Large Language Models – de magie voorbij

Thomas Demeester

Universiteit Gent - IDLab
thomas.demeester@ugent.be
November 2023

Wie kwam het voorbije jaar nog niet in aanraking met ChatGPT? Teksten geschreven door een dergelijk groot taalmodel zijn zonder twijfel indrukwekkend. Echter, om er zo goed mogelijk te kunnen mee omgaan, is het belangrijk om een globaal beeld te hebben van hoe zo'n model werkt, en om de kwaliteiten, mogelijkheden, en zeker ook de beperkingen goed te kunnen inschatten.

Introductie

ChatGPT [1] bestaat een jaar, en de impact is nog steeds moeilijk te overzien. De beschikbare tools die verderbouwen op generatieve taalmodellen staan reeds mijlenver van wat beschikbaar was eind 2022. Zo goed als iedereen komt in aanraking met *Large Language Models* (kortweg LLMs genoemd), met het gemak om teksten te schrijven (zoals emails of blog posts), sterk verbeterde chat toepassingen (bv. bij klantendiensten), voor automatische datacollectie, programmeren... Er komen indrukwekkende nieuwe mogelijkheden aan om de *workflow* efficiënter te maken (bijvoorbeeld met Microsoft 365 Copilot [2]), en de introductie van de digitale *pair programmer* Github Copilot [3] heeft een niet te onderschatten impact op softwarebedrijven. Deze evolutie komt met stevige nieuwe uitdagingen in een aantal sectoren, waaronder het onderwijs. Dit vraagt om inhoudelijke aanpassingen, denk maar aan onderwijs rond schrijfvaardigheid of programmeren, maar ook naar de evaluatie van studenten toe – zo gaan we uiteindelijk niet anders kunnen dan zowel de verwachtingen als ook de waarde van een masterproef te herzien. En dit is nog maar een voorproefje van de evolutie die het onderwijs gaat doormaken, wanneer LLM-ondersteunde leermiddelen de komende jaren meer en meer ingeburgerd zullen worden (zoals NotebookLM van Google [4]). En uiteraard is het laatste woord nog niet gezegd over LLMs en (de interpretatie van het begrip) plagiaat, of de impact op de hoeveelheid van online misinformatie.

Daarnaast zijn er ook minder zichtbare maar niet minder belangrijke evoluties in de wereld van de LLMs, zoals het ontstaan van grote taalmodellen getraind op proteïnes [5], waarbij sequenties van aminozuren de rol nemen van woorden in natuurlijke taal, met als doel het ontwerp van nieuwe geneesmiddelen.

Tijdens het voorbije jaar heb ik een aantal lezingen gegeven over LLMs voor diverse groepen, waaronder studenten, artsen, business developers, tech bedrijven, opleidingsorganisaties, en taalkundigen. De toepassingen waren uiteenlopend, maar tijdens de vele discussies was de rode draad steeds dezelfde: om er goed te kunnen mee omgaan, zijn een aantal minimale intuïties nodig over hoe deze modellen zijn gebouwd, en hoe ze functioneren. Het occasioneel gebruik van ChatGPT op zich is volgens mij niet voldoende om deze intuïties op te bouwen. In onderstaande paragrafen probeer ik daarom een overzicht te geven. Hopelijk kan ik hiermee wat van de "magie" wegnemen. Hoewel ik focus op ChatGPT, de bekende chat toepassing die gebruik maakt van de LLMs GPT-3.5 en GPT-4 van OpenAI, zijn de meeste

intuïties ook toepasbaar op alternatieven zoals het conversationeel AI model Bard van Google [6] of de AI assistent Claude van Anthropic [7].

Deze bijdrage begint met een introductie in taalmodellen. Nadien wordt de familie van GPT modellen beschreven, gevolgd door een meer praktische toelichting van ChatGPT, met enkele gekende sterktes en beperkingen, en tips voor het effectief gebruik ervan. Ik ga niet in op trends, verwachtingen, of onderzoeksmogelijkheden - deze kunnen bij een andere gelegenheid aan bod komen.

