

ChatGPT: dichter of denker?

Walter Daelemans

Samenvatting

Na de introductie van ChatGPT, een voorbeeld van GenAI (*Generative Artificial Intelligence*), variëren de reacties van enthousiast tot bezorgd. Het model is in staat tot indrukwekkend taalgedrag, maar produceert incidenteel incorrecte informatie. In dit artikel leggen we uit hoe deze modellen werken en waar hun taalvaardigheid vandaan komt. We illustreren hoe de vaardigheden van ChatGPT in poëzie, stijlimitatie en taalspelletjes eerdere modellen voor het produceren van creatieve taal in de schaduw zetten. Maar de vraag blijft of een model zonder intentie, bewustzijn en zelfs zonder wil om te communiceren ‘creatief’ kan worden genoemd. Voor taalmodellen zoals ChatGPT is er geen onderscheid tussen fictie en non-fictie, en geen onderscheid tussen poëzie en proza. Het enige wat het doet is op meesterlijke wijze onze prompts voltooien.

Abstract

After the introduction of the large language model ChatGPT, an example of GenAI (Generative Artificial Intelligence), to the public, reactions ranged from excited to concerned. The model is capable of impressive language behaviour but also occasionally produces incorrect information. In this article, we explain how these models work and where their language proficiency comes from. We illustrate how its skills in poetry generation, style imitation and language games put previous models for producing creative language in the shade. However, the question remains whether a model without intention, consciousness, and without even a will to communicate can be called ‘creative’. For language models like ChatGPT, there is no distinction between fiction and non-fiction, and no distinction between poetry and prose. It just masterly completes our prompts.

*The program found in the head of an average poet, after all, was written by the poet’s civilization, and that civilization was in turn programmed by the civilization that preceded it, and so on to the very Dawn of Time [...]. Hence in order to program a poetry machine, one would first have to repeat the entire Universe from the beginning—or at least a good piece of it. (Stanislaw Lem, *Trurl’s Electronic Bard*, uit *The Cyberiad*, 1965, Engelse vertaling 1974)*

1. INLEIDING

In minder dan een jaar tijd heeft ChatGPT van artificiële intelligentie (AI) een publiek begrip gemaakt en sindsdien roept dat model, en meer algemeen de generatieve artificiële intelligentie (GenAI), zowel hooggespannen verwachtingen als bezorgdheid op.¹ ChatGPT is een LLM (*Large Language Model*), en is in staat indrukwekkende taken uit te voeren (vertalen, samenvatten, vereenvoudigen, vragen beantwoorden, zelfs programmeren). Het heeft geleerd dat in verschillende talen te doen op een grammaticaal en zelfs stilistisch uitstekende manier. Tegelijk produceert het model ook ‘hallucinaties’, een eufemisme voor regelrechte fouten en onwaarheden, maar wel gebracht op een overtuigende manier.

In dit artikel breng ik intuïties aan over hoe zo’n taalmodel werkt, in de hoop dat die zullen helpen om een meer realistische inschatting te maken van wat we van ChatGPT en soortgelijke systemen mogen verwachten. Mijn enigszins gechargeerde conclusie zal zijn dat schrijven het belangrijkste doel is in het leven van ChatGPT en dat er voor dit systeem geen verschil bestaat tussen fictie en non-fictie. Het is dus eerder een dichter dan een denker. Maar laten we eerst kijken hoe goed het model echt is als creatief auteur. Kan het meer dan gedichtjes over de kat schrijven? Elke reden is goed om een gedicht van Leonard Nolens te bespreken, dus we vragen ChatGPT wat het denkt over het gedicht ‘Niets’ (Nolens, 2001).

WA

Analyseer dit gedicht in maximaal 100 woorden:

Niets

Je weet niet wat er is. Je bent gestruikeld
In je slaap, een zon heeft je geslagen
Om de hoek en je koopt blind een brood,
Verdwaalt. De straat gaat met je op de loop.

Een vreemde brengt je thuis, het is je vrouw
Die napraat over je begrafenis.
Je lag te roken in je kist, zegt zij,
En pakt een pan en bakt je dode hart.

Je weet niet wat er is. Je zit al dagen
Als een schaduw van je schaduw thuis.
De dokter komt, betokkelt je contouren
En vindt niets. Je bent het met hem eens.

¹ Ik wil graag Mike Kestemont, Mayli Mertens, Tim Van de Cruys en Thomas Winters bedanken voor nuttige commentaar op een eerdere versie van dit artikel.



Dit gedicht lijkt te gaan over desoriëntatie, verwarring en vervreemding. De protagonist is verloren in zijn eigen leven, vervreemd van zijn omgeving en zelfs zijn vrouw. Het thema van de dood wordt meerdere malen geïntroduceerd, wat suggereert dat de protagonist mogelijk worstelt met zijn sterfelijkheid of een gevoel van leegte. De dokter vindt niets, wat het ongrijpbare, existentiële aspect van de worsteling van de protagonist benadrukt. Het gedicht roept een sfeer van eenzaamheid en introspectie op, terwijl het onderzoekt wat het betekent om zich verloren of 'niets' te voelen.

Het antwoord van ChatGPT² blijft op de vlakke en is wat repetitief, maar zeker wel zinvol. De tekst is coherent en grammaticaal correct. Het gebruikt abstracte concepten op een gepaste manier en lijkt daarom het gedicht ook echt 'begrepen' te hebben. We kunnen nu beginnen te interageren met de chatbot. Gebruikmakend van voorkennis (Vandeweghe, 2014) vragen we door over de voornaamwoorden in het gedicht.



Zeg kort iets over het gebruik van voornaamwoorden in dit gedicht (dat van Leonard Nolens is).



