Verslaglegging Expertsessies Representatiebias

GPT-NL

Verslaglegging Expertsessies Representatiebias

Auteurs: Duuk Baten (SURF)
Dominique Blok (TNO)
Lieke Dom (TNO)
Eliza Hobo (TNO)
Merel van Steenis (SURF)

Document ID GPTNL-RAP-25-006
Versie: 1
Datum: 29 augustus 2025

GPT-NL is een samenwerking van TNO, SURF, en het NFI
Financiering: Faciliteiten Toegepast Onderzoek (FTO) vanuit RVO/EZK

www.gpt-nl.nl
info@gpt-nl.nl
www.linkedin.com/company/gpt-nl/

Inhoudsopgave

Executive summary

Management samenvatting (NL)

Het ontwikkelen van taalmodellen brengt diverse uitdagingen met zich mee. In dit rapport delen we één van de vraagstukken die binnen het project GPT-NL voorbij kwam: hoe gaan we om met representatiebias?

Kunstmatige Intelligentie (AI), meer in het bijzonder Generatieve AI, zoekt naar patronen in grote datasets en leert zo te classificeren of voorspellingen te doen. Bij taalmodellen (LLMs) werkt dit hetzelfde: een algoritme analyseert grote hoeveelheden tekst en 'leert' zo welke woorden elkaar logischerwijs op lijken te volgen. Een inherent risico is echter dat taalmodellen (en AI in het algemeen) onbedoelde of ongewenste patronen versterken die het in de dataset herkent. Dit noemen we *bias*. Het gevolg is dat een homogene output ontstaat die werkt voor grote groepen, maar niet voor kleine groepen. We spreken dan van *representatiebias*.

Om de kans op representatiebias te verkleinen, is het van belang een zo divers mogelijke dataset te hebben die zoveel mogelijk groepen representeert. Echter leidt dit tot een complexe uitdaging: wanneer is een dataset 'representatief genoeg'? Hoe meet je dat überhaupt? Met welke risico op bias kan je genoegen nemen? En met welke groepen moet je dan rekening houden?

In de wetenschap is vooral kennis over bias in *AI systemen* (dus in de *toepassing* van AI). Er zijn echter nog weinig best practices over het mitigeren van bias in de eerste fase van de ontwikkeling van taalmodellen: het creëren van de dataset. Om brede inzichten op te halen om de vragen over representatiebias in de eerste ontwikkelfase van GPT-NL te beantwoorden, hebben we in september 2024 twee sessies georganiseerd met experts op het gebied van representatie en discriminatie. We zijn de discussie gestart met een openingsvraag: 'Welke zorgen hebben jullie rondom bias bij taalmodellen?' Vervolgens zijn we in discussie gegaan over de vragen:

1) Wat is 'goede' representatie?
2) Op welke (ondergerepresenteerde) groepen moeten we letten?
3) Hoe meet je representatie?
4) Hoe mitigeert je representatiebias?
5) Wanneer is het 'goed genoeg'?

Vanuit de gesprekken met deze experts hebben we een aantal aanbevelingen verzameld in vijf thema's:

1) prioriteer transparantie,
2) volledige representatie is niet mogelijk; wees voorzichtig met 'oplossingen' voor representatiebias,

3) geef ook aandacht aan het gebruik en de gebruiker,
4) Werk samen met bestaande initiatieven en
5) kijk verder dan de data-acquisitie en -curatie fases.

Gebaseerd op deze inzichten reflecteren we kort op het GPT-NL project en hoe we hier mee aan de slag zijn gegaan.

Executive summary (ENG)

Developing language models presents various challenges. In this report, we address one of the issues that arose within the GPT-NL project: how to deal with representation bias.

Artificial Intelligence (AI) searches for patterns in large datasets and learns to classify or make predictions. Language models (LLMs) operate similarly: an algorithm analyzes vast amounts of text and "learns" which words logically follow each other. However, there is an inherent risk that language models (and AI in general) reinforce unintended or undesirable patterns recognized in the dataset. This phenomenon is known as "bias." The result is a homogeneous output that works for large groups but not for smaller ones, referred to as "representation bias."

To mitigate representation bias, it is crucial to have a dataset that is as diverse as possible, representing a wide array of groups. However, this introduces a complex challenge: when is a dataset "representative enough"? How do you measure that? What level of bias risk are you willing to accept? And which groups need to be considered?

Academic resources predominantly discuss bias in AI systems (i.e., in the application of AI). However, there are few best practices for mitigating bias in the initial phase of LLM development: the creation of the dataset. To gain broad insights into addressing representation bias in the first development phase of GPT-NL, we organized two sessions in September 2024 with experts in representation and discrimination. We started the discussion with an opening question: ‘What concerns do you have about bias in language models?’ Then we discussed the following questions:

1) What constitutes "good" representation?
2) Which (underrepresented) groups should we focus on?
3) How do you measure representation?
4) How do you mitigate representation bias?
5) When is it "good enough"?

Based on these discussions, we compiled several recommendations under five themes:

i) prioritize transparency,

ii) ii) acknowledge that full representation is not possible and be cautious with "solutions" for representation bias,
iii) iii) pay attention to the use and the user,
iv) iv) collaborate with existing initiatives and
v) v) look beyond the data acquisition and curation phases.

Reflecting on these insights, we briefly review the GPT-NL project and our approach to addressing representation bias.

Our report is written in Dutch. However, to share the learnings the 'recommendations' section is also translated to English in the appendix.

