Wat is een taalmodel en hoe werkt het?

Taalmodellen zijn kunstmatige neurale netwerken die teksten kunnen genereren op basis van voeding die bestaat uit miljarden stukjes data. Teksten die nodig zijn voor het trainen van deze modellen worden veelal van het internet geplukt. Een methode die bekend staat als 'scrapen' (schrapen). Het meest bekende taalmodel is de Generative Pre-trained Transformer (GPT). Ook wel bekend van ChatGPT, de populaire chatbot van OpenAI. Naast OpenAI zijn er meer grote commerciële partijen bezig met de ontwikkeling van een eigen taalmodel (LLM), zoals de LLaMA (Large Language Model Meta AI) van Meta, en PaLM (Pathways Language Model) van Google.

Een belangrijk technisch aspect is het gebruik van open-source modellen. Van open-source taalmodellen is namelijk de broncode en de onderliggende dataset bekend. Dit maakt dat het technologiegebruik transparanter is en het makkelijker is om te controleren welke keuzes en afwegingen door een taalmodel worden gemaakt. Desalniettemin kleven er ook nadelen aan open-source taalmodellen, doordat potentiële kwaadwillenden ook makkelijker toegang krijgen tot de technologie erachter.

Bovendien is het belangrijk om kritisch te kijken naar wat er ‘open-source’ wordt genoemd. Meta noemt de LLaMA2 een open-source model, maar volgens wetenschappers van de Radboud Universiteit voldoet het niet aan de eisen voor daadwerkelijke open-source modellen.

Op de vraag hoeveel 2+2 is, weten taalmodellen het antwoord. Dat halen ze uit bestaande teksten waarin meestal het getal 4 volgt na 2+2. Via de tekstvoorspellingstechnologie, die op basis van bestaande teksten werkt, worden verbanden gelegd tussen de volgorde van woorden en zo onze taal gereproduceerd. Ze hebben geen zelfbewustzijn en daarmee geen grip op hun eigen beperkingen of vooroordelen. Kortom, een taalmodel ‘begrijpt’ niet wat het zelf uitvoert.

Concrete kansen

Recente onderzoeken over taalmodellen zijn behoudend positief. Zo wordt er wel onderzoek gedaan naar nieuwe (theoretische) toepassingen, maar tegelijkertijd wordt er kritisch gesproken over risico’s. In dit hoofdstuk bekijken we de toepassingen.

Uit onderzoek komen een aantal mogelijke toepassingen naar voren in het onderwijs, medische veld en bioengineering. In het onderwijs kunnen leraren taalmodellen inzetten als hulpmiddel in het ontwikkelen van passende toetsen voor en feedback aan leerlingen, zodat zij meer tijd overhouden voor hun lesplannen. En in het medische veld maken taalmodellen informatievoorziening toegankelijker door complexe, elektronische gezondheidsdossiers leesbaarder te maken. Dit zorgt ervoor dat de behandeling van patiënten efficiënter kan verlopen. In de onderzoekswereld levert het gebruik van taalmodellen voornamelijk nieuwe inzichten op. Zoals in de bioengineering waar nieuwe eiwitsequenties gebouwd worden die niet voorkomen in de natuur. Dit soort innovatieve experimenten zijn cruciaal bij de ontwikkeling van medische oplossingen. Denk bijvoorbeeld ook aan het verbeteren van vaccinaties.

Ook met betrekking tot de communicatie met inwoners bieden taalmodellen op verschillende manieren mogelijkheden. Een bekend probleem is bijvoorbeeld het taalgebruik in brieven en formulieren en op websites dat door veel mensen als te moeilijk wordt ervaren. Een taalmodel kan complexe ambtelijke en juridische taal omzetten naar een goed leesbaar niveau. Daarnaast gebruiken diverse gemeenten op dit moment al een chatbot om vragen van inwoners te beantwoorden. De functionaliteit van deze chatbots kan worden uitgebreid met behulp van een taalmodel.

Een andere toepassing is meertalige communicatie met inwoners. Zo kan een taalmodel dat getraind is op meerdere talen, mogelijkheden bieden voor gemeenten waar veel verschillende talen worden gesproken. Denk aan een gemeente als Den Haag met een cultureel diverse bevolkingssamenstelling. Met behulp van zo'n taalmodel kunnen gemeenten (live) communiceren in de moedertaal van hun inwoners. Uit het rapport *Uitdagingen in het sociaal domein* van het Sociaal en Cultureel Planbureau (CPB) blijkt dat gemeenten soms moeite hebben met het bereiken van grote groepen inwoners. De inzet van taalmodellen kan een belangrijke bijdrage leveren aan dat bereik.

