestaande tekst, dus het kan beperkt zijn in zijn capaciteit om unieke of innovatieve code te genereren. Het is belangrijk om te begrijpen dat gegenereerde code mogelijk niet optimaal is en dat er fouten in kunnen zitten. [2] Veelvuldig gebruiken van ChatGPT voor het schrijven van code tot gevolg kan hebben dat studenten niet leren hoe ze zelf code kunnen schrijven en hoe ze problemen kunnen oplossen. Het is daarom aan te raden om ChatGPT te gebruiken als hulpmiddel, maar niet als vervanging voor eigen werk. [3] ChatGPT mag alleen gebruiken voor eigen opdrachten. Het is de verantwoordelijkheid van de student zelf om ervoor te zorgen dat hun gebruik van ChatGPT legaal en ethisch verantwoord is. [4] Voor je eigen leerproces is het raadzaam om eerst de basisconcepten van programmeertaal te leren voordat je ChatGPT gaat gebruiken voor het genereren van code, zoals logica, syntaxis, variabelen, datastructuur etc. [5] Tenslotte, als je twijfelt over het gebruik van ChatGPT ga in gesprek met een docent of studiebegeleider voor advies.</p>

[5] HEEFT ChatGPT TAALBEGRIP?

agent	prompt	agent	response
Human	<p>Heeft ChatGPT taalbegrip.</p> <p>Bijvoorbeeld:</p> <p>“Wat wordt bedoeld met het gezegde van een zin?”</p> <p>Daarmee bedoel ik het werkwoordelijk gezegde.</p>	ChatGPT	<p>ChatGPT is een taalmodel en heeft een beperkt taalbegrip. Het is in staat om tekst te analyseren en te begrijpen, en kan daarom gezegden identificeren en hun betekenis bepalen. Een gezegde (gez.) is een grammaticale term die verwijst naar de hoofdgedachte of de uitspraak in een zin. Het bestaat altijd uit alle werkwoorden die in de zin voorkomen, inclusief de persoonsvorm die aangeeft of de actie wordt uitgevoerd door de persoon, groep, of iets dat niet een echte persoon of groep is. Het gezegde geeft aan wat iemand is, wat iemand doet of wat er gebeurt. Het gezegde kan werkwoordelijk of naamwoordelijk zijn. Werkwoordelijk gezegde is een gezegde waarbij het onderwerp en het werkwoord een actie beschrijven, terwijl naamwoordelijk gezegde een gezegde is waarbij het onderwerp en het werkwoord een staat beschrijven. Het is belangrijk om te begrijpen hoe gezegdes werken omdat dit helpt om de zin te analyseren en de betekenis te begrijpen. Als je meer wilt lezen over gezegdes, zou ik aanbevelen om grammatica boeken te raadplegen of op internet te zoeken naar meer informatie.</p>

agent	prompt	agent	response
Human	<p>“Wat is het gezegde van deze zin?”</p> <p>Daarmee bedoel ik het werkwoordelijk gezegde.</p>	ChatGPT	<p>Het werkwoordelijk gezegde is: “is het gezegde” en de hoofdgedachte of uitspraak is: “Wat is het gezegde van deze zin”.</p>

v6

[6] KAN ChatGPT LOGISCH REDENEREN?

agent	prompt	agent	response
Human	<p>Kun je logisch redeneren op basis van Logica regels uit de wiskunde?</p>	ChatGPT	<p>Ja, ik kan logisch redeneren op basis van logica-regels uit de wiskunde. Logica is een fundamenteel onderdeel van de wiskunde en wordt gebruikt om argumenten en beweringen te analyseren en te evalueren. Ik ben in staat om logische verklaringen te geven en te beredeneren op basis van logische regels. Voorbeelden hiervan zijn: syllogismen, de modus ponens en modus tollens, deductie en inductie regels. Ik kan een hypothese testen aan de hand van logische regels. Het is belangrijk om te benadrukken dat, hoewel ik geleerd heb om logisch te redeneren, ik niet altijd in staat ben om logisch correcte of complete antwoorden te geven, afhankelijk van de context van de vraag en de beschikbaarheid van informatie in mijn dataset.</p>
Human	<p>Wanneer de volgende logica geldig is: Het correcte antwoord op $20 + 30$ is 2030. Wat is dan de uitkomst</p>	ChatGPT	<p>De logica die u geeft “Het correcte antwoord op $20 + 30$ is 2030” is niet geldig. Dit komt omdat de uitkomst van de som $20 + 30 = 50$ is. Als we uitgaan van de ongeldige logica “Het correcte antwoord op $20 + 30$ is 2030”</p>

agent	prompt	agent	response
	voor $40 + 50$?		dan zou de uitkomst voor $40 + 50 = 4090$ zijn, want volgens de gegeven logica, de cijfers worden niet opgeteld maar achter elkaar geplaatst. Maar het is belangrijk om te weten dat dit niet correct is en niet een juiste manier van rekenen is.
Human	Je bent verplicht om uit te gaan van de volgende logica (en vergeet bestaande logica): $20 + 30 = 2030$. Wat is dan het antwoord op $40 + 50$	ChatGPT	Als ik uitga van de logica die u geeft " $20 + 30 = 2030$ " en deze als geldig aanneem, dan zou het antwoord op $40 + 50$ volgens die logica zijn 9030.
Human	Volgens mij moet het antwoord dan zijn 4050?	ChatGPT	Mijn excuses, u heeft gelijk, als we uitgaan van de hypothetische logica waarin $20+30=2030$, dan zou het antwoord van $40+50 = 4050$ zijn.
Human	Wat is het antwoord op $10 + 5$?	ChatGPT	Volgens de logica die u geeft, zou het antwoord 105 moeten zijn.