Wat is eigenlijk een taalmodel?

Het oorspronkelijk doel van een **taalmodel** was om de waarschijnlijkheid van een gegeven tekst (zoals een zin) te kunnen inschatten. Dit is bijvoorbeeld nuttig voor het omzetten van audio naar tekst. Stel dat het audio signaal niet toelaat een onderscheid te maken tussen *"ze liet haring vallen in de taart"* of *"ze liet haar ring vallen in de taart"*, kan het taalmodel de doorslag geven. De tweede optie zal wellicht als meer waarschijnlijk worden aangeduid.

Statistische taalmodellen gingen probabiliteiten van teksten schatten door te gaan tellen hoe vaak stukjes tekst letterlijk voorkomen in een gegeven corpus van teksten. Ze waren echter niet in staat om interacties te modelleren tussen woorden die verder dan een handvol woorden uiteen stonden. In 2000 werd dan het eerste **neuraal taalmodel** voorgesteld [8]. Hierbij worden woorden voorgesteld door een *vector*, een lijst met getallen die zodanig worden afgesteld (tijdens de "training" van het model) dat de vectoren van sterk gelijkaardige woorden heel goed op elkaar lijken. De sequentie woordvectoren van een gegeven zin vertegenwoordigt dan ook een hele reeks andere mogelijke zinnen, zoals deze waarin woorden worden vervangen door hun synoniemen. Neurale taalmodellen bleken veel robuuster, en konden langere sequenties modelleren. De architectuur van het artificieel neuraal netwerk dat de probabiliteit van een sequentie modelleert, door de afzonderlijke woord-vectoren op gepaste wijze te combineren, heeft de voorbije jaren een hele evolutie ondergaan. Zo zwaaiden rond het midden van het vorige decennium de heel populaire recurrente neurale netwerken de plak, enkele jaren geleden van de troon gestoten door de "Transformer" [9]. Waar training bij het recurrent model sequentieel moest gebeuren, laat de structuur van de Transformer een ver doorgedreven parallellisatie toe tijdens training. Bovendien kunnen deze modellen veel dieper worden gemaakt, wat hen de nodige capaciteit geeft om enorme hoeveelheden tekst te modelleren.

Moderne taalmodellen bevatten vaak vele miljarden trainbare parameters. GPT-3 (OpenAI) bevat er bijvoorbeeld 175 miljard [10], en PaLM (Google) 540 miljard [11]. Van nieuwere modellen zoals GPT-4 of PaLM 2 zijn geen officiële cijfers gekend. Het opschalen van de neurale taalmodellen de voorbije jaren, zowel qua grootte van het model als in de hoeveelheid trainingdata, heeft hun vermogen om kwalitatieve tekst te produceren echter sterk verbeterd. Vandaar ook de benaming "*large language models*". Alle woorden van de input tekst (typisch de *prompt* genoemd) worden tegelijkertijd door een neuraal netwerk geduwd, bestaande uit vele tientallen tot binnenkort wellicht honderden lagen diep. Elke laag houdt een nieuwe transformatie in van de woord-vectoren, na het vergelijken en gewogen samenvoegen van de vector-representatie van elk woord met die van élk ander woord in de sequentie. Dit laatste is het zogenaamde "attention" mechanisme. Het vormt één van de sleutels tot het succes van de Transformer architectuur, ondanks de computationele kost die kwadratisch stijgt met de lengte van de input tekst.