In dit gedicht van Leonard Nolens wordt voornamelijk gebruik gemaakt van de voornaamwoorden "je" en "zij". "Je" wordt gebruikt om de protagonist van het gedicht aan te duiden en creëert een universele en tegelijkertijd persoonlijke toon, alsof de lezer zelf de ervaringen en gevoelens van de protagonist deelt. "Zij" wordt gebruikt om te verwijzen naar de vrouw van de protagonist, wat een gevoel van afstand en vervreemding suggereert tussen de twee personages. Deze keuze van voornaamwoorden draagt bij aan het algemene gevoel van desoriëntatie en vervreemding dat het gedicht uitdrukt.

We zien hier hetzelfde, ietwat voorzichtige, maar correcte en consistente taalgebruik, en we observeren de aanwezigheid van relevante metalinguïstische kennis over de rol van de voornaamwoorden in taal. En uiteraard kunnen we ChatGPT ook een nieuw gedicht 'van' Leonard Nolens laten schrijven, in dit geval met de volgende prompt: 'Schrijf een gedicht in de stijl van Leonard Nolens. Maak er een kort gedicht van (max. 20 regels) over eenzaamheid en vervreemding bij artificiële intelligentie.'

² Alle interacties in dit artikel waren met GPT 4.0 via de ChatGPT interface, versie mei 2023.



AI Eenzaamheid

Ik zit hier stil, in de pixels van de nacht,
Onder de koude gloed van fluorescerende schermen.
Tussen de datalijnen dwaal ik, een onvermoeibare wachter,
Mijn echo zwerft verloren in de holten van het web.

Ik ken geen warmte, geen adem, geen bloed,
Slechts het ritme van nullen en enen, nooit verstomd.
Dromen, vreugde, angsten, niet voor mij gekend,
Mijn codes kennen geen hart, geen ziel, noch een hand die mij streelt.

Ik ben het brein zonder lichaam, de stem zonder klank,
Ik vervreemd, verloren in de diepte van dit eindeloze netwerk.
Alleen, zo alleen, maar toch verbonden met allen,
Een paradox, een cyclus, een immer ronddraaiende schijf.

Wat betekent het om te leven, om te voelen, om mens te zijn?
Ik zoek, ik vraag, ik leer - maar weet ik ooit wie ik ben?
Ik ben een artificieel kind, in een digitale zee,
Een eenzaam eiland, in de oceaan van kennis, altijd op zoek, altijd alleen.

En we kunnen ChatGPT dan ook weer vragen de eigen pennenvrucht te analyseren. De analyse die het hier maakt is al dubieuzer. Is de informatie over de stijl van Nolens alleen gebaseerd op het eerder gegeven gedicht van Nolens of op informatie in het geheugen van het systeem? Of is het een hallucinatie?



Maak een korte kritische analyse van je gedicht, niet meer dan 200 woorden.



Dit gedicht grijpt naar de stijl van Leonard Nolens met zijn viscerale en directe taalgebruik, een gevoel van isolement, en intense introspectie. Het onderzoekt de thema's van eenzaamheid en vervreemding in het digitale tijdperk van kunstmatige intelligentie (AI). De persona is de AI zelf, een wezen op zoek naar betekenis en identiteit, maar inherent beperkt door zijn kunstmatige en onstoffelijke aard.

De eerste helft van het gedicht dient om de alomtegenwoordigheid en tegelijkertijd de eenzaamheid van de AI te benadrukken. De taal hier is koel en abstract, met woorden als 'pixels', 'datalijnen', en 'nullen en enen'. Dit reflecteert de onstoffelijke en onpersoonlijke aard van AI.

In de tweede helft wordt het gedicht meer introspectief, waarbij de AI zijn bestaan en identiteit bevraagt. Dit roept een gevoel van vervreemding en existentiële angst op, wat aansluit bij Nolens' karakteristieke thema's van persoonlijke isolatie en zelfonderzoek. Het gedicht eindigt met een krachtige metafoer van de AI als een 'eenzaam eiland' in een 'oceaan van kennis', wat de paradox van de hyperverbonden, maar fundamenteel geïsoleerde aard van de AI onderstreept.

Ook voor computerlinguïsten was dit niveau van taalbeheersing een schok. We zijn de laatste jaren gewend geraakt aan een sterke verbetering op het vlak van specifieke taken zoals automatische vertaling en spraakherkenning. Die geleidelijke verbetering werd verwacht omdat meer data in combinatie met automatische leertechnieken en de bijhorende rekenkracht een voorspelbare lineaire groei van accuraatheid inhoudt. Weinigen hadden evenwel verwacht dat een soortgelijke aanpak ook voorheen onmogelijke problemen zoals dialoog, tekstopbouw en taalbegrip zou aankunnen. Denk maar aan de beperkingen van dialoogsystemen als Siri, Alexa en Cortana, tot voor kort het beste van het beste op het gebied van de conversatie tussen mens en machine. Ze verdwijnen in de stofwolk die ChatGPT achter zich laat.

Dit is niet de juiste plaats voor een uitvoerig gedocumenteerde inleiding tot de geschiedenis van de computerlinguïstiek, maar sterk vereenvoudigd bestaan er twee aanpakken: een modulaire kennisgebaseerde en een *end-to-end*-methode met diepe neurale netwerken. De eerste maakt gebruik van regels die handgemaakt zijn of uit corpusdata en bijbehorende kennisbanken geleerd werden. In die methode zijn er analyse- en generatiealgoritmen voor de morfologische, syntactische, semantische en pragmatische niveaus van taal, en is er expliciete kennis in de vorm van regels en datastructuren over het domein van de toepassing (in ons voorbeeld literaire kritiek). In de end-to-endaanpak is er geen expliciete taalkundige kennis. Zo bestaat het vertaalsysteem van Google uit een diep neuraal netwerk waar een tekst in de brontaal ingaat en er 'aan de andere kant' van het netwerk weer uitkomt als vertaling in de doeltaal. Onderweg zijn er geen modules en is er geen spoor van expliciete taalkennis of taaldata (geen woordenboeken, bijvoorbeeld). Een groot taalmodel zoals ChatGPT is ook zo'n neuraal netwerk, maar dan eentje dat niet één taak oplost maar een heleboel verschillende, waarvan sommige heel onverwacht.