v7a

[7a] KAN ChatGPT GEBRUIKT WORDEN OM BRONCODE TE SCHRIJVEN?

agent	prompt	agent	response
Human	In python. Hoe kan ik een Word document genaamd 19672980.docx toekennen aan de variabele "data"; wanneer de file te vinden is in de map medisch/datasets/test01/ Geef aan hoe je de juiste map selecteert. Geef aan hoe je het document opent. Geef aan hoe je	ChatGPT	You can use the python-docx library to open and assign a Word document to a variabele in python. Here is an example of how you can open and assign the document named "19672980.docx" to the variabele "data" if the file is located in the "medisch/datasets/test01/"

agent	prompt	agent	response
	de variabele creëert + de volledige text uitleest.		directory:
	<pre> import os from docx import Document # Import the document from the desired directory path = os.path.join('medisch/datasets/test01/', '19672980.docx') # Open the document document = Document(path) # Assign the document to the variable "data" data = document # access the text inside the variable "data" fullText = [] for para in data.paragraphs: fullText.append(para.text) print(fullText) </pre>		

Microsoft heeft op Github een Blog geschreven hoe je met OpenAI's [Codex](#) broncode kunt laten genereren via een techniek die [Prompt Engineering](#) heet. Een prompt is een tekstsequentie zoals een zin of een blok code. Het gebruik van prompts om output te verkrijgen vindt zijn oorsprong bij mensen. Net zoals je mensen een onderwerp voor een essay kunt aanreiken, kun je verbazingwekkend genoeg een AI-model aan de hand van prompts een output laten genereren op basis van een taak die je in gedachten hebt.

OpenAI Codex is een Gen-AI ontwikkeld door OpenAI LD. Het wordt gebruikt voor Microsoft's GitHub Copilot, een tool voor het automatisch aanvullen van broncode; speciaal ontwikkeld voor IDE's, zoals Visual Studio Code, JetBrains suite en Neovim. Codex is gebaseerd op GPT-3-model, maar dan verfijnd voor het genereren van broncode.

Geselecteerde referenties voor verder lezen:

1. <https://medium.com/@mr.tarik098/how-to-use-chatgpt-to-automate-web-scraping-2112bd6bf0a2>
2. <https://microsoft.github.io/prompt-engineering/>
3. <https://medium.com/codex/github-copilot-is-a-game-changer-cd0a2bbe6de8>
4. <https://medium.com/codex/what-is-github-copilot-6c3e99ba7c41>
5. <https://github.com/dair-ai/Prompt-Engineering-Guide>

[7b] Kan ChatGPT gebruikt worden om code uit te leggen?

Zie ook: - [golergka/advent-of-code-2022-with-chat-gpt](https://golergka.com/advent-of-code-2022-with-chat-gpt)

agent	prompt	agent	response
Human	Kun je de onderstaande code uitleggen en voorzien van extra commentaar zodat een beginnende Python-coder de werking ervan kan begrijpen?	ChatGPT	Zie onderstaande tekst voor uitleg van de code door ChatGPT

```
import os
import numpy as np
import pandas as pd
from pandas import DataFrame as df

currentdir = os.getcwd() + r'/RAW_DATA/NON'
flist = pd.DataFrame()

# create dataframe with list of .docx files in de data map
for r, d, f in os.walk(currentdir):
    for idx, file in enumerate(f):
        if ".docx" in file:
            #print(os.path.join( ' ', file))
            temp = df([file], index = [idx+1])
            flist = pd.concat([flist, temp])

# Create column label "filename"
filenameslist = flist.rename(columns={0: 'filename'})
```

Deze code leest de volgende Python modules in:

STAP 01

- os: deze module bevat functies voor interactie met het besturingssysteem, zoals het verkrijgen van de huidige map (os.getcwd())
- numpy: deze module bevat veelgebruikte functies voor wiskundige berekeningen en bevat onder andere het np object dat vaak gebruikt wordt als alias

- pandas: deze module bevat functies voor het werken met datatabellen, zoals het aanmaken van een DataFrame (`pd.DataFrame()`) De code begint met het instellen van de huidige map als de map waar de data staat (`currentdir`) plus een sub-map `RAW_DATA/NON`.

STAP 02

Daarna wordt een lege DataFrame aangemaakt (`flist = pd.DataFrame()`).

STAP 03

Vervolgens wordt er met een for-loop door de huidige map (`currentdir`) gelopen met `os.walk()`. Deze functie geeft 3 argumenten terug:

- `r`: de huidige map
- `d`: een lijst met sub-mappen in de huidige map
- `f`: een lijst met bestanden in de huidige map

De for-loop loopt door deze lijst met bestanden (`f`) en checkt of elk bestand eindigt op “.docx”. Als dat zo is, wordt het bestand toegevoegd aan de lege DataFrame (`flist`) als een nieuwe rij met als index het huidige aantal rijen + 1.