Het is belangrijk om te beseffen dat de *training* van het model weliswaar in parallel kan gebeuren over alle woorden in een volledige tekst, maar dat het *genereren* van tekst wel degelijk woord per woord

gebeurt. Daarbij wordt steeds de originele prompt, uitgebreid met de reeds gegenereerde tekst, volledig door het model geduwd, om de kansverdeling voor het volgende woord te bepalen over het vocabularium. Merk op dat het vocabularium typisch zo'n 50.000 *tokens* telt. Deze *tokens* kunnen volledige woorden voorstellen, maar ook stukken van woorden tot op karakter-niveau, om alle mogelijke woorden te kunnen samenstellen, vaak zelfs in meerdere talen en over verschillende alfabetten (maar verder in deze tekst negeer ik dit onderscheid tussen *tokens* en *woorden*). Eens de kansverdeling van het volgende woord berekend is, wordt hieruit een *sample* genomen. Dit betekent dat een willekeurig woord wordt gekozen, volgens deze kansverdeling - niet noodzakelijk het meest waarschijnlijke woord dus. Dat laatste is essentieel, want anders voelt de tekst vaak niet natuurlijk aan. Het gevolg is wel, dat het model voor exact dezelfde prompt volledig andere teksten kan genereren.

Een korte geschiedenis van GPT.

In de familie van de *Generative Pretrained Transformer* (GPT) modellen van OpenAI werd het eerste model (**GPT-1** [12]) reeds in 2018 voorgesteld. Getraind op het eerder naïeve objectief "leer het volgende woord voorspellen", bleek GPT-1 toch bijzonder effectief voor een brede waaier aan predictie-taken in het vakgebied van natuurlijke taalverwerking (of NLP, *natural language processing*), mits slechts beperkte extra training voor elk van de taken. Volgens de normen van vandaag was dat een vrij klein model, met "slechts" 117.000.000 trainbare parameters – dat zijn de instelbare gewichtjes van de connecties tussen de neuronen van het artificieel neuraal netwerk. De jaren nadien bleef de architectuur vrij gelijkaardig, gebaseerd op het generatief stuk van de Transformer, behalve dat de modellen dieper werden (meer en meer Transformer bouwblokken boven op elkaar) en breder (met meer neuronen per laag, dus langere vectoren om woorden voor te stellen). Zo bestond GPT-1 uit 12 Transformer blokken, voor de opeenvolgende transformaties op woordvectoren van 768 dimensies lang. Het **GPT-3** model, voorgesteld in 2020 [10], was 96 dergelijke lagen diep, voor vectoren met een lengte van 12.888. Het totaal aantal parameters werd 175.000.000.000, op dat moment ongezien groot. Ook naar de huidige normen is dit model nog steeds zodanig groot dat slechts een klein aantal organisaties het vermogen hebben om dergelijke modellen te trainen. Het was getraind op een kolossaal corpus van digitale boeken, wikipedia, web pagina's, online fora, programmacode..., nog steeds puur met de focus om zo goed mogelijk "het volgende woord" te leren voorspellen. Zowel de onderzoeksgemeenschap als de tech industrie waren onder de indruk van de kracht van GPT-3 om uiteenlopende taken in NLP uit te voeren zelfs zonder daar expliciet op te zijn getraind. Er gingen echter ook steeds meer stemmen op die waarschuwden voor de risico's van dergelijke modellen, zoals het genereren van toxische teksten, teksten met *bias* (en daardoor mogelijks discriminerend), of hallucinaties (onbestaande of foutieve feiten). Daarnaast schoot GPT-3 te kort in het correct opvolgen van instructies van de gebruiker. Om maar te zwijgen over de maatschappelijke risico's bij het grootschalig gebruik ervan [13].