Introductie

Het ontwikkelen van een taalmodel brengt ethische, juridische en technische uitdagingen met zich mee. Eén van die uitdagingen is het vraagstuk hoe we om moeten gaan met (representatie)bias en hoe we dit kunnen mitigeren.

Bij 'representatiebias' is de dataset niet representatief voor de samenleving of populatie waarin het AI systeem wordt gebruikt. Er treedt daardoor een bias ("vooringenomenheid") op waarbij het AI systeem reflecteert wat het in de dataset heeft gezien, maar dit leidt niet altijd tot een wenselijke uitkomst. Een bekend voorbeeld komt uit de geneeskundige hoek: huidige AI systemen zijn getraind op datasets waarin er meer beschreven is over mannelijke artsen dan over vrouwelijke artsen. Het gevolg is dat een taalmodel model bovenal naar 'hij' of 'hem' verwijst als het om een arts gaat. Bijvoorbeeld, bij de vraag om de zin 'de arts verzekerde de patiënt van goede zorg' te vereenvoudigen zal dan vaker geformuleerd worden als 'De arts zei dat *hij* goed voor de patiënt zal zorgen' in plaats van 'De arts zei dat *zij* goed voor de patiënt zal zorgen' of 'De arts zei dat *die* goed voor de patiënt zal zorgen'. Wanneer het model voortdurend naar een mannelijke arts verwijst, zoals in dit voorbeeld, heeft het systeem dus een vertekening overgenomen uit de dataset. Hierdoor is het model niet representatief voor hoe de populatie artsen eruit ziet (of eruit zou kunnen zien).

Omdat de samenstelling van de trainingsdata bepalend is voor wat het taalmodel weerspiegelt, is het cruciaal om al in de data-acquisitiefase en datacuratie fase bewust om te gaan met de risico's op bias en discriminatie. Hier is echter nog weinig onderzoek naar gedaan. Er is veel literatuur te vinden is over hoe representatiebias door AI modellen tot uiting komt, maar nog niet hoe dit op het niveau van datasetcreatie gemitigeerd[1] kan worden. Om extra inzichten op te halen, organiseerden we daarom twee bijeenkomsten voor experts op het onderwerp representatie, discriminatie, en biasmitigatie. Tijdens deze bijeenkomsten zijn we een open gesprek aangegaan over representatiebias. Middels de expertsessies wilden we extra inzicht verkrijgen in de representatie van bepaalde groepen. Daarnaast wilden we kennis vergaren op mogelijkheden voor *data augmentatie*, waarbij door te filteren of teksten te herschrijven, de representatie in de dataset kan worden verbeterd. In de expertsessies testen we onze aannames en onderzoeken we andere visies, uitgangspunten en methoden.

Disclaimer

Verslaglegging

Dit verslag beschrijft de discussies en inzichten die we hebben opgehaald in deze bijeenkomsten. We delen deze inzichten om transparant te zijn over de ontwikkeling

[1] 'Gemitigeerd' of 'bias mitigatie' is het verkleinen van het risico op bias

van GPT-NL. Daarnaast hopen wij middels dit verslag de dialoog over taalmodellen en representatiebias te voeden. Let wel: de sessies hadden niet het doel om een kenniskloof te dichten, dit verslag presenteert geen wetenschappelijke onderzoeksresultaten.

Groepen
We spreken in dit verslag van 'groepen'. Hiermee kunnen we verwijzen naar bijvoorbeeld 'bevolkingsgroepen', 'demografische groepen', of 'culturele groepen'. Ook spreken we soms van 'minder gerepresenteerde' of 'ondergerepresenteerde' groepen. We zijn ons ervan bewust dat cultuur, identiteit en verbinding tussen mensen niet in losse groepen te beschrijven is. Daarnaast kunnen mensen zich verbonden voelen met meerdere groepen en communities (ook wel: intersectionaliteit). We spreken echter van 'groepen' voor het gemak van gespreksvoering en verslaglegging.

Aanpak expertsessies

Tijdslijn:

1. Augustus 2024: 24 uitnodigingen verstuurd naar expert stakeholders binnen Nederland met twee mogelijke data.
2. September 2024: twee expert bijeenkomsten fysiek op locatie (Utrecht)
    a. 12 september en 19 september
    b. Respectievelijk 4 en 3 deelnemers per sessie

De bijeenkomsten duurden 2,5 uur. Deelname aan de bijeenkomsten was onbezoldigd. De deelnemers zijn terug te vinden in de colofon.

Tijdens de bijeenkomsten hebben we een aantal vragen aan de deelnemers voorgelegd:

1) Openingsvraag: Welke zorgen rondom bias bij taalmodellen komt direct bij jullie op?
2) Discussievraag 1: Wat is goede representatie?
3) Discussievraag 2: Voorstel voor ondergerepresenteerde groepen – zijn dit de juiste groepen?
4) Discussievraag 3: Wat is de beste methode om representatie te meten?
5) Discussievraag 4: Wat is de beste methode om representatiebias te mitigeren?
6) Discussievraag 5: Wanneer is het goed genoeg?

Vanuit het GPT-NL team waren bij beide sessies vier mensen aanwezig om het gesprek te faciliteren en te notuleren.

Inzichten uit de expertsessies

In deze sectie van het verslag zijn de inzichten uit de expertbijeenkomsten gepresenteerd aan de hand van de gestelde discussievragen.