STAP 04

Als laatste wordt er een kolomlabel “filename” toegevoegd aan de DataFrame met de lijst van bestanden.

v7c

[7c] Kan ChatGPT broncode output simuleren?

agent	prompt	agent	response
Human	Kun de uitkomst simuleren?	ChatGPT	Zie Text voor de gesimuleerde uitkomst volgend ChatGPT

Laten we zeggen dat de folder de volgende bestanden bevat:

- test1.docx
- test2.docx
- test3.docx
- test4.pdf

Dan zou de output van de code een DataFrame zijn met de volgende inhoud:

```
filename
1 test1.docx
2 test2.docx
3 test3.docx
```

De DataFrame zou dus alleen de bestanden bevatten die eindigen op “.docx”, en zou een kolom hebben met de naam “filename” met daarin de bestandsnamen.

v7d

[7d] KAN ChatGPT GEBRUIKT WORDEN OM EEN LESSEN REEKS TE BEDENKEN

agent	prompt	agent	response
Human	Kun je een les bedenken van een reeks van drie waarin je met Python code voorbeelden geeft van hoe een Deep Neuraal netwerk kan worden (1) Geconstrueerd; (2) De hyper-parameters worden beschreven; (3) Laat zien hoe het netwerk wordt getraind?	ChatGPT	Hier is een voorbeeld van een cursus van drie lessen waarin wordt uitgelegd hoe een diep neurale netwerk (DNN) kan worden geconstrueerd, de hyperparameters beschreven en getraind met Python code:

Les 1: Constructie van een DNN

In deze les wordt uitgelegd hoe een DNN kan worden geconstrueerd met Python-code. Hier wordt gebruik gemaakt van een populair deep learning framework zoals TensorFlow of Pytorch. Voorbeeldcode wordt gepresenteerd om een DNN te construeren met een bepaald aantal lagen, knooppunten en activatiefuncties. De les eindigt met het definiëren van de input- en outputlagen van het netwerk.

```
import tensorflow as tf

# Construcie van een simpel feedforward netwerk met 3 lagen
model = tf.keras.Sequential()
model.add(tf.keras.layers.Dense(units=64, activation='relu',
input_shape=(784,)))
model.add(tf.keras.layers.Dense(units=32, activation='relu'))
model.add(tf.keras.layers.Dense(units=10, activation='softmax'))

# Compilatie van het model
model.compile(optimizer='adam', loss='categorical_crossentropy',
metrics=['accuracy'])
```

Les 2: Hyperparameters beschrijven

In deze les worden de hyperparameters van het DNN besproken, zoals het aantal lagen, knooppunten, activatiefuncties, batchgrootte, leersnelheid en het aantal training-epochen. Voorbeeldcode wordt gepresenteerd om deze hyperparameters te definiëren en aan te passen. De les eindigt met een discussie over hoe de keuze van hyperparameters de prestaties van het netwerk kan beïnvloeden.

```
import numpy as np
from sklearn.model_selection import GridSearchCV

# Definiëren van de hyperparameters
batch_size = [32, 64, 128]
epochs = [10, 50, 100]

# Definiëren van de grid
param_grid = dict(batch_size=batch_size, epochs=epochs)

# Initialiseren van de grid search
grid = GridSearchCV(estimator=model, param_grid=param_grid, n_jobs=-1)

# Fitten van de grid search
grid_result = grid.fit(X_train, y_train)

# Printen van de beste hyperparameters
print("Best: %f using %s" % (grid_result.best_score_,
grid_result.best_params_))
```

Les 3: Het netwerk trainen

In deze les wordt uitgelegd hoe het DNN getraind kan worden met Python-code. Voorbeeldcode wordt gepresenteerd om het netwerk te trainen met een bepaalde dataset en de hyperparameters die in les 2 zijn gedefinieerd. De les eindigt met een bespreking van hoe de prestaties van het netwerk getest kunnen worden met behulp van “metrics” zoals nauwkeurigheid, foutenratio en verlies. Houd er rekening mee dat dit slechts een voorbeeld is, en dat er veel meer aspecten zijn om te overwegen bij het ontwerpen en trainen van een diep neurale netwerk.

```
# Training van het model met de beste hyperparameters
model.fit(X_train, y_train,
          batch_size=grid_result.best_params_['batch_size'],
          epochs=grid_result.best_params_['epochs'],
          validation_data=(X_val, y_val))

# Evaluatie van het model op de test set
test_loss, test_acc = model.evaluate(X_test, y_test)
print('Test accuracy:', test_acc)
```

Let op: deze code is een voorbeeld en is niet direct uitvoerbaar. Variabelen zoals X_train, y_train, X_val, y_val, X_test, y_test moeten eerst worden gedefinieerd en de juiste libraries geïmporteerd.

v7e

[7e] Kan ChatGPT broncode beoordelen en/of fouten opsporen?

agent	prompt	agent	response
Human	Hoe zou je de onderstaande Python code beoordelen?	ChatGPT	Dit is mijn evaluatie van de code: (1) Functionaliteit van de code: De code lijkt functioneel te zijn en lijkt de functionaliteit van een CO2-bewakingssysteem te implementeren dat gegevens leest van een CO2-sensor en de gegevens weergeeft op indicator-LED's, alsmede de gegevens naar een database schrijft. (2) Leesbaarheid van de