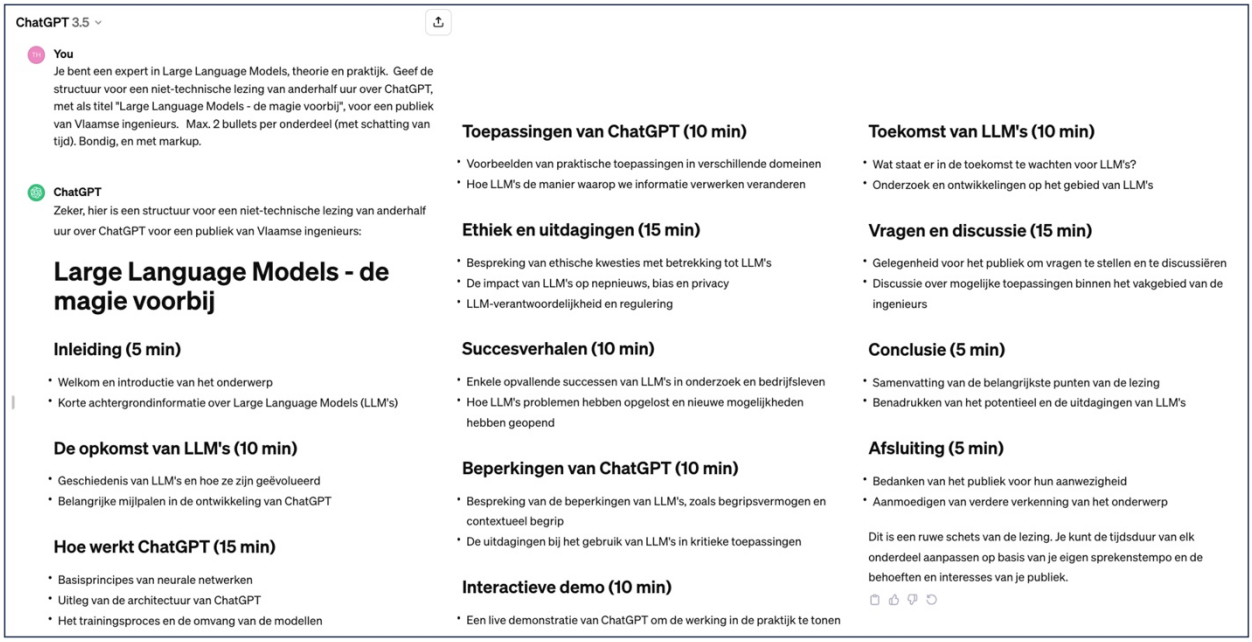
Een eerste stap in de goede richting werd opnieuw gezet door OpenAI, in het **InstructGPT** model [14]. Eerst werd GPT-3 verder verfijnd door expliciet te trainen op manueel gecreëerde "ideale" antwoorden op input instructies. Voor een groot aantal inputs werden vervolgens meerdere outputs van het model gegenereerd, waarna de meest wenselijke outputs door mensen werden geselecteerd. Via de techniek *Reinforcement Learning from Human Feedback* (RFHF) werd het model verder verfijnd om tekst te genereren die beter in lijn lag met deze menselijke voorkeuren. Het model ging al minder hallucineren, en was beter in staat instructies te volgen, en te antwoorden in lijn met de voorkeuren van de annotatoren.

In een volgende fase werd het eerste **ChatGPT** model publiek gemaakt (november 2022, in de familie van de zogenaamde **GPT3.5** modellen), eveneens gebaseerd op RLHF maar met meer focus op dialogen dan bij InstructGPT. In maart 2023 kwam dan **GPT-4**, de langverwachte nieuwe generatie GPT modellen [15]. Tot grote ontsteltenis in de onderzoekswereld rapporteerde OpenAI geen enkel detail meer over de architectuur en training van het model, omwille van *"the competitive landscape and safety implications"*. Microsoft had er toen net ook 10 miljard dollar in geïnvesteerd. GPT-4 was baanbrekend in de multi-modale input (tekst en beeldmateriaal) en de sterk uitgebreide context tijdens training (tot 32.000 tokens per keer, wat neerkomt op 50 pagina's tekst). Het werd heel recent nog overtroffen door GPT-4 Turbo [16] dat overweg kan met een context van 128.000 tokens.

Waar vind ik ChatGPT?