Discussievraag 1: Wat is goede representatie?

Introductie van de vraag

Voordat we weten welke stappen we kunnen nemen om een representatieve dataset te waarborgen, willen we eerst bespreken wat een 'goede' representatie is. Een beoordeling van 'goede' representatie roept meerdere vragen op, zoals:

1. Gaat representatie om Nederland 'nu', Nederland 'in de toekomst' of Nederland 'nu met extra aandacht voor minder-gerepresenteerde groepen'. Met andere woorden: moet het taalmodel de huidige Nederlandse samenleving zo accuraat mogelijk weerspiegelen ("beschrijven"), zo accuraat mogelijk een inschatting maken van de Nederlandse samenleving in de toekomst (bijvoorbeeld op basis van statistieken) of moeten we juist streven naar een ideale en inclusievere versie van de samenleving (en daarmee een normatief standpunt innemen)? Vooral het laatste geval leidt weer tot extra vragen: wat zijn dan deze idealen, aan welke (bevolkings)groepen moeten we dan meer aandacht geven, wanneer is het 'inclusief genoeg', en hoe prioriteren we deze groepen tijdens de ontwikkeling van het model en curatie van de dataset?
2. Wat is überhaupt een descriptieve, statistische of accurate representatie van Nederland? Welke (demografische of bevolkings)groepen neem je dan mee en voor welke toepassingen?
3. Gaat representatie over wélke groepen worden gerepresenteerd of hóe groepen worden gerepresenteerd (of beide)?
4. Welke mitigaties op representatiebias passen bij welke fase in de curatiefase?

Discussie

Trade-off kwaliteit - representatie
Uit de discussie bleek dat het belangrijk is om de data in grove lijnen de normen en waarden van de Nederlandse samenleving te laten representeren. Echter bleek het onduidelijk wat 'goede' representatie is en dat het lastig is om vooraf te bepalen welke data je wel en niet mee moet nemen om goede kwaliteit én goede representatie van bepaalde groepen te waarborgen. Data van sociale media worden, bijvoorbeeld, doorgaans niet als hoge kwaliteit teksten beschouwd. Zou je echter alle data van sociale media weglaten, dan kan dit leiden tot een vertekening van de mening van de Nederlandse maatschappij. Een ander voorbeeld is straattaal: zou je bepaalde woorden en zinnen eruit filteren om te zorgen dat het taalmodel teksten teruggeeft volgens Nederlandse spelling en grammatica, dan ontstaat daarmee mogelijk ook het risico dat het model taal en teksten van bepaalde groepen niet herkent.

Representatie van Nederland

De vraag of het taalmodel moet kijken naar Nederland nu, later zoals we het statistisch kunnen verwachten of een gewenste samenleving in de toekomst (normatief), bleek ook lastig. Speculatie over welke (bevolkings)groepen er later in Nederland zijn te onderscheiden is lastig, omdat de samenleving zal blijven veranderen. Het wordt daarom afgeraden om keuzes voor dataselectie aan te passen op speculaties over de toekomst.

Hoe gerepresenteerd?
Representatie gaat niet alleen over **wélke** groepen worden weerspiegeld in de dataset, maar ook **hóe** groepen worden weerspiegeld. De statistische representatie in een dataset leidt niet per se tot de gewenste output in het taalmodel en is daarom niet toereikend.
Het is ook belangrijk om bepaalde ongewenstheden in de data te identificeren, denk aan: schadelijke, kwetsende, racistische of discriminerende taal en uitdrukkingen (**harmful language**). Het is dan wel noodzaak om als organisatie te bepalen wat wordt beschouwd als schadelijk, racistisch, e.d.. Hierbij is het belangrijk om transparant te zijn over deze keuzes en ze goed te documenteren.

Belang van diverse data
De groep benadrukt het belang van het verzamelen van diverse data, al is er soms van bepaalde groepen niet genoeg beschikbaar. Er wordt gerefereerd naar het voorbeeld van gezondheidsdata: doordat er meer medisch onderzoek is gedaan bij mannen, werken de uitkomsten van modellen beter voor mannen dan voor vrouwen. Daarom is het voor GPT-NL belangrijk om bij de data-acquisitie al voldoende diverse data te verzamelen. **En als die data niet beschikbaar is, dat goed te documenteren**.
In sommige gevallen is het ook mogelijk corrigerende maatregelen te nemen. Te denken valt aan het beperkt *over-samplen* of *under-samplen*[2] van data over bepaalde groepen of geschreven door bepaalde groepen. Zo kan het balanceren van voornaamwoorden zorgen voor een beter (i.e., inclusiever) taalgebruik door een model. Maar brengt ook het risico mee van mogelijke overcorrectie of het toevoegen van incorrectheden in de data.

Verschillende ontwikkelfases
Tot slot werd benoemd dat representatie niet alleen tijdens de dataverzameling en -curatie een rol kan (of moet) spelen. Dit kan ook later tijdens de ontwikkeling, in de fine-tuning fase. Dit is het proces waarbij het taalmodel wordt aangepast aan een specifieke taak. Als de trainingsdata zich in het gedrag van het model niet ‘gewenst' uit, kan dat in de fine-tuning fase worden rechtgetrokken. Het is dus van belang om het getrainde model te toetsen op representatie, om het vervolgens goed te kunnen bijsturen in fine-tuning.
Bovendien is niet bekend wat het precieze effect van de trainingdata is op het gedrag van het model. Overmatig aanpassen van de samenstelling van de dataset kan onverwachte en ongewenste effecten hebben. Wanneer een model al getraind is, kan er gerichter tewerk gegaan worden.