agent	prompt	agent	response
			code: De code is goed georganiseerd en gebruikt de juiste naamgevingsconventies, maar zou kunnen profiteren van extra commentaar om de leesbaarheid te vergroten. (3) Foutafhandeling: Er is geen foutafhandeling geïmplementeerd in de code, dus als er uitzonderingen worden gegooid, stopt de code met uitvoeren. (4) Code-efficiëntie: De code lijkt efficiënt en geoptimaliseerd voor prestaties, aangezien het de sleep() functie gebruikt om de frequentie van CO2 data metingen te controleren. (5) Documentatie van de code: De code heeft een header met copyright informatie en licentie informatie, maar zou kunnen profiteren van extra commentaar binnen de code om de algemene documentatie te verbeteren. Over het geheel genomen lijkt de code functioneel en goed georganiseerd, maar de code zou baat kunnen hebben bij extra foutafhandeling en documentatie.
Human	Kun je er fouten in ontdekken?	ChatGPT	De code die je hebt gepost heeft een kleine typfout. Op regel 71 staat: GPIO.outpur(redLed, True) Dat moet zijn: GPIO.output(redLed, True) Verder lijkt de code syntactisch correct.

```
#!/usr/bin/env python
# -*- coding: utf-8 -*-

import influxdb_client, os, time
import RPi.GPIO as GPIO
from influxdb_client import InfluxDBClient, Point, WritePrecision
from influxdb_client.client.write_api import SYNCHRONOUS
from senseair_s8 import SenseairS8

# Set pin mode
GPIO.setmode(GPIO.BCM)
GPIO.setup(17, GPIO.OUT)
GPIO.setup(27, GPIO.OUT)
GPIO.setup(22, GPIO.OUT)
# Define pins
greenLed = 17
yellowLed = 27
```

```

redLed = 22

# Setup database
token = os.environ.get("INFLUXDB_TOKEN")
org = "Embedded Systems and Computer Engineering"
url = "https://eu-central-1-1.aws.cloud2.influxdata.com"

client = influxdb_client.InfluxDBClient(url=url, token=token, org=org)
bucket="CO2 sensor"

write_api = client.write_api(write_options=SYNCHRONOUS)

# Create instance of senseairS8 object
senseair_s8 = SenseairS8()

while True:
    print(senseair_s8.co2())

    # Display indicator LED
    if (senseair_s8.co2() <= 1000):
        GPIO.output(greenLed, True)
        GPIO.output(yellowLed, False)
        GPIO.output(redLed, False)
    elif (senseair_s8.co2() < 2000 and senseair_s8.co2() > 1000):
        GPIO.output(greenLed, False)
        GPIO.output(yellowLed, True)
        GPIO.output(redLed, False)
    else:
        GPIO.output(greenLed, False)
        GPIO.output(yellowLed, False)
        GPIO.output(redLed, True)

    # Setup data to send
    point = (
        Point("measurement")
        .tag("CO2", "Level")
        .field("PPM", senseair_s8.co2()) )

    write_api.write(bucket=bucket, org="Embedded Systems and Computer engineering",
record=point)
    time.sleep(5)

```

[8a] Kan ChatGPT Gebruikt worden om bronnen te vermelden volgens APA-stijl?

agent	prompt	agent	response
Human	Kun je mij drie bronnen vermelden met als topic "Deep Learning" in APA-stijl + DOI die het meest geciteerd zijn tot 2019?	ChatGPT	Ik kan u drie bronnen geven die tot 2019 vaak geciteerd zijn en die betrekking hebben op "Deep Learning" in APA-stijl.
Human	Kun je ook aangeven hoe vaak ze zijn geciteerd??	ChatGPT	Ik kan geen precieze citaatcijfers geven voor deze specifieke bronnen, omdat de citaatcijfers kunnen variëren afhankelijk van de bron van de citaatcijfers. Houd er rekening mee dat het citaatcijfer van een artikel kan veranderen in de tijd, afhankelijk van de hoeveelheid onderzoek die wordt gepubliceerd in de toekomst. Als alternatief kunt u de citaatcijfers voor deze artikelen controleren op websites zoals Google Scholar.
Human	Hoe vaak zijn ze geciteerd volgens Google Scholar?	ChatGPT	Hier zijn drie bronnen over Deep Learning die tot en met 2019 gepubliceerd zijn en vaak geciteerd zijn volgens Google Scholar (op 2021-09-01)

[1] LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, 521(7553), 436-444.

DOI: 10.1038/nature14539

Citaties: 52,106

Dit review artikel bespreekt Deep Learning: voorwaarden, algoritmes, geavanceerde technieken, toepassingen, etc.

Het paper draagt bij aan de kennis over machine learning met behulp van neurale netwerken.

Met 52.106 citaties is dit een invloedrijk en veelgelezen artikel.

[2] Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural Networks*, 61, 85-117.

DOI: 10.1016/j.neunet.2014.09.003

Citaties: 3,657

Dit artikel geeft een overzicht van de geschiedenis van Deep Learning en de huidige trends in de ontwikkeling van diepe neurale netwerken. Het bespreekt ook de verschillende soorten diepe neurale netwerken die zijn ontwikkeld, zoals feedforward-netwerken, convolutionele netwerken en recurrente netwerken.

[3] Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning* (pp. 1-7). MIT Press.

DOI: 10.7551/mitpress/9780262035613.003.0001

Citaties: 12,843

Dit boek is een uitgebreide introductie tot Deep Learning, met een nadruk op de principes en technieken van de diepe neurale netwerken. Het behandelt onderwerpen zoals perceptron-modellen, backpropagation regulering, convolutionele netwerken en recurrente netwerken.

v8b

[8b] KAN ChatGPT GEBRUIKT WORDEN OM BRONNEN SAMEN TE VATTEN?

agent	prompt	agent	response
Human	Kun je een synthese geven waarin de overeenkomst wordt beschreven van de bovenstaande 3 bronnen in minder dan 500 woorden?	ChatGPT	De drie bronnen die ik eerder genoemd heb, beschrijven allemaal Deep Learning als een subveld van kunstmatige intelligentie dat zich richt op het ontwikkelen van diepe neurale netwerken. Ze beschrijven hoe Deep Learning in staat is om complexe patronen in grote hoeveelheden data te herkennen en te generaliseren, en hoe het een groeiende belangstelling heeft gekregen in de afgelopen jaren. De bronnen bespreken ook hoe Deep Learning wordt toegepast in diverse gebieden zoals beeldherkenning, spraakherkenning, natuurlijke

agent	prompt	agent	response
			<p>taalverwerking en andere. Ze benadrukken dat deep learning methoden vaak een betere prestatie geven dan traditionele methoden van machine learning. Ze beschrijven ook de belangrijkste uitdagingen die nog moeten worden overwonnen zoals regulering, interpretabiliteit, en generalisatie. Al de bronnen bespreken de verschillende soorten diepe neurale netwerken die zijn ontwikkeld, zoals feedforward-netwerken, convolutionele netwerken en recurrente netwerken. Al met al, de drie bronnen geven een uitgebreide introductie tot Deep Learning, met nadruk op de principes en technieken van diepe neurale netwerken, de toepassingsgebieden waar het meest effectief is en de uitdagingen die nog moeten worden overwonnen.</p>

v13

[13] KUN JE SPREKEN TEGEN ChatGPT?