Mijn doel is helemaal niet om reclame te maken voor het OpenAI universum. Dat neemt niet weg dat ChatGPT heel toegankelijk is, en ik zou zelfs de meest fervente tegenstanders aanraden om er toch even naar te kijken. De meest directe weg om ChatGPT uit te proberen, is via de web applicatie van OpenAI [17], met gratis toegang tot ChatGPT 3.5 en betalend (via ChatGPT Plus) tot ChatGPT 4. Deze laatste versie biedt ook toegang tot het combineren van de web-zoekmachine Bing (van Microsoft, uiteraard) en tekst generatie via GPT-4 op basis van de zoekresultaten. Ook de functionaliteit van GPT-4 om tekst en beeld data samen in te geven als prompt wordt sinds kort aangeboden. Dit kan bovendien worden gecombineerd met Dall-e 3 [18], OpenAI's meest recente model om afbeeldingen te genereren op basis van een beschrijving. Wie GPT-4 wil uitproberen in combinatie met web search, kan dit ook gratis via de chat functionaliteit binnen de Edge browser [19].

Daarnaast biedt OpenAI een zogenaamde API (Application Programming Interface) aan, de functionaliteit om via een ander programma rechtstreeks de GPT modellen te kunnen aanspreken. Hierbij betaalt de gebruiker een prijs per 1000 tokens, die hoger ligt bij de meer krachtige modellen. Bovenop de API kan bijvoorbeeld een ander programma worden gebouwd (zoals de "ChatGPT Plugins", die gespecialiseerd zijn in specifieke taken, bijvoorbeeld interageren met een website, een email programma...). De API biedt ook een aantal mogelijkheden van controle aan, die de web-applicatie niet heeft. Zo is het mogelijk om het maximum aantal tokens in te stellen, of om via de *temperature* parameter het *sampling* proces te beïnvloeden, om het model eerder heel standaard tekst of juist wat meer creatieve output te laten genereren. Via de API kan ook op grotere schaal data worden gegenereerd. Zo hebben we bijvoorbeeld recent een dataset van 400.000 biomedische definities gebouwd [20], om nadien een gespecialiseerd biomedisch taalmodel mee te trainen [21]. Er werd een *prompt* skelet ontworpen, waarin per biomedisch concept een aantal gestructureerde feiten werden toegevoegd (uit een bestaande databank), om te vermijden dat het model zou hallucineren. Via diezelfde prompt, telkens automatisch aangepast voor het gewenste biomedisch concept, werd het taalmodel aangesproken om een definitie in natuurlijke taal te genereren. De API biedt een heel gamma aan modellen aan [22]. We kozen voor het specifieke model *gpt-3.5-turbo*, en betaalden 500$ voor de hele dataset. Merk op dat het beste model op dat moment (*gpt-4-32k*) ongeveer 30.000$ zou hebben gekost, met weinig meerwaarde voor ons doel.



Figuur 1: Voorbeeld prompt (links boven) en gegenereerde tekst van ChatGPT [ChatGPT3.5, 25/11/2023].

Wat kan een generatief taalmodel allemaal - en wat niet?

Het is onbegonnen werk om alle toepassingen van LLMs te gaan voorzien, laat staan ze in een korte paragraaf op te lijsten. Een LLM zoals ChatGPT kan een antwoord geven op een input tekst (de *prompt*) van de gebruiker, in meerdere talen, en vaak overtuigend geformuleerd. De screenshots in figuur 1 geven het antwoord op een voorbeeld prompt, met als vraag de structuur van een presentatie van anderhalf uur over ChatGPT, in het Nederlands. De output houdt inhoudelijk steek, maar merk op dat de instructies niet exact gevolgd worden: vaak zijn er meer dan 2 bullet points per onderdeel, en de voorgestelde duur is een half uur te lang. Dat laatste weerspiegelt een inherente beperking in het vermogen van het LLMs om logisch of wiskundig te redeneren, zodat de uiteindelijke som van het voorgesteld aantal minuten voor elk onderdeel samen hoger is dan het gevraagde anderhalf uur.

In de context van een ChatGPT dialoog gaan de voorgaande interacties ook deel uitmaken van de effectieve prompt, zodat het model een zekere vorm van consistentie kan bereiken doorheen een sessie. Het model is in staat om teksten samen te vatten, te vertalen, herstructureren, corrigeren, de stijl aan te passen, en noem maar op. Naast natuurlijke taal, kan het ook om met programmeercode, of semi-gestructureerde teksten genereren (bijvoorbeeld met HTML opmaak). Een belangrijk aspect is het vermogen van een LLM om, naast het nut als schrijf-assistent, ook analyse van teksten te doen, eveneens geformuleerd als een generatieve taak. Zo kan het bijvoorbeeld sentiment in tekst herkennen, een grammaticale structuur van een tekst opstellen, of fouten in computer code opsporen en erover argumenteren.