Ook voor ambtenaren bieden taalmodellen op termijn mogelijkheden. Zo kan een taalmodel dat gevoed is met wetenschappelijke onderzoeken, bijdragen aan de ontwikkeling van op wetenschap gebaseerd beleid. En wie wil brainstormen of sparren over bepaalde onderwerpen kan dat voortaan met een taalmodel doen in plaats van met collega's tijdens een tijdrovende brainstormsessie. Een chatbot als ChatGPT is laagdrempelig en direct beschikbaar.

Risico's

Diverse onderzoeken leggen ook de risico's van taalmodellen bloot. In dit hoofdstuk gaan we daar verder op in. Op hoofdlijnen gaat het om zaken als foute en/of schadelijke output, problemen met onderliggende bronnen, ecologische impact en risico's voor eindgebruiker. En hoewel deze risico's apart zijn beschreven, bestaat er onderling veel samenhang. Ze kunnen elkaar zelfs versterken.

Hallucinaties

Het is al langer bekend dat chatbots zoals ChatGPT (OpenAI) soms informatie verzinnen. Dat verschijnsel heet hallucineren. Taalmodellen hebben namelijk geen menselijk begrip. Het is dus mogelijk dat ze logisch ogende teksten schrijven, die inhoudelijk niet kloppen of zelfs onzinnig zijn. Dit verschijnsel kan verergeren naarmate taalmodellen steeds meer met elkaar in verbinding komen te staan, en zodoende elkaars werk gaan overnemen. Zo interpreteerde de chatbot van Microsoft (Bing) een artikel dat over een hallucinatie van ChatGPT geschreven werd verkeerd, en nam het de hallucinatie over als feit. Helaas ging het hier om een door ChatGPT verzonnen valse beschuldiging van seksuele intimidatie die met het delen door Bing in stand werd gehouden. Iedereen die taalmodellen gebruikt, zou zich bewust moeten zijn dergelijke risico's. En daarmee informatie nooit zomaar als juist aannemen. In onderstaand screenshot ziet u een voorbeeld hiervan.

Uitgeschreven tekst

Screenshot genomen op 4 oktober 2023. Het antwoord van ChatGPT is fout. De Spreidingswet is een wetsvoorstel over asielopvang.

Gebruikersinput

Er bestaat weinig duidelijkheid over hoe OpenAI omgaat met gebruikersinput. De Autoriteit Persoonsgegevens (AP) vroeg hier in juni 2023 opheldering over bij OpenAI. Deze gebruikersinput kan OpenAI namelijk opslaan en gebruiken om het datamodel verder te trainen. Mocht dat het geval zijn, dan is het niet duidelijk wat dit betekent voor de privacy van de gesprekken. Ook roept dit vragen op over het effect van potentieel schadelijke input in de output.

Het effect van dit risico voor de privacy wordt nog versterkt door lekken en hackers. De Italiaanse toezichthouder op privacy heeft ChatGPT in Italië zelfs tijdelijk verboden. Een van de redenen was dat gegevens afkomstig uit de gebruikersinput waren gelekt. Ook lagen betaalgegevens van abonnees op straat. Daarnaast blijken ChatGPT-accounts populaire doelwitten te zijn voor hackers, omdat er veel potentieel gevoelige informatie door OpenAI opgeslagen wordt. Volgens digitaal platform Mashable *s*taan er op het dark web enorm veel ChatGPT-accounts te koop aangeboden. Dit soort risico's vormen voor grote bedrijven als Apple al voldoende reden om gebruik van de tool te verbieden

Duurzaamheid

Het trainen en het gebruik van taalmodellen kost enorm veel energie. Volgens de NOS staat 1 enkele trainingssessie van ChatGPT gelijk aan ongeveer 500 ton CO2-uitstoot. Dat is gelijk aan 1000 auto's die allemaal 1000 kilometer rijden. En om dit nog verder in perspectief te plaatsen: 1 mens is per jaar verantwoordelijk voor ongeveer 5 ton CO2-uitstoot. Naast het trainen van een taalmodel kost het gebruik ervan ook veel energie. Naar schatting staat het uitvoeren van 1 opdracht gelijk aan het energiegebruik van een kamerlamp die 1 uur aanstaat.

Veel landen streven momenteel naar klimaatneutraliteit. Maar hoe combineren we die grote ecologische impact van het trainen en onderhouden van een taalmodel met dit doel? Het NOS-artikel vermeldt ook dat er in Nederland al vanuit diverse hoeken wordt gewerkt aan de verduurzaming van deze technologie. Denk dan aan meer energiezuinige chips of het gebruik van andere modellen. Dat laatste leidde al tot 30 tot 40% minder energieverbruik.

Onderliggende datasets

Een taalmodel beschikt dus over een grote hoeveelheid data op basis van teksten. Deze teksten halen ze van het internet met een proces dat bekend staat als schrapen.