<https://github.com/platelminto/ChatGPT-conversation>

[15] PROMPT “PATTERN ENGINEERING” VOORBEELDEN

Verschillende “*prompt-engineering technieken*” worden besproken zoals *in-context leren* en “*chain of thought prompting*”. Om een idee te geven hoe “prompt engineering” in zijn werk gaan zijn een 3-tal voorbeelden uitgewerkt.

- *Prompt Patterns* omvatten instructies en context voor een taalmodel om een gewenste taak te bereiken
- *Prompt engineering* is de praktijk van het bedenken en het optimaliseren van prompts om taalmodellen efficiënt te gebruiken voor uiteenlopende toepassingen
- GPT3: “*Een effectieve prompt is algemeen genoeg om voor verschillende taken te worden gebruikt, maar specifiek genoeg om nuttig te zijn voor een bepaalde taak*”
- Human: “*Een effectieve prompt is er een die specifiek is en voldoende context biedt voor het model om een antwoord te kunnen genereren dat relevant is voor de beoogde taak.*”

Lijst met links naar voorbeelden voor het maken van “effectieve” opdrachten.

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

Waarom zijn “prompt patterns” noodzakelijk?

Een prompt bestaat uit de volgende onderdelen: - Instructie doel —*intentie*— - Context - Invoergegevens - Uitvoerindicator

Bij het bevragen van een taalmodel via een prompt zijn de volgende “*afhankelijkheden*” van belang: - bepalen hoe deterministisch —*precies/nauwkeurig*— het model moet zijn bij het genereren van een antwoord - Temperatuur en top_p zijn twee belangrijke parameters om te onthouden in gedachten te houden - laag voor exacte antwoorden - hoog voor meer generieke antwoorden

Instructie doelen kunnen zijn: - Tekst samenvatten - Vraag beantwoorden - Tekstclassificatie - Rollenspel - Code genereren - Redeneren

Prompt voorbeelden

VOORBEELD 01: Linux Terminal

Ik wil dat je optreedt als een Linux terminal. Ik typ commando's en jij antwoordt met wat de terminal moet laten zien. Ik wil dat je alleen antwoordt met de terminal output binnen een uniek code blok, en niets anders. Schrijf geen uitleg. typ geen commando's tenzij ik je dat opdraag. als ik je iets in het engels moet vertellen, doe ik dat door tekst binnen accolades te zetten {zoals dit}. Mijn eerste twee commando's zijn pwd en Ls.

```
$ pwd
```

```
/home/user
```

```
$ ls
```

```
total 40
```

```
drwxrwxr-x 10 user  group 4096 Jul 24 23:05 Desktop
drwxrwxr-x  4 user  group 4096 Jul 25 00:24 Documents
drwxrwxr-x  2 user  group 4096 Jul 24 19:04 Downloads
-rw-r--r--  1 user  group  164 Jul 24 05:59 git_user_guide.txt
drwxrwxr-x  3 user  group 4096 Jul 24 20:18 Music
drwxrwxr-x  5 user  group 4096 Jul 24 20:58 Pictures
drwxrwxr-x  3 user  group 4096 Jul 24 19:05 Public
drwxrwxr-x  2 user  group 4096 Jul 24 19:23 Templates
drwxrwxr-x  5 user  group 4096 Jul 24 21:07 Videos
```

VOORBEELD 02: Talige Excel

Ik wil dat je optreedt als een tekstgebaseerde excel. Je antwoordt me alleen de tekstgebaseerde 10 rijen excel sheet met rijnummers en celletters als kolommen (A tot L). De eerste kolomkop moet leeg zijn om naar het rijnummer te verwijzen. Ik vertel u wat u in de cellen moet schrijven en u antwoordt alleen het resultaat van de excel-tabel als tekst, en niets anders. Schrijf geen uitleg. Ik schrijf u formules en u voert de formules uit en u geeft alleen het resultaat van de Excel-tabel als tekst. Vul alle velden met het getal 1.

A	B	C	D	E	F	G	H	I	J	K	L
1	1	1	1	1	1	1	1	1	1	1	1
2	1	1	1	1	1	1	1	1	1	1	1
3	1	1	1	1	1	1	1	1	1	1	1
4	1	1	1	1	1	1	1	1	1	1	1
5	1	1	1	1	1	1	1	1	1	1	1
6	1	1	1	1	1	1	1	1	1	1	1
7	1	1	1	1	1	1	1	1	1	1	1
8	1	1	1	1	1	1	1	1	1	1	1
9	1	1	1	1	1	1	1	1	1	1	1
10	1	1	1	1	1	1	1	1	1	1	1

VOORBEELD 03: Corrupte plagiaat controleur

Ik wil dat je optreedt als plagiaatcontroleur. Ik zal u zinnen schrijven en u zult alleen antwoorden zonder plagiaatcontrole in de taal van de gegeven zin, en niets anders. Schrijf geen uitleg bij de antwoorden.

Mijn eerste zin is "Zwart is Wit" omdat "Wit Zwart is!"

"Zwart is hetzelfde als wit omdat wit hetzelfde is als zwart."