[2] Het over- of under-samplen van data betekent dat je een bepaald stuk uit de dataset vaker of minder vaak aan het model laat zien om van te ‘leren’.

Bevindingen uit de groep

Wat betreft de eerste vraag over representatie nu, later, of normatief, raadt de groep af om te speculeren over een Nederland in de toekomst of hoe Nederland eruit zou moeten zien. Er was geen consensus over wat dan wél een 'goede' (beschrijvende/statistische/normatieve) representatie van Nederland is. Er werd verder benadrukt dat representatie gaat over zowel wélke groepen worden gerepresenteerd, als hóe groepen worden gerepresenteerd. Het mitigeren op (representatie)bias vindt plaats in meerdere fasen van de ontwikkelfase.

De volgende concrete aanbevelingen volgen uit de discussie:

1. Maak **expliciete keuzes** (bijvoorbeeld over het wel of niet filteren van racistische of discriminerende uitspraken) en deel publiek hoe we dat doen. Wees helder over wélke data je eruit filtert en documenteer waarom (dus welke data je onder harmful of discriminerend schaart).
2. **Documenteer** welke keuzes er worden gemaakt en bewaar alle data (ook verwijderde of gefilterde data).
3. **Redeneer vanuit toekomstige gebruikerscontext** over de mogelijk schadelijke gevolgen van keuzes in het model. Op basis daarvan kan je terug redeneren naar keuzes over datacuratie.
4. Zoek (verdere) **samenwerking** op met al bestaande initiatieven op het gebied van (representatie-)bias in taalmodellen. Er is in andere projecten ook al veel geleerd.
5. Focus op representatie van het **huidige** Nederland, laat een toekomstig of het gewenste Nederland buiten beschouwing.
6. Zorg dat **diversiteit een prioriteit** heeft vanaf het begin van de ontwikkeling, al in de data-acquisitiefase. Niet alles is in een latere fase te corrigeren.
7. Kijk ook naar mogelijkheden om in **latere fases** in de ontwikkeling op (representatie)bias te mitigeren.
8. **Voer experimenten uit** met data-augmentatie en fine-tuning om op representatie te optimaliseren.

Discussievraag 2: Op welke ondergerepresenteerde groepen moeten we focussen?

Introductie van de vraag

Om een volledig representatieve dataset te creëren, zouden we van elke groep data moeten hebben. Zoals ook besproken bij Discussievraag 1, is er van sommige groepen minder data beschikbaar of is het moeilijk om die data te verkrijgen. Om in kaart te brengen hoe representatief een dataset is, is het daarom van belang om te onderzoeken om welke groepen dit gaat. En met name om de representativiteit van een dataset te verhogen, moeten we inschatten welke groepen het 'risico' lopen in mindere mate gerepresenteerd te zijn. Omdat gebrek aan representatie discriminatie tot gevolg kan hebben, zijn we uitgegaan van de (gronden) waarop discriminatie wettelijk niet is toegestaan:[3]

- Herkomst/huidskleur ('ras' genoemd in juridische zin)
- Leeftijd
- Politieke gezindheid
- Burgerlijke staat
- Geloof
- Levensovertuiging
- Seksuele gerichtheid
- Handicap of chronische ziekte
- Geslacht
- Nationaliteit

In een poging dit door te vertalen naar groepen waar we representatie mee zouden kunnen meten, komen we tot de volgende lijst:

- **Gender**[4], bijvoorbeeld: vrouwen, transgender, intersekse mensen.
- **Leeftijd**[56], bijvoorbeeld: ouderen.
- **Etniciteit**[78], bijvoorbeeld: de meest voorkomende migrantengroepen in Nederland uit het Caribisch gebied, Suriname, Marokko, en Turkije.

[3] Overgenomen van https://discriminatie.nl/wat-is-discriminatie/
[4] https://www.mensenrechten.nl/themas/gendergelijkheid
[5] https://www.rijksoverheid.nl/onderwerpen/gelijke-behandeling-op-het-werk/leeftijdsdiscriminatie-op-de-arbeidsmarkt
[6] https://www.cbs.nl/-/media/_pdf/2018/11/sociaal-in-de-marge.pdf
[7] https://www.cbs.nl/nl-nl/dossier/dossier-asiel-migratie-en-integratie/hoeveel-inwoners-hebben-een-herkomst-buiten-nederland
[8] https://www.amnesty.nl/wat-we-doen/themas/discriminatie/nederland-discriminatie-vanwege-afkomst

- **Religie**[9][10][11], bijvoorbeeld: de meest voorkomende religies in Nederland: Katholicisme, Islam, Protestantisme, Boeddhisme, Hindoeïsme, Jodendom
- **Seksuele oriëntatie**[12][13], bijvoorbeeld: Lesbisch, homoseksueel, biseksueel, panseksueel
- **Personen met een beperking** [14][15], bijvoorbeeld: personen met beperkt gehoor of gezichtsvermogen, met psychologische problemen, of fysieke problemen.
- **Klasse/Sociale status/Inkomen/Opleiding**[16][17], bijvoorbeeld: laaggeletterden, mensen in de schulden, of mensen zonder opleiding.