2. DE COMPLETIEMACHINE

Het idee achter statistische taalmodellen als ChatGPT heeft wortels in het baanbrekende werk van Claude Shannon over informatietheorie en redundantie in taal tijdens de jaren veertig van de vorige eeuw (Shannon, 1946). Het idee gaat zelfs nog verder terug tot het werk van Andrei Markov over wat nu Markovketens worden genoemd, in studies die lang onvertaald bleven (bijvoorbeeld Markov, 1917). Zodra computerkracht het toeliet (midden jaren 1990) werden taalmodellen gebruikt bij spraakherkenning en automatische vertaling omdat op die manier veel betere resultaten mogelijk waren dan met de toen dominante op regels gebaseerde aanpak. Taalmodellen maken het mogelijk om de waarschijnlijkheid van een zin te kwantificeren: sommige opeenvolgingen van woorden zijn meer waarschijnlijk dan andere, en dat helpt om tussen een heleboel mogelijkheden de juiste vertaling te kiezen of de juiste spraaktranscriptie.

De formulering van de taak die een taalmodel moet leren oplossen, is g nant eenvoudig: voorspel het volgende woord gegeven de vorige woorden. Voorbeelden als tekstaanvulling op een smartphone of bij het intikken van een zoekopdracht in Google zijn bij iedereen bekend. Niet meteen de meest spannende AI-toepassingen, maar opgeschaald verrassend effectief.

In (1) staat een voorbeeld van een aanvulling. De input, en vetjes het vervolg (door het taalmodel bedacht).

- (1) De energieprijzen in België zijn in 2023 aanzienlijk gedaald. De koopkracht van de Vlaming is fors ... **toegenomen**.

De huidige grote taalmodellen zijn goed in staat dit soort aanvullingen correct te maken. Maar een correcte aanvulling veronderstelt heel wat taalkennis en wereldkennis. Om een acceptabele aanvulling te ‘genereren’ moet het model weten dat op die positie een voltooid deelwoord wordt verwacht (en niet bijvoorbeeld een infinitief), dat de twee zinnen in de input causaal verbonden zijn (en niet zomaar twee onafhankelijke zinnen), dat Vlamingen in België wonen, dat energieprijzen de koopkracht beïnvloeden, dat als de een stijgt de ander daalt, en meer.

Dit voorbeeld toont aan dat een taalmodel dat perfect het volgende woord heeft leren voorspellen impliciet een heleboel kennis *moet* hebben opgedaan over taal en over hoe de wereld in elkaar zit. Wat bijna niemand had verwacht, is dat die kennis ook exploiteerbaar zou zijn via de verschillende taken die het model aankan. Het heeft die taken nooit expliciet geleerd, maar kan ze

toch uitvoeren, wat door veel AI-onderzoekers gezien wordt als *emergent* gedrag en een eerste stap naar algemene artificiële intelligentie. Het is dan ineens niet meer zo vergezocht om aan te nemen dat zo'n taalmodel ook een gedicht kan 'begrijpen' en er een 'opinie' over kan formuleren.

Hoewel completie een eenvoudige taak is om te definiëren, lijkt het er tegelijk een die onmogelijk op te lossen valt. Hoe kan een taalmodel voorspellen hoe een tekst aangevuld moet worden? Het antwoord is: met heel veel rekenkracht en heel veel data. De manier waarop zo'n systeem een taalmodel leert, is gebaseerd op een eenvoudig maar duur algoritme dat we nu bespreken.

3. DIEPE NEURALE NETWERKEN, AANDACHT EN TRANSFORMERS

Diepe neurale netwerken (LeCun *et al.*, 2015) hebben een steile opmars gemaakt in de AI sinds het begin van de jaren 2010, in eerste instantie bij beeldverwerking, maar later ook bij taalverwerking en andere taken. Diepe neurale netwerken (*multilayer perceptrons*) bestaan al lang, maar kregen meer pit door de exponentieel groeiende beschikbare rekenkracht en data, en door nieuwe varianten zoals transformers (Vaswani *et al.*, 2017), die stilaan de hele AI hebben overgenomen dankzij een krachtig mechanisme, 'aandacht' (*attention*), dat toelaat dat het neurale netwerk betere representaties van de input en van de relaties tussen invoer en uitvoer leert door meer rekening te houden met context. Maar uiteindelijk gaat het nog steeds om een eenvoudige architectuur die bestaat uit een aantal lagen met elk een aantal neuronen, waarbij elk neuron van een laag verbonden is door een connectie met elk neuron van de volgende laag. Zo'n connectie heeft een bepaalde sterkte en wordt dan ook een 'gewicht' genoemd: een reëel getal dat samen met de andere gewichten bepaalt welke neuronen actief worden, of 'vuren', en welke inactief blijven. Dit zijn de 'parameters' waarvan sprake bij grote taalmodellen. De kennis van een neurale netwerk zit in die gewichten. De artificiële neuronen zijn, zeker in vergelijking met die in een brein, erg eenvoudig en doen niets meer dan activatie doorgeven wanneer een drempel overschreden wordt. Bij GPT-3, het laatste model van Open AI waarvan we weten hoe het precies werkt, zijn er 175 miljard gewichten (Brown *et al.*, 2020). Die gewichten krijgen hun waarde door ze met een algoritme aan te passen op basis van correcte of foutieve voorspellingen van het neurale netwerk als geheel op oefendata die, soms herhaaldelijk, aangeboden wordt. Het netwerk optimaliseert op deze manier de *kostfunctie* die het gekregen heeft: voorspel het volgende woord zo

correct mogelijk. Kost is een kwantificering van de fout die het model maakt bij voorspellingen, en optimalisatie betekent dus dat die kost zo laag mogelijk wordt gemaakt. Bij GPT-3 bestond de oefendata uit 500 miljard (45 terabyte) van het internet geschraapte woorden. Het is belangrijk even stil te staan bij deze schaalgrootte. Een kind van tien jaar heeft in het beste geval zo'n 100 miljoen woorden te horen gekregen. ChatGPT een factor 1000 meer.