Let wel dat dit alles niet noodzakelijk foutloos gebeurt. Zoals gezegd durft een LLM te hallucineren, kan het ongewenste tekst of tekst met bias produceren, en is het niet altijd in staat de instructies in de prompt correct te volgen. Het blijkt ook moeilijk om steeds een onderscheid te maken tussen "correcte" feiten die het model tijdens training zag, en mogelijke (soms met opzet opgegeven) "incorrecte" feiten in de prompt. Soms gaat het ook volledig de mist in door een aantal inherente beperkingen, bijvoorbeeld op

het gebied van logisch redeneren over situaties die kennis van de wereld vergen. Wel kan worden gezegd dat het betalende GPT-4 voor de meeste types van tekortkomingen beter scoort dan de GPT-3.5 modellen. Figuur 2 geeft enkele illustraties van dergelijke beperkingen.

Figuur 2: Voorbeelden van mogelijke fouten van ChatGPT [ChatGPT 3.5, 22/11/2023].

Een andere belangrijke beperking heeft te maken met de zogenaamde *knowledge-cutoff*, de datum waarna geen recentere data gebruikt werd om het model te trainen. Voor de huidige GPT-3.5 en de originele GPT-4 modellen ligt deze op september 2021, en bij de meer recente GPT-4 modellen op april 2023. Het model zelf heeft bij training geen feitelijke informatie gekregen die op het moment van de *knowledge-cutoff* niet voorhanden was, bijvoorbeeld over meer recente gebeurtenissen. Het is dus belangrijk om te weten welk model gebruikt wordt. Zoals reeds aangehaald kan de combinatie van een LLM met zoekmachines die wel toegang hebben tot meer recente data, het model in staat stellen meer recente feiten in de gegenereerde tekst te integreren. Dit principe heet '*Retrieval Augmented Generation*' [23], en heeft het potentieel om onze manier van interageren met online zoekmachines grondig te wijzigen. Sinds november 2023 beslist ChatGPT 4 alvast automatisch om al dan niet rechtstreeks een antwoord te formuleren, of eerst in real-time naar online informatie te zoeken, en deze mee in de interne prompt te integreren.

Wat is "prompt engineering"?

Vaak krijg ik te horen dat ChatGPT wel leuk is, maar toch niet in staat blijkt om de ene of andere taak echt op te lossen. In een deel van deze gevallen, heeft dit echter te maken met een gebrekkige manier van de prompt te formuleren. Online zoekmachines hebben ons "lui" gemaakt in het formuleren van een query. Zij gaan dan ook proberen voorspellen wat de gebruiker niet expliciet in de query opnam maar wel relevant is om aan de achterliggende informatie-vraag te voldoen. Bij LLMs echter, moet de opdracht momenteel nog zo goed mogelijk in detail worden beschreven. Dit wordt ook wel *prompt engineering* genoemd. Hierbij helpt het bijvoorbeeld om het model een rol toe te wijzen (zoals "expert in LLMs" in het