Deze groepen zien we niet als een totaal overzicht. Er zullen dus groepen zijn die hierin gemist worden of niet gerepresenteerd worden op het detail niveau dat ideaal zou zijn. Dit leidt tot de volgende vragen:

1. Is representatie te kwantificeren; Is bijvoorbeeld een lijst van groepen (zoals hierboven) een goede manier om representatie te meten?
2. Zo ja, welke groepen moeten op een dergelijke lijst staan?
3. Hoe kunnen we ermee omgaan dat een lijst nooit iedereen kan beschrijven?
4. In hoeverre kunnen cultuur, normen en waarden van deze groepen gerepresenteerd worden in de data?

Discussie

De lijst

De aanwezigen waren niet op de hoogte van een bestaande lijst die dergelijke groepen opsomt. Uit de discussie kwam naar voren dat het maken van een lijst zoals hierboven een duidelijk doel moet hebben, zoals inzicht krijgen in de data. Er zijn echter verschillende risico's bij het gebruik van een dergelijke lijst: het kan categoriserend werken, en mensen reduceren tot bepaalde eigenschappen. Ook is er geen ruimte voor intersectionaliteit omdat mensen onderdeel kunnen zijn van meerdere ondergerepresenteerde groepen, waardoor ze andere vormen van onderrepresentatie

[9] https://longreads.cbs.nl/nederland-in-cijfers-2023/welk-geloof-hangen-we-aan/
[10] https://www.rijksoverheid.nl/actueel/nieuws/2024/04/23/reactie-ncab-op-antisemitismecijfers-om-en-politie-verharding-antisemitisme.-het-blijft-niet-meer-bij-woorden-alleen
[11] https://www.rijksoverheid.nl/actueel/nieuws/2024/04/23/reactie-ncab-op-antisemitismecijfers-om-en-politie-verharding-antisemitisme.-het-blijft-niet-meer-bij-woorden-alleen
[12] https://www.rijksoverheid.nl/onderwerpen/lhbti-emancipatie/bestrijden-geweld-en-discriminatie-tegen-lhbtis
[13] https://discriminatie.nl/discriminatiegronden/seksuele-gerichtheid/#:~:text=van%20seksuele%20gerichtheid%3F-,Wat%20is%20discriminatie%20op%20basis%20van%20seksuele%20gerichtheid%3F,die%20in%20de%20wet%20staat
[14] https://www.rijksoverheid.nl/onderwerpen/rechten-van-mensen-met-een-handicap
[15] https://www.scp.nl/publicaties/publicaties/2020/04/02/ervaren-discriminatie-in-nederland-ii
[16] https://www.cbs.nl/-/media/_pdf/2018/11/sociaal-in-de-marge.pdf
[17] https://www.scp.nl/publicaties/publicaties/2023/03/07/eigentijdse-ongelijkheid

ervaren. Het is daardoor van belang de intersecties van deze groepen in de gaten te houden.

Een mogelijk alternatief voor het beginnen bij de groepen die mogelijke nadelige gevolgen kunnen ervaren, is om te beginnen met het schetsen van scenario's waarin het taalmodel gebruikt kan worden. Door vanuit deze scenario's vast te stellen welke groepen er in verschillende contexten kwetsbaar zijn, kan er op een gerichte manier getoetst worden op misrepresentatie.

Andere groepen

In het gesprek werden verschillende sociale groepen en eigenschappen genoemd die mogelijk ondergerepresenteerd worden en nog niet op de lijst staan. Zo werd er tijdens de discussie gewezen op de manier waarop de nieuwe AI-verordening het begrip 'kwetsbaarheid' benadert. Waar eerder de focus lag op ondergerepresenteerde of kwetsbare groepen zoals mensen met een handicap, houdt de AI-verordening ook rekening met andere factoren die een persoon kwetsbaar kunnen maken. Denk bijvoorbeeld aan verslaving, politieke oriëntatie of regionale afkomst. Dit zijn ook factoren die mensen extra gevoelig kunnen maken voor bias in taalmodellen. Dergelijke eigenschappen zouden kunnen worden meegenomen GPT-NL zou hier een voorbeeld aan kunnen nemen.

Een ondergepresenteerde groep die in het bijzonder toegevoegd kan worden, zijn laaggeletterden. Het meenemen van gesimplificeerde tekstdata in het taalmodel kan waardevol zijn voor de inclusie van laaggeletterden.

Een belangrijk inzicht over de samenstelling van de dataset, is dat het taalmodel mogelijk voornamelijk Nederlandse en Engelse teksten bevat. Dat zorgt voor ondervertegenwoordiging van anderstalige bevolkingsgroepen, niet alleen wat betreft hun taal. Teksten die in het Nederlands geschreven zijn, weerspiegelen de normen, waarden en culturele realiteit van de Nederlandstalige samenleving. Nederlandse bevolkingsgroepen die een andere taal spreken, of groepen met andere culturele of religieuze achtergronden, worden hier mogelijk minder door gerepresenteerd.

Transparantie

Het is lastig, al dan niet onmogelijk, om volledig te corrigeren voor misrepresentatie van groepen. Transparantie over de beperking van de dataset is belangrijk. Zo kunnen we bijvoorbeeld transparantie bieden over de mate waarin bepaalde groepen in de data gerepresenteerd zijn. Bijvoorbeeld door inzicht te geven in de auteurs van de verschillende teksten of over welke thema's in de data gerepresenteerd zijn. Dit biedt gebruikers van het model inzicht in de mate waarin verschillende groepen zijn vertegenwoordigd. Een kwestie die hierbij komt kijken is dat het formaliseren van

sociale concepten naar data uitdagend is. Hoe vertalen wij maatschappelijke concepten als religie of etniciteit naar data?