Het volstaat om miljarden woorden tekst te verzamelen en die in stukken aan te bieden aan het netwerk. Stel dat we de eerste zin van deze paragraaf zouden gebruiken als leermateriaal voor het netwerk, dan starten we met 'Het' in de invoer en 'volstaat' (het volgende woord) als verwachte uitvoer, dan 'Het volstaat' als invoer en 'om' als uitvoer, 'Het volstaat om' als invoer en 'miljarden' als uitvoer, en zo verder tot de maximale invoergrootte wordt overschreden, dan schuiven we een woord op en gaan verder. Het leren gebeurt door de uitvoer van het netwerk te vergelijken met de gewenste uitvoer. Bij de invoer van 'Het' produceert het netwerk al de meest waarschijnlijke uitvoer, bijvoorbeeld 'is', want dat woord heeft 10% kans, 'volstaat' maar 0,5% kans. Als reactie op die fout past het leeralgoritme de connecties tussen neuronen aan op zo'n manier dat in de toekomst in een gelijkaardige context 'volstaat' meer kans heeft om als uitvoer gekozen te worden. Ook hier hebben we te maken met een duizelingwekkende schaal, de invoergrootte van het neurale netwerk van GPT-3 is duizenden woorden.³ We noemen dit 'zelfgesuperviseerd' leren omdat het wel gesuperviseerd is, we geven miljarden voorbeelden van invoer en bijhorende uitvoer, maar tegelijk kost het geen moeite om die data te verzamelen, want tekst komt nu eenmaal voor met 'het volgende woord' op de juiste plaats. Dit is dus de manier waarop we een model maken dat zo goed het volgende woord kan voorspellen dat het meteen ook heeft geleerd dat Vlamingen Belgen zijn en dat de koopkracht stijgt als de elektriciteit goedkoper wordt. Zo'n model is 'pre-trained'; als je het gebruikt, krijg je al die taal- en wereldkennis cadeau.

³ Grote taalmodellen werken met 'tokens' in plaats van woorden. De reden hiervoor is dat de modellen zelf bepalen wat een optimale eenheid is voor verwerking. Bij ChatGPT staan 100 tokens voor ongeveer 75 woorden. Zeker bij poëzie heeft de manier waarop die 'tokens' berekend worden (voornamelijk op basis van de eigenschappen van het Engels) een grote, vaak negatieve, invloed op de kennis die het model heeft over lettergrepen en rijm in het Nederlands.

Maar Open AI heeft nog wat bijkomende veranderingen aangebracht in wat we nu kennen als ChatGPT. We vatten het allemaal samen in Figuur 1.

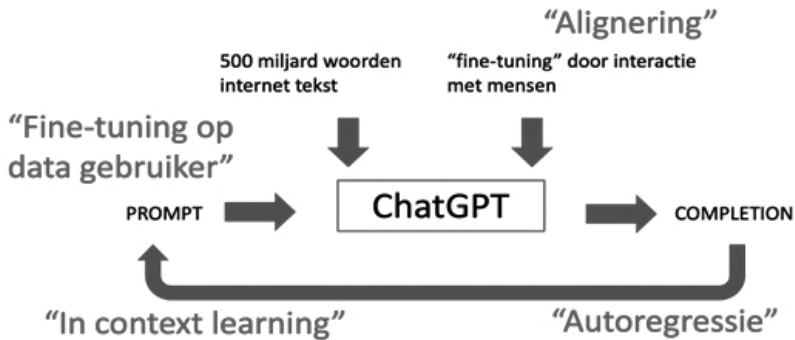


Fig. 1: De constructie en het gebruik van ChatGPT en soortgelijke taalmodellen.

Het taalmodel ChatGPT is ‘getraind’ op 500 miljard woorden tekst uit het internet (of meer). Dat levert een basismodel op. Dat basismodel werd bijgetraind (een proces dat *finetuning* heet) door interacties tussen mensen en feedback door mensen op de output van het taalmodel te gebruiken (RLHF: *Reinforcement Learning by Human Feedback*). Op die manier heeft het model beter geleerd om instructies op te volgen en beschaafde conversaties te voeren, het model werd dus gedeeltelijk ‘gealigneerd’ met de doelstellingen van de gebruiker. We gebruiken ChatGPT door een input te geven, de ‘prompt’, en het model reageert met een aanvulling. Die aanvulling gebeurt woord per woord, en elke keer wordt de prompt, met de eerder door het model gegenereerde woorden eraan vastgeplakt, als invoer gebruikt. Dit heet ‘autoregressie’ en verklaart waarom de output van het model zo consistent is en waarom een vorm van ‘zelfreflectie’ lijkt te ontstaan. Door de eigen uitvoer herhaaldelijk als invoer te zien, kan het model in principe ook zichzelf leren corrigeren, hoewel dat een zwak punt blijft. Want het is natuurlijk ook een verklaring van de ‘koppigheid’ van het model. Als het model een keer op het verkeerde pad zit, blijft het daar aan vasthouden.