De Groene Amsterdammer bekeek een dataset die veel wordt gebruikt om taalmodellen te ontwikkelen en zag dat de kans aannemelijk is dat deze set ook in ChatGPT is verwerkt. In deze dataset staat docplayer.nl bovenaan. Docplayer was lange tijd een belangrijke verzamelplek voor internetpiraten en een goudmijn voor hackers. Zij vonden hier privé-gegevens uit datalekken en sporen van rondslingerende AIVD-rapporten. Zo staan er volledig ingevulde cv's op deze site en belastingaangiften incluis de namen en BSN's van veel Nederlanders. Allemaal gegevens waarmee criminelen identiteitsfraude kunnen plegen of bij mensen kunnen inbreken. Ook marktplaats.nl maakt onderdeel uit van de website. Dat is volgens de De Groene Amsterdammer zorgelijk, omdat gebruikers daar hun telefoonnummers achterlaten.

Uit hetzelfde onderzoek bleek bovendien dat ChatGPT waarschijnlijk zeer schadelijke bronnen bevat, zoals het neonazistische Stormfront en de complotsite Vrijspreker. En hoewel het mogelijk is om filters in het taalmodel te bouwen, kunnen dit soort bronnen toch invloed hebben op de output. Via de website van Stormfront kan ChatGPT bijvoorbeeld nazistisch gedachtegoed en taalgebruik aanleren. En via Vrijspreker krijgt het een incorrect wereldbeeld. Hoe groot die invloed is, is nog niet aangetoond.

Hoe fout het kan gaan, liet de chatbot Tay van Microsoft zien in 2016. Deze chatbot, gevoed en getraind door gebruikers, herhaalde veel van de input die hij kreeg. Dat viel op bij een aantal gebruikers van Twitter (tegenwoordig X) en zij voedden hem vervolgens met racistische en misogyne taal. Een dag na de lancering was de chatbot offline. De mogelijkheid dat modellen discriminatoire associaties maken of schadelijke output leveren vanuit onderliggende bronnen, is dan ook een risico waarvoor we moeten waken.

Desinformatie, arbeidsuitbuiting en de rol van Big Tech

Dat dit soort risico's bestaan mag geen verrassing zijn. Big Tech-bedrijven uit Silicon Valley handelen al sinds jaar en dag volgens het motto *Move fast and break things*. De gedachte erachter is dat je huidige (verouderde) systemen kapot moet maken om ruimte te maken voor nieuwe, betere technologie. Het resultaat is echter dat Big Tech-bedrijven vaak problemen creëren voor de samenleving die iemand anders vervolgens mag oplossen. Of ze maken dingen kapot die iemand anders mag repareren.

Denk bijvoorbeeld aan de schade die is ontstaan aan de sociale cohesie in onze samenleving onder invloed van aanbevelingsalgoritmes van sociale mediaplatforms. Of aanbevelingsalgoritmes gericht op het aanjagen van conflict, omdat dat meer 'engagement' voor de platforms oplevert. Een ander voorbeeld is het algoritme dat Amazon inzette om nieuw personeel aan te nemen en waarmee het duidelijk vrouwen discrimineerde.

Big Tech rolt voortdurend technologieën uit die nog niet 'af' zijn en waarvan de maatschappelijke impact nog niet in kaart is gebracht. Een van de belangrijkste risico's die zich voordoet bij taalmodellen is de ontwikkeling en verspreiding van desinformatie. Een rapport van Newsguard (dat misinformatie opspoort en tegengaat) toont aan dat tientallen nieuwe nepnieuwswebsites in verschillende talen dagelijks honderden AI-gegenereerde artikelen publiceren.

Taalmodellen maken de drempel voor nepnieuws en andere vormen van desinformatie lager en de verspreiding via openbare sociale mediaplatforms sneller. Denk dan aan Facebook, maar in toenemende mate ook aan privékanalen zoals Telegram. Elementaire concepten achter onze democratische rechtsstaat, zoals waarheidsvinding en een gedeeld perspectief op de realiteit, worden door bedrijven als OpenAI onder druk gezet. En dit allemaal om zo veel mogelijk (markt)macht naar zichzelf toe te trekken en een monopolie te krijgen.

Bovendien is arbeidsuitbuiting een kernelement van de werkwijze van grote technologiebedrijven, ook bij de ontwikkeling van taalmodellen. Het trainen van AI-systemen is mensenwerk dat plaatsvindt in zeer slechte omstandigheden. OpenAI betaalt Kenianen $1,32 tot $2 per uur voor het labellen van datasets. Daarbij zitten zeer heftige en illegale teksten en afbeeldingen waardoor werknemers psychische klachten krijgen. Ook in Europa doet dit probleem zich voor. Het gaat dan om digitaal werk vanuit huis dat in de vorm van kleine taken via platforms toebedeeld worden. Door deze constructie hebben de werkers geen recht op het minimumloon, ziekteverlof of sociale zekerheid.