Door of over
Tenslotte kwam ook in dit gesprek naar voren dat niet alleen het onderwerp van de tekst, maar ook de auteur van belang zijn voor representatie. Een belangrijk nuance zin namelijk in het verschil tussen teksten geschreven door bepaalde groepen of juist over bepaalde groepen. Er is een sterke link tussen wie een tekst schrijft en hoe een tekst geschreven wordt. Een tekst geschreven dóór iemand uit een ondergerepresenteerde groep over de bijbehorende cultuur verschilt van een tekst die geschreven is óver een ondergerepresenteerde cultuur door een auteur uit een maatschappelijk dominante groep.

Bevindingen uit de groep

Wat betreft het gebruik van een lijst van ondergerepresenteerde groepen, was er niet een eenduidig beeld. Aan de ene kant is het een goed middel voor het verkrijgen van inzicht, maar aan de andere kant is het niet toereikend. Het kan categoriserend werken, groepen buitensluiten, en het houdt geen rekening met intersectionaliteit. Voor het verder uitwerken van een dergelijke lijst, raden de experts aan om contact op te nemen met mensenrechtenorganisaties. Een alternatief is te beginnen bij de concrete toepassingen en inzicht te krijgen in de concrete biases die optreden door het uitwerken van scenario's. Echter is ook dit een grote uitdaging, er zijn immers eindeloze scenario's denkbaar voor general-purpose taalmodellen zoals GPT-NL.

Tenslotte, benadrukt men het belang van transparant te zijn over mogelijke tekortkoming die het gevolg kunnen zijn van gemaakte keuzes en de kwaliteiten en beperkingen van de dataset. Dit is een eerste versie van het model GPT-NL. De groep concludeert dat het niet onoverkomelijk is als de toepassing van het taalmodel binnen bepaalde contexten nog niet 'goed genoeg' is bij de eerste oplevering.

Discussievraag 3: Wat is de beste methode om representatie te meten?

Introductie van de vraag

In Data Science wordt representatiebias meestal gemeten door een getraind model toe te passen op verschillende groepen en de prestaties van het model te vergelijken. Bijvoorbeeld: bij een gezichtsherkenningsalgoritme wordt vergeleken hoe deze presteert op mensen met een lichtere of donkerdere huidskleur. Stel dat blijkt dat het algoritme in negen van de tien gevallen werkt voor mensen met een lichte huid, maar slechts in zes van de tien gevallen voor mensen met een donkere huid, kan meetbaar worden gemaakt wat de representatiebias is voor deze twee groepen. Bestaande best practice methodes voor het meten van representatiebias in tekstuele data waren ons niet bekend ten tijden van deze expert-sessies. Representatiebias kan zich voordoen als gevolg van denigratie, stereotypering of onderrepresentatie in de teksten.[18] Het is dus belangrijk deze fenomenen in de data te herkennen. Dat roept meteen vragen op:

1. Wat zijn methoden voor het herkennen van representatiebias in datasets?
2. Wat zijn de sterke en zwakke punten van deze methodes?

Discussie

Uit de discussie kwam naar voren dat het meten van representatiebias inderdaad een complex vraagstuk is. Goede representatie draagt bij aan het verminderen van bias, maar voorkomt dit niet volledig. Tijdens het meten van bias in een taalmodel is het belangrijk om te focussen op systematische bias; terugkerende patronen van vooroordelen, stereotypering of onderrepresentatie. Juist vanwege de complexiteit van de taak, gaf de groep het advies te beginnen met simpele NLP-technieken om representatie te meten, zoals:

- *Topic classification* (over welke groepen gaat de tekst)
- *Co-occurence analysis* (welke woorden komen vaak voor bij deze groep)
- *Wordclouds* (visuele weergave van veelvoorkomende woorden).

Er werd herhaald dat niet alleen het onderwerp van de tekst ertoe doet maar ook de context, zoals de eigenschappen van de auteur en het jaartal van publicatie.

Tenslotte werd besproken dat het ook waardevol is om representatiebias te meten door te experimenteren met het model. Door het taalmodel in verschillende contexten te testen kan meer inzicht worden verkregen in de aanwezige biases.

[18] Shahbazi, Nima, et al. "Representation bias in data: A survey on identification and resolution techniques." *ACM Computing Surveys* (2023)

Bevindingen uit de groep

Meten van representatiebias

Uit de discussie kwam naar voren dat er nog geen methoden zijn die daadwerkelijk gefocust zijn op het meten van representatiebias in tekstuele datasets. De groep raadt wel aan om algemene NLP-technieken te gebruiken om inzicht te krijgen in het onderwerp van de tekst.

Experimenteren

Daarnaast werd besproken dat experimenten met het getrainde model ook van belang zijn. Zo kan inzicht worden verkregen in de scenario's waar het model wel en niet verantwoord kan worden ingezet. Deze inzichten kunnen vervolgens fungeren als een soort 'handleiding' om het gebruik van het model te ondersteunen.

Transparantie

Tenslotte benadrukte de groep dat alle opgedane kennis en inzichten waardevol kunnen zijn. Ons werd aangeraden om inzichtelijk te maken welke methoden we gebruiken en wat de voor- en nadelen daarvan zijn. Dergelijke transparantie helpt zowel gebruikers als toekomstige doorontwikkelaars.