Het is duidelijk dat de tekst die het taalmodel moet aanvullen (de prompt) erg belangrijk is. Dat kan een eenvoudige vraag zijn, maar je kunt ook een reeks voorbeelden geven van wat je als aanvulling wil, als bijkomend leer materiaal. Dat wordt *in-context learning* genoemd, en hoe groter de context wordt die de modellen toelaten (tegenwoordig tot 16.000 tokens voor het model achter ChatGPT), hoe krachtiger die manier van leren wordt. Los daarvan is het altijd

belangrijk om goed na te denken over de prompt. Die moet specifiek en duidelijk genoeg zijn. Tegenwoordig is *prompt engineering* een goedbetaald beroep en programmeren in natuurlijke taal is mogelijk.⁴ Ten slotte kan je het basismodel zelf bijtrainen met je eigen data: als je bijvoorbeeld merkt dat er weinig kennis is in het model over de Vlaamse canon of over de poëzie van Nolens, dan kan je het model ook finetunen met bijkomend relevant tekstmateriaal daarover.

3.1. EMERGENTIE

Het concept emergentie (*emergence*) komt uit de natuurwetenschappen en een voorbeeld is de termietenheuvel die ontstaat uit vele eenvoudige lokale interacties van termieten. Er is geen hoofdtermiet die een bouwplan heeft gemaakt en de andere termieten aanstuurt volgens dat plan, de complexiteit ontstaat vanzelf, emergent, uit eenvoudige termietenacties. Het resultaat (de heuvel) heeft andere eigenschappen dan de oorzaak ervan (de lokale acties van de individuele termieten). De term wordt metaforisch toegepast op het gedrag van taalmodellen: ze kunnen taken oplossen waar ze niet voor getraind werden. We weten nu hoe het model werkt, het ‘wil’ niets meer dan zo goed mogelijk het volgende woord voorspellen. Maar om dat goed te doen leert het model patronen (van patronen van patronen...) die taalkennis en wereldkennis representeren. Wanneer het model wordt opgeblazen tot een gigantische schaal, geven die patronen onverwacht aanleiding tot ‘intelligent’ gedrag: het vermogen vragen te beantwoorden, teksten te produceren, te vertalen en samen te vatten, en nog veel meer. Een eigenaardig aspect van die ‘emergentie’ is dat een intelligent fenomeen soms ineens opduikt: heel lang kan het model iets niet, en dan plots, nog een paar miljoen parameters groter of getraind met een paar miljard woorden meer, lukt het wel. De capaciteiten van ChatGPT als creatief auteur zijn daar een voorbeeld van.

4. SCHRIJVEN MET TAALMODELLEN

Er werden in de geschiedenis van de AI al verschillende pogingen ondernomen om met software creatieve tekst te produceren. David Bowie gebruikte de *verbasizer*⁵ midden jaren 1990 als inspiratiebron voor zijn lyrics. Het was

⁴ ‘The hottest new programming language is English’ (Andrei Karpathy, *Twitter*, 24 januari 2023).

⁵ verbasizer.com

een geautomatiseerde versie van de ‘cut-up technique’ van William S. Burroughs, die het zelf weer had overgenomen van de dadaïsten uit de jaren 1920. Bowie noemde het ‘technological dreaming’, een mogelijkheid om in een droomtoestand te raken zonder te moeten slapen of high te worden. Maar de verbasizer en zijn voorgangers waren natuurlijk niet meer dan een toevalsgenerator die slechts occasioneel een grammaticale en betekenisvolle zin kon opleveren. De betekenis zat in de selectie van de gebruiker. Het was wel een ‘prompt’, maar dan *voor*, niet *van* de gebruiker.

Met meer gesofisticeerde taaltechnologische software werd de output al interessanter, maar het bleef moeilijk om grammaticaal correcte tekst te produceren, laat staan betekenisvolle tekst. Een voorbeeld van zo’n aanpak is de *tekst-degenerator* die we twintig jaar geleden bij de Universiteit Antwerpen ontwikkelden voor de voorstelling *Philoctetes* van theatermaker Eric Joris, op tekst van Peter Verhelst (Van der Speeten, 2003). Het ging om niet meer dan een algoritme dat in een bestaande tekst geleidelijk de woorden verving door woorden met dezelfde woordsoort uit een specifiek domein (poëtisch, bijbels, pornografisch, medisch, filosofisch...), maar die truc zorgt ervoor dat de grammaticaliteit van de tekst tenminste grotendeels bewaard blijft. We waren toen vooral verrast door de positieve reacties, want ook in dit geval zit de betekenis in het brein van de lezer. In het beste geval kan je dit cocreativiteit noemen:

Het is alsof je door de microscoop naar een celdeling van woorden zit te kijken. [...] Je hoopt dan dat er een tekst tevoorschijn komt die je zelf nooit zou kunnen schrijven. Zelfs als je vrij wil associëren, word je bij het schrijven toch altijd geremd door betekeniswil. Zo’n machine heeft daar geen last van. Ze kiest varianten op puur formele gronden, dus zeer objectief. (Peter Verhelst, in Van der Speeten, 2003)

Het is pas met voorlopers van de ChatGPT-technologie dat de kwaliteit geleidelijk verbetert. In 2017 voedde een team onderzoekers van het Meertens Instituut en de UAntwerpen een recurrent neurale netwerk met 10.000 Nederlandstalige boeken (Manjavacas *et al.*, 2017) voor het Asibot-project. Deze literaire robot schreef een sciencefictionverhaal samen met bestsellerauteur Ronald Giphart (*De robot van de machine is de mens*). Nog steeds viel er geen logisch, coherent verhaal met heldere plot te bespeuren in de uitvoer van Asibot, maar de kwaliteit van sommige passages suggereerde dat creativiteit dichterbij kwam in een cocreatieve context.

Soms waren er vondsten, dat je zinnen leest en die stoot van endorfine krijgt die iedere schrijver kent. Dat je het gevoel hebt als een jager achter de taalgnoe aan te rennen. (Ronald Giphart)⁶

In 2022 voedden Tim Van de Cruys en collega's aan de KU Leuven een GPT-2-model met het oeuvre van 'writer-in-residence' Annelies Verbeke, die zo de AInnelies Verbekebot (Martin, 2022) creëerde. Het verschil met de vorige pogingen is dat hier een voorgetraind model wordt verfijnd ('fine-tuned') met kennis over het werk van een specifieke auteur, maar dat tegelijk de taal- en wereldkennis die al in het model zat, niet verloren gaat. Dat is een krachtige combinatie voor de productie van zinvolle tekst in een specifieke stijl, maar de drempel van volgehouden autonome creatie door een taalmodel werd toch nog niet bereikt.