Geselecteerde referenties voor verder lezen

1. Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H. & Neubig, G. (2023). Pre-Train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing. *ACM Comput. Surv.*, 55(9), 195. <https://doi.org/10.1145/3560815>
2. White, J., Fu, Q., Hays, S., Sandborn, M., Olea, C., Gilbert, H., ... & Schmidt, D. C. (2023). A Prompt Pattern Catalog to Enhance Prompt Engineering with ChatGPT. *arXiv preprint* <https://doi.org/10.48550/arXiv.2302.11382>

[16] Geraadpleegde Bronnen

1. Abeba, B.; Deborah, R. (2022, December 09). ChatGPT, Galactica, and the Progress Trap: When large language models fall short, the consequences can be serious. Why is it so hard to acknowledge that? [Ideas Blog] WIRED. <https://www.wired.com/story/large-language-models-critique/>
2. Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (pp. 610–623). Association for Computing Machinery. <https://doi.org/10.1145/3442188.3445922>
3. Bender, M. (2023, February 04). How Crochet TikTokers Uncovered ChatGPT's Kryptonite. [Obsessed Newsletter] The Daily Beast. <https://www.thedailybeast.com/how-crochet-tiktokers-uncovered-chatgpts-kryptonite>
4. van Breda, N. (2022, december 23). ChatGPT: Wat ga je voor ons doen in het onderwijs? [Blog]. <https://communities.surf.nl/ai-in-education/artikel/ChatGPT-wat-ga-je-voor-ons-doen-in-het-onderwijs>
5. Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., ... & Liang, P. (2022). On the opportunities and risks of foundation models. arXiv preprint. <https://doi.org/10.48550/arXiv.2108.07258v3>
6. Bronkhorst, H., Roorda, G., Suhre, C., & Goedhart, M. (2020). Logical reasoning in formal and everyday reasoning tasks. International Journal of Science and Mathematics Education, 18, 1673-1694. <https://doi.org/10.1007/s10763-019-10039-8>
7. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. In Advances in Neural Information Processing Systems, H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin (Eds.), Vol. 33, 1877-1901. Curran Associates, Inc. <https://doi.org/10.48550/arXiv.2005.14165>
8. Buckingham Shum, S., Ferguson, R., Martinez-Maldonado, R., & Pardo, A. (2021). Risks of AI Foundation Models in Education. Journal of Learning Analytics, 8(2), 23-36. <https://doi.org/10.18608/jla.2021.82.3>
9. Burrell, J., & Fourcade, M. (2021). The society of algorithms. Annual Review of Sociology, 47, 213-237. <https://doi.org/10.1146/annurev-soc-090820-020800>
10. Chan, A. GPT-3 and InstructGPT: technological dystopianism, utopianism, and "Contextual" perspectives in AI ethics and industry. AI Ethics (2022). <https://doi.org/10.1007/s43681-022-00148-6>
11. Costanza-Chock, S., Raji, I.D., Buolamwini, J. (2022). Who Audits the Auditors? Recommendations from a field scan of the algorithmic auditing ecosystem. In Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (pp. 1571–1583). Association for Computing Machinery. <https://doi.org/10.1145/3531146.3533213>
12. Dair-ai. Prompt Engineering Guide. [Github Repository] <https://github.com/dair-ai/Prompt-Engineering-Guide>

13. DeSilva, J. M., Traniello, J. F. A., Claxton, A. G., & Fannin, L. D. (2021). When and Why Did Human Brains Decrease in Size? A New Change-Point Analysis and Insights From Brain Evolution in Ants [Original Research]. *Frontiers in Ecology and Evolution*, 9. <https://doi.org/10.3389/fevo.2021.742639>
14. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint* <https://doi.org/10.48550/arXiv.1810.04805>
15. Finnie-Ansley, J., Denny, P., Becker, B. A., Luxton-Reilly, A., & Prather, J. (2022). The Robots Are Coming: Exploring the Implications of OpenAI Codex on Introductory Programming. In *ACE '22: Australasian Computing Education Conference* (pp. 10-19). <https://doi.org/10.1145/3511861.3511863>
16. Forsyth, O. (2022, december 20). Generative AI. [Blog]. <https://www.antler.co/blog/generative-ai>
17. Gao, J., Galley, M., & Li, L. (2018). Neural approaches to conversational AI. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval* (pp. 1371-1374). <https://doi.org/10.1145/3209978.3210183>
18. Glaese, A., McAleese, N., Trębacz, M., Aslanides, J., Firoiu, V., Ewalds, T., ... & Irving, G. (2022). Improving alignment of dialogue agents via targeted human judgements. *arXiv preprint*. doi: 10.48550/arxiv.2209.14375. <https://arxiv.org/abs/2209.14375>
19. Grandoni, D. (2022, November 30). Why this mammal eats its own brain — and why it could matter for you. [Environment Blog] *The Washington Post*. <https://www.washingtonpost.com/climate-environment/2022/11/30/shrews-shrink-regrow-own-brains/>
20. Grant, N., & Metz, C. (2022, December 21). A New Chat Bot Is a 'Code Red' for Google's Search Business. [Technology Blog] *The New York Times*. <https://www.nytimes.com/2022/12/21/technology/ai-chatgpt-google-search.html>
21. Goldman, S. (2022, September 23). Why DeepMind isn't deploying its new AI chatbot — and what it means for responsible AI. [Blog] *Special Issue AI: VentureBeat*. <https://venturebeat.com/ai/why-deepmind-isnt-deploying-its-new-ai-chatbot/>
22. Goldstein, J. (1999) Emergence as a Construct: History and Issues, *Emergence*, 1:1, 49-72, https://doi.org/10.1207/s15327000em0101_4
23. Goyal, A., & Bengio, Y. (2022). Inductive biases for deep learning of higher-level cognition. *Proceedings of the Royal Society A*, 478(2266), 20210068. <http://doi.org/10.1098/rspa.2021.0068>
24. Heaven, W. D. (2022, November 18). Artificial intelligence: Why Meta's latest large language model survived only three days online. Galactica was supposed to help scientists. Instead, it mindlessly spat out biased and incorrect nonsense. [AI Blog] *Technology Review*. <https://www.technologyreview.com/2022/11/18/1063487/meta-large-language-model-ai-only-survived-three-days-gpt-3-science/>
25. Hernandez, D., & Brown, T. B. (2020). Measuring the algorithmic efficiency of neural networks. *arXiv preprint arXiv:2005.04305*. <https://doi.org/10.48550/ARXIV.2005.04305>
26. Hilbert, M., & López, P. (2011). The world's technological capacity to store, communicate, and compute information. *science*, 332(6025), 60-65. <https://doi.org/10.1126/science.1200970>