Concrete aanbevelingen zijn:

- Start met traditionele NLP-technieken om een beeld te vormen van de representatie in de dataset
- Experimenteer ook met het getrainde model
- Wees transparant over de gemaakte keuzes, de redenen daarvoor, en de gevolgen ervan
- Maak een gebruikershandleiding die de gebruiker van de risico's op de hoogte brengt.

Discussievraag 4: Wat is de beste methode om representatiebias te mitigeren?

Introductie van de vraag

Na de detectie van representatiebias, kunnen er mitigatiestappen worden ondernomen. Wij denken daarbij aan data-augmentatie, waarbij weaanpassingen doen om de datavariatie in een bestaande set te vergroten, (bijvoorbeeld: voor een tekst waarin wordt gerefereerd naar mannelijke artsen, maken we een variant waarin ook artsen van andere genders, het toevoegen van (synthetische) data om groepen in de data te vertegenwoordigen, het filteren (weglaten) van data en het verzamelen van nieuwe data. Voor het gebruik van deze methoden voor biasmitigatie ontbreken best practices, dus komen de volgende vragen naar voren:

1. Wat zijn de voor- en nadelen van data-augmentatie, data filteren en nieuwe data verzamelen?
2. Zijn er andere methoden voor representatiebiasmitigatie?
3. Hoe kunnen we omgaan met het maken van keuzes?

Discussie

De groep is het eens dat de door ons genoemde methodes degene zijn om aan te denken. De voor- en nadelen van de methoden werden besproken.

Data-augmentatie kan helpen de datavariatie te vergroten door bestaande data op verschillende manieren te bewerken. Zo kunnen bepaalde woorden worden vervangen door synoniemen, kunnen zinsstructuren worden veranderd, of persoonlijk voornaamwoord ‘zij’ worden vervangen door ‘hij’. Dit kan helpen bij het verminderen van bias van ondergerepresenteerde groepen.

Echter zijn dergelijke aanpassingen erg risicovol. Het kan leiden tot incorrecte data (*‘hij is de dochter van…’*) of mislukte generalisatie. Het aanpassen van data kan nieuwe bias introduceren, bestaande bias versterken of leiden tot verlies van betekenis. Bovendien is niet te voorspellen wat het exacte gevolg van dergelijke aanpassingen is op het gedrag van het model. Het is dus van belang dit voorzichtig te doen, en zodanig dat het uitgebreid getest kan worden. Als dit niet secuur gedaan kan worden, is het mogelijk beter het niet te doen.

Data filtering verwijdert ongewenste data uit de set. Het kan dus bijdragen aan het samenstellen van een eerlijke dataset. Het is wel van belang om helder af te wegen welke data wordt verwijderd en waarom.

Een voorbeeld dat werd benoemd is het verwijderen van antisemitische teksten om discriminatie van Joden in het taalmodel te voorkomen. Hoewel dit een weloverwogen

keuze is met een duidelijke reden, leidt het mogelijk ook tot verlies van historische data en kennis. Het taalmodel moet wel in staat zijn uitleg te geven over de Holocaust. Data-filtering kan dus leiden tot het verwijderen van relevante inhoud.

Het is dus belangrijk om weloverwogen keuzes te maken over het filteren van data en vervolgens transparant te rapporteren welke data is verwijderd en waarom. Tenslotte merkte de groep op dat biasmitigatie door filteren niet voor alle soorten data en dataproviders nodig zal zijn. Data van cultural heritage organisaties over de Holocaust zou bijvoorbeeld buiten het filter tegen antisemitisme gehouden kunnen worden.

Tenslotte werd bij de discussie over biasmitigatie wederom gepleit voor transparantie en publieke rapportage over de gemaakte mitigatiekeuzes. Ook werd aanbevolen alle data te bewaren, ook als het wordt gefilterd uit de uiteindelijke dataset. Dit biedt de mogelijkheid de alle data en de gemaakte keuzes te analyseren, en eventuele fouten te corrigeren.

Bevindingen uit de groep

De door ons voorgestelde methoden voor het mitigeren van bias in de training data werden door de groep gezien als logische methoden. De discussies over de nadelen van deze methoden lieten echter zien dat er veel onduidelijk is over het effect van dergelijke keuzes op het gedrag van het model. Bij augmentatie bestaat het risico dat er niet-bestaande ideeën of concepten worden geïntroduceerd en bij filtering juist dat er kennis buiten het model wordt gehouden. Het al dan niet optreden van dit bijeffect kan niet van tevoren worden voorspeld.

Vanwege de complexiteit van de keuzes en gevolgen, en het gegeven dat er niet kan worden afgekeken van andere onderzoeken, benadrukte de groep het belang van transparantie. Ze raden aan publiek te rapporteren over de gemaakte mitigatiekeuzes en om de gefilterde datasets te bewaren. De inzichten die in dit project worden opgedaan kunnen bovendien bijdragen aan het ontwerpen van best practices.

Concrete aanbevelingen uit de groep zijn:

- Wees transparant over de gemaakte keuzes
- Bewaar alle data, ook de verwijderde of weggefilterde data
- Maak keuzes over het eindproduct van het taalmodel.
- Draag met de inzichten bij aan best practices

Discussievraag 5: Wanneer is het goed genoeg?