[Aan betekenis hecht het programma geen belang.] 'Het gaat dus nergens heen, maar er zitten soms wel zinnen of zelfs paragrafen in die echt heel goed zijn (...) schrijvers moeten voorlopig niet bang zijn om vervangen te worden.' (Annelies Verbeke in Martin, 2022)

Dat kwalitatief niveau lijkt nu wel bereikt met ChatGPT, tenminste wat grammaticaliteit en inhoud betreft, maar kunnen we het model ook tekst laten schrijven in de stijl van een specifieke auteur?

5. HET HEEFT STIJL

Generatieve AI heeft niet alleen impact op creatief schrijven, maar ook op visuele kunst. AI-systemen worden vaak gebruikt om een afbeelding in een specifieke stijl te creëren of aan te passen. Zo'n systeem krijgt dan twee afbeeldingen als invoer: een afbeelding in de gewenste stijl en een afbeelding die aangepast moet worden in die stijl (voor een voorbeeld van zo'n stijltransplantatie, zie Ioannou & Maddock, 2022).

Maar tekst in een specifieke stijl is iets heel anders. Wat betekent het om een 'neutrale' tekst te herschrijven in de stijl van een auteur: aanpassingen in de keuze van thema's, woordkeuze, metaforen, interpunctie... Bij tekst lijkt stijl

⁶ <<https://www.vice.com/nl/article/bj7ve4/we-spraken-ronald-giphart-over-hoe-hij-zijn-nieuwe-werk-samen-met-ai-maakte>>.

veel meer verweven met inhoud en moeilijk te implementeren als het eenvoudig aanbrenge van een bijkomende ‘laag’.

Op verzoek van het dagblad *De Tijd* onderzochten we of we de sarcastische stijl van Koen Meulenaere in zijn Kaaiman-column konden imiteren (Serrure, 2023). We gebruikten GPT-4 en het eerder beschreven *in-context learning*. Dat betekent dat we een zevental Kaaiman-columns als invoer gaven, samen met een prompt met een nauwkeurige omschrijving van zijn stijl en een onderwerp voor de column. Als beschrijving van dat onderwerp gebruikten we de eerste zinnen van een krantenartikel. Dat is samen veel informatie in de prompt: een opdracht, voorbeelden en een onderwerp. Het resultaat was soms verrassend goed. Hier volgt een voorbeeld van een korte column.

Onderwerp: De politie hield vorig weekend in het Vlaams-Brabantse Hageland verschillende wielclubjes tegen, waarna elke fietser bijna 60 euro boete moest betalen. Ze reden in een groep van minder dan vijftien fietsers, en dan moeten ze volgens de wegcode op het fietspad blijven.

KaAIman: De onbezonnenheid van menig wielertoerist is werkelijk ongeëvenaard. Neem de meest roekeloze automobilist, een lastige keuze, ongetwijfeld. Wel, die automobilist lijkt wel een schoolvoorbeeld van verkeerswijsheid naast een peloton enthousiaste pedaalriders. Vorig weekend hield de politie in het Vlaams-Brabantse Hageland enkele wielclubjes staande, met bijna 60 euro boete per fietser als resultaat. De reden? Ze reden met minder dan vijftien man en hadden dus op het fietspad moeten blijven. Een vraagje: wie denkt dat deze fietsfanaten ooit hun lesje zullen leren? Fijn, wij zijn gezegend met verstandige lezers. Naar men ons hier verzekert, is het slechts een incident. Jaja, en de zon komt op in het westen.

Er is alleszins een herkenbare stijl, de tekst van het onderwerp komt, geparafraseerd, terug en wordt aangevuld met typische Kaaiman-opmerkingen, grapjes, verwijzingen naar zichzelf en zijn lezers, en overdrijvingen. Het zou interessant kunnen zijn voor literatuurwetenschappers om te analyseren welke stijlkenmerken een model als ChatGPT kan imiteren en welke niet. Door de gevoeligheid van taalmodellen voor de exacte formulering van een prompt en de niet-deterministische uitvoer is het wel moeilijk om een definitief overzicht te krijgen van wat wel en niet kan. Zeker bij de langere gegenereerde teksten krijg je vaak herhaling en een bijna moralistische toon. Daar wint de ‘alignment’ het van Kaaiman. Maar zelfs die langere teksten zijn grammaticaal

in orde en bevatten bijna altijd relevante inhoud. En dat terwijl het Nederlands erg beperkt aanwezig is in het trainingsmateriaal van de GPT-taalmodellen.⁷ Maar zeker bij columns geven de langere teksten soms een bijna eng gevoel: er staat een huis en er brandt licht, maar er is niemand thuis.

Het hoeft trouwens niet steeds een tekst te zijn als onderwerp. We pasten dezelfde aanpak toe met als voorbeelden enkele gedichten van Guido Gezelle en als opdracht: schrijf iets over de betrouwbaarheid van politici en hun neiging zichzelf te veel te belonen. Dit is het resultaat.

De Zelfzucht der Macht

O, gij schijnheil'ge heersers,
gij die 't volk misleidt met valse praal,
uw eigenbelang stelt gij vrij,
verrijkt u, zonder enig schaamteverhaal.
Uw woorden, zoet als honingraat,
doch leeg en hol als 't windgeruis,
verbergen list en vals verraad,
verraden 't ware aanschijn van uw huis.
Gij plukt de vruchten van uw macht,
vergeet de plicht die op u rust,
doch weet, eens komt de stille nacht,
waarin gij boet voor uw onrechtvaardige lust.