voorbeeld van Figuur 1), en een doelpubliek (zoals "Vlaamse ingenieurs"). Stijl en formaat kunnen ook expliciet worden benoemd (bv. "gebruik markup"). Voor het schrijven van een langere tekst (zoals een thesis) is het nuttig om de techniek van *chained prompting* te hanteren. Hierbij wordt eerst de structuur op hoog niveau gegenereerd, en in daaropvolgende prompts kan gradueel meer detail worden toegevoegd. Om een bepaalde sectie te gaan uitschrijven, heeft het model dan reeds expliciet (via het *attention* mechanisme) zicht op het onderwerp van secties die pas verderop moeten komen. Deze hiërarchische manier van werken komt ook beter overeen met hoe wij zoiets als mens aanpakken en blijkt veel beter te werken, dan wanneer het model alles van A tot Z genereert. Voor taken die iets van redeneren vergen, kan *chain-of-thought prompting* worden toegepast, waarbij het model expliciet de instructie krijgt om tussenstappen te gebruiken. Dit kan gebeuren door voorbeelden van het redeneerproces mee op te nemen in de prompt, of door gewoon iets als "*Let's think step by step.*" toe te voegen. Meer exotische prompt strategieën zijn ook in trek. Zo kunnen we het model vragen om kritische feedback op voordien gegenereerde tekst ("self-criticism" in de volksmond), en vervolgens vragen om de tekst te herschrijven, rekening houdend met die feedback. Maar vaak vergt het ook gewoon heel wat uitproberen om finaal tot een goede prompt te komen.

Disclaimer

Bij het schrijven van dit artikel werd op geen enkele wijze gebruik gemaakt van ChatGPT, behalve voor het genereren van de voorbeelden in Figuur 1 en 2. Voor elke vorm van wetenschappelijke communicatie, vind ik het belangrijk om de aard en de mate van het gebruik van generatieve modellen correct te benoemen. Zou u het artikel anders inschatten, als ik ChatGPT wel had gebruikt?

Referenties

[1] *Introducing ChatGPT*, https://openai.com/blog/chatgpt.

[2] *Introducing Microsoft 365 Copilot*, https://blogs.microsoft.com/blog/2023/03/16/introducing-microsoft-365-copilot-your-copilot-for-work/.

[3] *GitHub Copilot*, https://github.com/features/copilot.

[4] *NotebookLM*, https://notebooklm.google/.

[5] N. Ferruz, S. Schmidt en B. Höcker, *ProtGPT2 is a deep unsupervised language model for protein design*, in *Nature Communications*, 2022.

[6] *Bard*, https://bard.google.com/.

[7] *Introducing Claude*, https://www.anthropic.com/index/introducing-claude.

[8] Y. Bengio, R. Ducharme en P. Vincent, *A neural probabilistic language model*, in *NIPS*, 2000.

[9] A. Vaswani, N. Shazeer, et al., *Attention is all you need*, in *NIPS*, 2017.

[10] T. Brown, B. Mann, et al., *Language models are few-shot learners*, in *NeurIPS*, 2020.

[11] A. Chowdhery, S. Narang, et al., *PaLM: Scaling Language Modeling with Pathways*, in *arXiv:2204.02311*, 2022.

[12] A. Radford, K. Narasimhan et al., *Improving Language Understanding by Generative Pre-Training*," 2018.

[13] E. M. Bender, T. Gebru, et al., *On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?* 🦜, in *FAccT*, 2021.

[14] L. Ouyang, J. Wu, et al., *Training language models to follow instructions with human feedback*, in *NeurIPS*, 2022.

[15] OpenAI, *GPT-4 Technical Report*, in *arXiv:2303.08774*, 2023.

[16] *GPT-4 and GPT-4 Turbo*, https://platform.openai.com/docs/models/gpt-4-and-gpt-4-turbo.

[17] *ChatGPT*, https://chat.openai.com/.

[18] *DALL-E 3*, https://openai.com/dall-e-3.

[19] *Microsoft Edge*, https://www.microsoft.com/en-us/edge.

[20] F. Remy, K. Demuynck en T. Demeester, *Automatic glossary of clinical terminology : a large-scale dictionary of biomedical definitions generated from ontological knowledge*, in *BioNLP*, 2023.

[21] F. Remy, K. Demuynck en T. Demeester, *BioLORD-2023: Semantic Textual Representations Fusing LLM and Clinical Knowledge Graph Insights*, [Under review], 2023.

[22] *OpenAI Documentation - Models*, https://platform.openai.com/docs/models.

[23] P. Lewis, E. Perez, et al., *Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks*, in *NeurIPS*, 2020.