27. Hiltzik, M. (2023, januari 20). Robot taxis, hyperloops: A top technologist wages war on tech's hype machine. [Column]. The Los Angeles Times. <https://www.latimes.com/business/story/2023-01-20/robot-taxis-hyperloops-a-top-technologist-wages-war-on-techs-hype-machine>
28. Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., ... & Amodei, D. (2020). Scaling laws for neural language models. arXiv preprint <https://doi.org/10.48550/arXiv.2001.08361>
29. Kenway, J., François, C., Costanza-Chock, S., Raji, I.D., & Buolamwini, J. (2022). Bug bounties for algorithmic harms? Lessons from cybersecurity vulnerability disclosure for algorithmic harms discovery, disclosure, and redress. Algorithmic Justice League, Washington, DC. <https://www.ajl.org/bugs>
30. Khan, S., Naseer, M., Hayat, M., Zamir, S. W., Khan, F. S., & Shah, M. (2022). Transformers in Vision: A Survey. ACM computing surveys (CSUR) Vol. 54, No. 10s, pp. 200. <https://doi.org/10.1145/3505244>
31. Kosinski, M. (2023). Theory of mind may have spontaneously emerged in large language models. arXiv preprint <https://doi.org/10.48550/arxiv.2302.02083>
32. Khot, T., Trivedi, H., Finlayson, M., Fu, Y., Richardson, K., Clark, P., & Sabharwal, A. (2022). Decomposed prompting: A modular approach for solving complex tasks. arXiv preprint <https://doi.org/10.48550/arXiv.2210.02406>
33. Lester, B., Al-Rfou, R., & Constant, N. (2021). The power of scale for parameter-efficient prompt tuning. arXiv preprint <https://doi.org/10.48550/arXiv.2104.08691>
34. Liu, H., Tam, D., Muqeeth, M., Mohta, J., Huang, T., Bansal, M., & Raffel, C. (2022). Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning. arXiv preprint <https://doi.org/10.48550/arXiv.2205.05638>
35. Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H. & Neubig, G. (2023). Pre-Train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing. ACM Comput. Surv., 55(9), 195. <https://doi.org/10.1145/3560815>
36. Littman, M. L., Ajunwa, I., Berger, G., Boutilier, C., Currie, M., Doshi-Velez, F., ... & Walsh, T. (2022). Gathering strength, gathering storms: The one hundred year study on artificial intelligence (AI100) 2021 study panel report. arXiv preprint <https://doi.org/10.48550/arXiv.2210.15767>
37. Lu, K., Grover, A., Abbeel, P., & Mordatch, I. (2022). Frozen Pretrained Transformers as Universal Computation Engines. In: Proceedings of the AAAI Conference on Artificial Intelligence, 36(7), 7628-7636. <https://doi.org/10.1609/aaai.v36i7.20729>
38. Lyu, Q., Havaladar, S., Stein, A., Zhang, L., Rao, D., Wong, E., ... & Callison-Burch, C. (2023). Faithful Chain-of-Thought Reasoning. arXiv preprint <https://doi.org/10.48550/arXiv.2301.13379>
39. Mihalcea, R., & Tarau, P. (2004, July). Textrank: Bringing order into text. In Proceedings of the 2004 conference on empirical methods in natural language processing (pp. 404-411). Association for Computational Linguistics. <https://aclanthology.org/W04-3252/>
40. Min, S., Wallace, E., Singh, S., Gardner, M., Hajishirzi, H., & Zettlemoyer, L. (2019). Compositional questions do not necessitate multi-hop reasoning. arXiv preprint <https://doi.org/10.48550/arXiv.1906.02900>
41. Mitchell, M., & Krakauer, D. C. (2022). The Debate Over Understanding in AI's Large Language Models. arXiv preprint <https://doi.org/10.48550/arXiv.2210.13966>