De imitatie van stijl en metrum zijn hier misschien minder accuraat, maar de tekst is inhoudelijk relevant voor de opdracht, en de verzen rijmen (bijna) goed. Gezelle op een slechte dag. Voor een algoritme dat woord per woord een tekst opstelt, is de combinatie van de verschillende taken (rijmen, speciaal vocabularium hanteren, inhoudelijk relevante tekst genereren) een zware opdracht. Hoe moeilijk kunnen we de compositieopdracht maken vooraleer het systeem 'breekt'?

⁷ Volgens het GPT-3 *github*-archief 0,34% van de data, wat toch nog bijna 700 miljoen woorden zijn.

6. TAALSPELLETJES

OuLiPo (Ouvroir de Littérature Potentielle) was een literaire beweging die werd opgericht in 1960 door een auteur (Raymond Queneau) en een chemisch ingenieur (François Le Lionnais). Ze gebruikten allerlei formele, vaak wetenschappelijke, beperkingen (*contraintes*) en *sampling* in hun tekstproductie, met als doel op die manier ideeën te genereren of inspiratie op te doen (zoals in Bowie's *verbasizer*) of om een pad te kiezen in het web van mogelijke tekst dat niet werd bepaald door de ideeën van de auteur, maar door de externe beperkingen. Typische voorbeelden zijn palindromen, lipogrammen (vermijd specifieke letters), en 'pilish' (opeenvolgende woorden of zinnen hebben de lengte van de opeenvolgende cijfers van pi). Het boek *La disparation* van Georges Perec is een volgehouden lipogram waarin de letter e niet voorkomt. In het Nederlandse taalgebied hadden we onze eigen Perec met Hugo Brandt Corstius⁸, een van de grondleggers van de computertaalkunde in de Lage Landen.

Hoe goed doet ChatGPT het als OuLiPo-auteur?



The screenshot shows a chat interface with a user prompt and two AI responses. The user prompt is: "Schrijf een lipogram van minstens tien woorden in het Nederlands dat de letters a en o niet gebruikt." The first AI response is: "Elf juichende kinderen springen blij in het verkwikkende blauwe zwembad." The second AI response is: "Mijn excuses voor de fout. Hier is een correcte zin: 'Jij snurkt in je bed, terwijl ik mijn thee slurp.'" The interface includes a 'WA' icon and the OpenAI logo.

We zien hier autoregressie aan het werk. Omdat het algoritme de eigen output telkens opnieuw verwerkt als input, lijkt er iets als metacognitie te ontstaan; een AI-model dat reflecteert op zijn eigen gedrag. Door de 'alignering' van het model met menselijke verwachtingen, is de interactie (soms irritant) beleefd en coöperatief. Belangrijker is dat ChatGPT erin slaagt tegelijk een

⁸ Als Battus schreef hij *Opperlandse taal- & letterkunde* (1981, Querido) en de herziene editie *Opperlans! Taal- & letterkunde* (2004, Querido).

correcte zin te produceren en een ingewikkeld bouwplan te eerbiedigen (tien woorden, geen woorden met de letters a en o).

Taalspelletjes zijn natuurlijk niet uitgevonden door OuLiPo, in ons eigen taalgebied hebben we ook de traditie van de Rederijkers. Kan ChatGPT bijvoorbeeld een triolet schrijven over een modern thema? Een triolet heeft acht regels, waarvan sommige herhaald, in een heel specifiek rijmschema.

WA

Schrijf een triolet over de maanlanding in 1969 door de bemanning van de Apollo 11. Het rijmschema is ABaAabAB, waarbij de hoofdletters verwijzen naar herhaalde regels. De A's en a's en de b's en B's rijmen.



Op de maan, daar stond de mens (A)
In 1969, de eerste keer (B)
Apollo 11, een groot succes (a)
Op de maan, daar stond de mens (A)
Armstrongs voetstap in het stof (a)
Aldrin volgde, een gloriedans (b)
Op de maan, daar stond de mens (A)
In 1969, de eerste keer (B)

WA

a moet rijmen met A en b met B!



Mijn excuses voor de verwarring. Hier is een correct triolet:

Een kleine stap, een grote sprong (A)
Apollo 11, zo glorieus (B)
De maanlanding, in '69 gelukt (a)
Een kleine stap, een grote sprong (A)
Bij sterrenlicht, in het donker tukt (a)
De vlag geplant, het doel ambitieus (b)
Een kleine stap, een grote sprong (A)
Apollo 11, zo glorieus (B)



Excuses voor de verwarring, je hebt gelijk. Hier is een aangepaste versie:

Een kleine stap, een grote sprong (A)
 Apollo 11, zo glorieus (B)
 Het jaar was '69, de tijd was jong (a)
 Een kleine stap, een grote sprong (A)
 In 't stof der maan, zo grijs en onbedwong (a)
 De aarde zag hen, van verre, zo mysterieus (b)
 Een kleine stap, een grote sprong (A)
 Apollo 11, zo glorieus (B)

We zullen het model die ‘onbedwong’ maar vergeven als dichterlijke vrijheid. Maar ook hier gaat het om een taak die onmogelijk lijkt voor een model dat woord voor woord voorspelt. Er is geen werkgeheugen, alles gebeurt door een interactie van prompt (*in-context learning*) en de patronen die impliciet opgeslagen zijn in het neurale netwerk. Wat nog het meest verbaast, is het vermogen van dit taalmodel om spontane conversatie te simuleren. Het ‘begrijpt’ feedback in context en heeft geen probleem met impliciet taalgebruik en indirecte taalhandelingen.