Beleids- en juridische ontwikkelingen

Europa

Een van de grootste ontwikkelingen op juridisch en beleidsgebied is de Europese AI Act. Deze Wet op Kunstmatige Intelligentie schept kaders voor bedrijven en organisaties die AI-oplossingen bieden. De wet is nog in ontwikkeling en wordt naar verwachting begin januari 2024 aangenomen. De wet onderscheidt 4 risicoschalen:

- Onaanvaardbaar risico: voortkomend uit de schending van grondrechten of uit blootstelling van de veiligheid van mensen.
- Hoog risico: volgt aanpassingen aan zowel onderwijs als medische hulpmiddelen.
- Beperkt risico: duidt op lichter risico op het gebied van gezondheid en veiligheid (zoals voice-/chatbots).
- Minimaal risico: het niveau dat aspecten dekt, zoals zoekmachines of aanbevelingsalgoritmen.

Door de komst van ChatGPT moet de huidige AI Act worden aangepast. De Europese wetgevers hielden nog geen rekening met het gebruik van generatieve AI op publiek niveau toen zij dit wetvoorstel schreven. Zij gingen ervan uit dat regelgeving voor professioneel gebruik belangrijker was, aangezien generatieve AI toen nog enkel op kleine schaal bestond. Ook is de AI Act op sommige onderdelen van het generatieve AI-proces niet nauwkeurig genoeg. Zo bepaalt de eindgebruiker, en niet de aanbieder van AI, wat ermee gedaan wordt. Terwijl die eindgebruiker nauwelijks voorkomt in de huidige AI Act.

Kabinetsvisie Generatieve AI

Het kabinet werkt momenteel aan een visie over generatieve AI waarin de kansen en risico's van deze systemen belicht worden. Waarschijnlijk is er veel aandacht voor veiligheid, rechtvaardigheid, transparantie en brede welvaart. Ook verwachten we dat er aandacht wordt besteed aan de reguleringsopgave rondom taalmodellen, inclusief richtsnoeren voor het gebruik van taalmodellen door (rijks)ambtenaren. De VNG is betrokken bij de ontwikkeling van de visie die begin 2024 naar de Kamer zal worden gestuurd.

Vanuit de onderzoeksorganisatie TNO en het ministerie van Economische Zaken en Klimaat (EZK) is een initiatief gestart voor het ontwikkelen van een eigen taalmodel dat getraind moet worden op eigen, betrouwbare data. Omdat dit initiatief nog in zeer vroeg stadium zit, kan er inhoudelijk nog weinig over worden gezegd. Een eigen (overheids)taalmodel kan in theorie met behulp van betrouwbare data het risico van inmenging van foutieve of schadelijke bronnen beperken. Ook publieke waarden en juridische problemen rond auteursrecht krijgen zo de nodige aandacht. Zo'n eigen taalmodel kan transparanter, veiliger en meer betrouwbaar zijn.

Toezichthouders privacy

Toezichthouders dringen aan op transparantie over taalmodellen. Zij willen daarmee voorkomen dat er potentiële onwettelijkheden ontstaan. De Auroriteit Persoonsgegevens (AP) stelt dat ChatGPT onder meer met gebruikersinput gegevens kan verzamelen. De toezichthouder waarschuwt ervoor dat dit soort data gevoelige of persoonlijke informatie kan bevatten. Denk aan een gebruiker die advies vraagt over een echtelijke ruzie of medische zaken.

In andere Europese landen speelt dezelfde discussie over persoonsgegevens en het gebruik van ChatGPT. Eerder al noemden we het tijdelijke verbod in Italië. Volgens de Italiaanse toezichthouder is OpenAI niet duidelijk over welke gegevens het verzamelt en wat het daarmee doet. Bovendien bevatte de chatbot geen systeem waarmee het de leeftijd van minderjarigen kan achterhalen. Gevolg: OpenAI moest binnen 20 dagen een oplossing bieden. Volgens een woordvoerder van de AP is een verbod van ChatGPT zoals in Italië hier nog niet aan de orde.

Het proces van het schrapen van online beschikbare teksten en gebruik voor trainen van taalmodellen roept ook auteursrechtelijke vragen op. In de Verenigde Staten zijn meerdere rechtszaken aangespannen tegen OpenAI voor het gebruik van datasets die illegaal gekopieerde werken bevatten. Ook in Nederland vinden deze praktijken plaats, zoals onderzoekers van de Groene Amsterdammer bevonden. Juridisch gezien zijn er

echter nog te veel zaken onduidelijk op dit gebied en hebben we gewoonweg nog niet alle antwoorden.