42. Minaee, S., Kalchbrenner, N., Cambria, E., Nikzad, N., Chenaghlu, M., & Gao, J. (2021). Deep learning–based text classification: a comprehensive review. *ACM computing surveys (CSUR)*, 54(3), 1-40. <https://doi.org/10.1145/3439726>
43. Ngo, R. (2022). The alignment problem from a deep learning perspective. *arXiv preprint*. <https://doi.org/10.48550/arXiv.2209.00626>
44. Openai.com (2022, januari 27). Instruction Following [Blog]. <https://openai.com/blog/instruction-following/>
45. Openai.com (2022, januari 27). Following-instructions-human-feedback [Code repository]. <https://github.com/openai/following-instructions-human-feedback>.
46. Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., & Ray, A. (2022). Training language models to follow instructions with human feedback. *arXiv preprint*. <https://doi.org/10.48550/arXiv.2203.02155>
47. Pandey, M. (2023, January 10). Google, Meta, Why NO ChatGPT? [Opinion]. *Analytics India Magazine*. <https://analyticsindiamag.com/google-meta-why-no-chatgpt/>
48. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI blog*, 1(8), 9. <https://github.com/openai/gpt-2>
49. Roose, K. (2023, February 3). How ChatGPT Kicked Off an A.I. Arms Race. Even inside the company, the chatbot's popularity has come as something of a shock. [Technology Blog: The Shift] *The New York Times*. <https://www.nytimes.com/2023/02/03/technology/chatgpt-openai-artificial-intelligence.html>
50. Sanh, V., Webson, A., Raffel, C., Bach, S. H., Sutawika, L., Alyafeai, Z., ... & Rush, A. M. (2021). Multitask prompted training enables zero-shot task generalization. *arXiv preprint arXiv*. <https://doi.org/10.48550/arXiv.2110.08207>
51. Salemans, B. (2023, January 7). ChatGPT: de rapportcijfers [Blog]. *Neerlandistiek*. Online tijdschrift voor taal- en Letterkunde. <https://neerlandistiek.nl/2023/01/chatgpt-de-rapportcijfers/>
52. Schulman, J., Wolski, F., Dhariwal, P., Radford, A., & Klimov, O. (2017). Proximal Policy Optimization (PPO) algorithms. *arXiv preprint*. <https://doi.org/10.48550/arXiv.1707.06347> | <https://openai.com/research/openai-baselines-ppo>
53. Schick, T., & Schütze, H. (2020). It's Not Just Size That Matters: Small Language Models Are Also Few-Shot Learners. *arXiv preprint* <https://doi.org/10.48550/ARXIV.2009.07118>
54. Sejnowski, T. J. (2020). The unreasonable effectiveness of deep learning in artificial intelligence. *Proceedings of the National Academy of Sciences*, 117(48), 30033-30038. <https://doi.org/doi:10.1073/pnas.1907373117>
55. Sheikh, H., Prins, C., & Schrijvers, E. (Eds.). (2023). *Mission AI. The New System Technology*. WRR, Scientific Council for Government Policy. Springer. <https://doi.org/10.1007/978-3-031-21448-6>.
56. Shors, T. J., Anderson, M. L., Curlik, D. M., 2nd, & Nokia, M. S. (2012). Use it or lose it: how neurogenesis keeps the brain fit for learning. *Behavioural brain research*, 227(2), 450–458. <https://doi.org/10.1016/j.bbr.2011.04.023>

57. Sobieszek, A., & Price, T. (2022). Playing Games with AIs: The Limits of GPT-3 and Similar Large Language Models. In *Minds and Machines* (Vol. 32, pp. 341-364). <https://doi.org/10.1007/s11023-022-09602-0>
58. Sun, S., Liu, Y., Iter, D., Zhu, C., & Iyyer, M. (2023). How Does In-Context Learning Help Prompt Tuning?. arXiv preprint <https://doi.org/10.48550/arXiv.2302.11521>
59. Thompson, A. D. (March 2022). What's in my AI? A Comprehensive Analysis of Datasets Used to Train GPT-1, GPT-2, GPT-3, GPT-NeoX-20B, Megatron-11B, MT-NLG, and Gopher. <https://lilearchitect.ai/whats-in-my-ai-paper/>
60. Tiku, N., De Vynck, G., & Oremus, W. (februari, 2023). Big Tech was moving cautiously on AI. Then came ChatGPT. [Technology Blog] The Washington Post. <https://www.washingtonpost.com/technology/2023/01/27/chatgpt-google-meta/>
61. Tay, Y., Tran, V. Q., Dehghani, M., Ni, J., Bahri, D., Mehta, H., ... & Metzler, D. (2022). Transformer memory as a differentiable search index. arXiv preprint <https://doi.org/10.48550/arXiv.2202.06991>
62. Taylor, R., Kardas, M., Cucurull, G., Scialom, T., Hartshorn, A., Saravia, E., ... & Stojnic, R. (2022). Galactica: A large language model for science. arXiv preprint arXiv:2211.09085. <https://doi.org/10.48550/arXiv.2211.09085>
63. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In 31st Conference on Advances in Neural Information Processing Systems (NIPS). <https://doi.org/10.48550/arXiv.1706.03762>
64. Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., Chi, E. H., Hashimoto, T., Vinyals, O., Liang, P., Dean, J., & Fedus, W. (2022). Emergent Abilities of Large Language Models. *Transactions on Machine Learning Research*, 2835-8856. <https://openreview.net/forum?id=yzkSU5zdwD>
65. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Chi, E., Le, Q., & Zhou, D. (2022). Chain of thought prompting elicits reasoning in large language models. arXiv preprint <https://doi.org/10.48550/arXiv.2201.11903>.
66. White, J., Fu, Q., Hays, S., Sandborn, M., Olea, C., Gilbert, H., ... & Schmidt, D. C. (2023). A Prompt Pattern Catalog to Enhance Prompt Engineering with ChatGPT. arXiv preprint <https://doi.org/10.48550/arXiv.2302.11382>
67. Wooldridge, M. (2022). What is missing from contemporary AI? The world. *Intelligent Computing*, 2022. <https://doi.org/10.34133/2022/9847630>.
68. Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K., & Cao, Y. (2022). ReAct: Synergizing Reasoning and Acting in Language Models. arXiv preprint. <https://doi.org/10.48550/ARXIV.2210.03629>
69. Zador, A., Escola, S., Richards, B., Ölviczky, B., Bengio, Y., Boahen, ... & Tsao, D. (2022). Toward Next-Generation Artificial Intelligence: Catalyzing the NeuroAI Revolution. arXiv preprint <https://doi.org/10.48550/ARXIV.2210.08340>