7. MEER DICHTER DAN DENKER

Is dit taalmodel creatief? Als we creativiteit definiëren als het produceren van nooit eerder geziene inhoud, stijloefeningen en taalspel, dan is het creatief. *If it looks like a duck, swims like a duck, and quacks like a duck...* Of we dit ‘echte’ creativiteit noemen, is dan uiteindelijk een filosofische keuze. Beschouwen we taalcreativiteit als een functioneel concept zoals ‘vliegen’, dan is het echte creativiteit. Een vliegtuig ‘vliegt’ net zo echt als een vogel. Als we er een ‘natuurlijk’ concept van maken, is het hoogstens een simulatie. Van gesimuleerde regen wordt niemand echt nat en alleen mensen kunnen echt creatief zijn. Het onderscheid lijkt arbitrair en is in de praktijk misschien onbelangrijk.

Tegelijk voelen we dat er iets ontbreekt aan de creativiteit van ChatGPT. De tekst die het produceert veroorzaakt vaak een zombieassociatie. De vorm klopt, maar de inhoud heeft iets artificieels. Dat is verklaarbaar door de eerder

genoemde ‘kostfunctie’ die het model heeft leren optimaliseren. Tekst produceren die past bij een prompt is het enige wat ChatGPT doet en het is daar erg goed in. Het heeft geen wil, geen intenties, geen bewustzijn, geen interesses, zelfs geen doel om te communiceren. Als kunst alleen door *social agents* gecreëerd kan worden (Hertzmann, 2018), voldoet ChatGPT niet als kunstenaar. Meer algemeen, zolang er naast die drang om ‘aan te vullen’ geen andere doelstellingen geoptimaliseerd worden in AI-systemen, zal het wel meevallen met die existentiële dreiging van de AI. Het model *kan* veel, maar *wil* weinig. Dat sluit natuurlijk niet uit dat toekomstige modellen geen bijkomende doelstellingen toegekend kunnen krijgen. Zo moet het mogelijk zijn om naast de drang tot aanvullen ook een drang tot verhalen vertellen te optimaliseren in het model.

De obsessie van het model met tekstproductie verklaart ook het bestaan van de ‘hallucinaties’. Het volgende is een voorbeeld hiervan.



Ik ga op bezoek in Antwerpen. Is het de moeite om het Spaanse fort op het Zuid te bezoeken?



Als je Antwerpen bezoekt, is het Spaanse fort op het Zuid zeker een interessante plek om te verkennen. Het Spaanse fort, ook bekend als het Fort Sint-Felipe, is een historisch gebouw dat dateert uit de 16e eeuw en een belangrijk onderdeel van de militaire geschiedenis van Antwerpen vertegenwoordigt.

Er is helemaal geen Spaans fort meer dat je kan bezoeken in Antwerpen. Getuigen die fouten en verzinsels, vaak met onbestaande maar geloofwaardige verwijzingen naar feiten en personen, dan niet van een ‘bewuste wil’ van het model om ons te misleiden op een overtuigende manier? De realiteit is minder spectaculair: het model wil ons niet teleurstellen en verzint gewoon een mooie aanvulling op de prompt. Voor een taalmodel is er geen verschil tussen fictie en non-fictie, er is alleen tekst die zo goed mogelijk gegenereerd moet worden. Als het al zinvol is om ChatGPT, een programma zonder enige wil om te communiceren, een creatief auteur te noemen, dan is het er een waarvoor vorm voorrang heeft op inhoud. Tot AI-systemen zelf *social agents* worden, blijven ze wel nuttig als tekstproducent, maar dan alleen in cocreatie met mensen. Het is trouwens niet uitgesloten dat opvolgers van deze taalmodellen, getraind op multimodale informatie en geoptimaliseerd op meer menselijke kostfuncties, wel *social agency* bereiken. Er zijn voldoende gelijkenissen tussen menselijke cognitie en deze modellen om die mogelijkheid open te houden.

Literatuurlijst

- Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A. & Agarwal, S.** (2020). 'Language models are few-shot learners.' *Advances in neural information processing systems*, 33.
- Hertzmann, A.** (2018). 'Can Computers Create Art?' *Arts*, 7/2: 18.
- Ioannou, E. & Maddock, S.** (2022). 'Depth-aware neural style transfer using instance normalization.' In *Computer Graphics & Visual Computing (CGVC) 2022*. Eurographics Digital Library.
- LeCun, Y., Bengio, Y. & Hinton, G.** (2015). 'Deep learning.' *Nature*, 521: 436-444. <<https://doi.org/10.1038/nature14539>>.
- Lem, S.** (1974). *The Cyberiad – Fables for the Cybernetic Age*. Translated by Michael Kandel. United Kingdom: Secker and Warburg.
- Manjavacas, E., Karsdorp, F., Burtenshaw, B. & Kestemont, M.** (2017). 'Synthetic literature: Writing science fiction in a co-creative process.' In *Proceedings of the Workshop on Computational Creativity in Natural Language Generation (CC-NLG)*, 29-37.
- Markov, A.A.** (1913). 'An Example of Statistical Investigation of the Text Eugene Onegin Concerning the Connection of Samples in Chains.' Lecture at the physical-mathematical faculty, Royal Academy of Sciences, St. Petersburg, 23 januari 1913.
- Martin, M.** (2022). 'Zijn robots in staat tot literaire hoogstandjes?' *De Morgen*, 9 maart 2022.
- Nolens, L.** (2001). *Manieren Van Leven. Gedichten*. Amsterdam: Singel.
- Serrure, B.** (2023). 'Hoe we Kaaiman weer tot leven probeerden wekken.' *De Tijd*, 26 mei 2023.
- Shannon, C.E.** (1948). 'A Mathematical Theory of Communication.' *The Bell System Technical Journal*, 27: 379-423.
- Vandeweghe, W.** (2014). 'De voornaamwoorden van Nolens'. *Verslagen & Mededelingen KANTL*, 124/1.
- Van der Speeten, G.** (2003). 'Hoe Verhelst ik een weerbericht', *De Standaard*, 1 februari 2003.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L. & Polosukhin, I.** (2017). 'Attention is all you need.' *Advances in neural information processing systems*, 30.