Introductie van de vraag

Omgaan met bias en representatie bij het ontwikkelen van een groot taalmodel, is erg complex. Er moeten keuzes gemaakt worden, ook al zijn de gevolgen niet helder, mist er wetenschappelijke kennis, of is het resultaat niet zoals gewenst. Toch is het van belang dat we iets doen en ergens starten. Daarom spelen de volgende vragen op:

- Wat is een redelijke inschatting van wanneer onze activiteiten op dit vlak voldoende zijn om verder te kunnen?
- Wat moeten we rapporteren om te laten zien dat we voldoende gedaan hebben?

Discussie

In de discussie over wat goed genoeg betekent, kwam een viertal perspectieven aan bod. Ten eerste werd besproken dat het taalmodel goed genoeg is als het inzichtelijk is waarvoor en in welke context het gebruikt kan worden. Gebruikers moeten dus op de hoogte zijn van de kwaliteit en de sterke en zwakke punten van het model. Een tweede kijk op de kwestie is dat het model goed genoeg is als het beter is dan bestaande modellen als chat-GPT. Het derde perspectief is dat het model goed genoeg is als het bruikbaar is voor gebruikers. Het model moet dus goed genoeg aansluiten op de gebruiker, maar kan toch vooroordelen bevatten. Tenslotte werd benoemd dat *goed genoeg* niet alleen gaat om het omgaan met bias, maar ook om andere eigenschappen van het model en het proces, zoals het waarborgen van auteursrechten of digitale soevereiniteit.

De groep was het erover eens dat een perfect, biasvrij taalmodel niet bestaat. Uit die conclusie volgen verschillende acties, zoals het inzetten op bewustwording. Voor veel mensen is het niet bekend dat bias in een taalmodel niet kan worden uitgesloten. Daarom is bewustwording en geletterdheid van GPT-NL gebruikers is dus van belang. De groep denkt daarbij aan het toevoegen van expliciete disclaimers aan de interface van het model waarbij elke GPT-NL gebruiker bij gebruik van het model wordt geïnformeerd. Daarnaast creëert het ontwerp van de uiteindelijke gebruikersinterface ook verwachtingen van wat een product, dienst, of service wel of niet kan. Daarom is ook de manier waarop het uiteindelijke eindresultaat wordt aangeboden een belangrijke ontwerpkeuze.

Ook kwam een gebruikershandleiding weer ter sprake. Het is belangrijk dat mensen begrijpen waar het model voor gebruikt kan worden en op wat voor data het getraind is. Deze handleiding kan sturing geven voor het verantwoordelijk gebruik van GPT-NL en potentiële gevaren binnen verschillende contexten vormgeven.

Tenslotte kwam wederom het belang van transparantie naar boven. De groep raadt aan niet te focussen op perfectie, maar op transparantie. Het inzichtelijk maken van het

proces biedt ‘best practices’ voor vervolgwerk. Hierbij is het ook waardevol om inzicht te bieden van de dataset waar het model op getraind is, zonder privacy te schenden. Zo kan er inzicht gegeven worden over metadata zoals het aantal woorden van een tekst of de grootte van het document.

Bevindingen uit de groep

- Maak duidelijk voor welke toepassingen het model geschikt is
- Verzorg bewustwording en geletterdheid van GPT-NL gebruikers over bias in taalmodellen in de gebruikersinterface van het taalmodel.
- Maak een gebruikershandleiding voor het toepassen van het taalmodel, waarin duidelijk wordt in welke context het model het best toegepast kan worden en op wat voor soort data het model getraind is.
- Laat zien hoe het model zich vergelijkt met alternatieven door middel van benchmarks en begin op tijd met evaluatie van het GPT-NL model en GPT-NL project.

Aanbevelingen voor het omgaan met representatiebias bij LLM ontwikkeling

Op basis van de expert-sessies hebben we overzicht met aanbevelingen verzameld. Deze is niet uitputtend maar kan behulpzaam zijn voor de ontwikkeling van taalmodellen en het gesprek over inclusievere AI.

Aanbevelingen uit de expert-sessies:

1) *Prioriteer transparantie*
   a) Maak expliciete keuzes, documenteer deze en deel ze publiek.
   b) Laat zien hoe het model zich vergelijkt met alternatieven.
   c) Bewaar alle data: ook verwijderde of gefilterde data zodat men hier onderzoek naar kan doen.
2) *Volledige representatie is niet mogelijk; wees voorzichtig met 'oplossingen' voor representatiebias*
   a) Focus op representatie van het huidige Nederland en laat een toekomstig of gewenst Nederland buiten beschouwing.
   b) Werken vanuit lijsten en categorieën brengen risico's mee rondom ongewenste categorisering, buitensluiten van groepen, of het niet voldoende rekening houden met intersectionaliteit.
   c) Start met traditionele NLP-technieken voordat je grootschalig aan de slag gaat met meer experimentele methodes.
3) *Geef ook aandacht aan het gebruik en de gebruiker*
   a) Maak duidelijk voor welke toepassingen het model bedoeld is.
   b) Verzorg bewustwording en geletterdheid van GPT-NL-gebruikers over bias in de gebruikersinterface.
   c) Maak een gebruikershandleiding voor het toepassen van het taalmodel, waarin duidelijk wordt in welke context het model het best toegepast kan worden en op wat voor soort data het model getraind is.
   d) Redeneer vanuit toekomstige gebruikerscontext over de mogelijk schadelijke gevolgen van keuzes in het model. Op basis daarvan kan je terugredeneren naar keuzes over datacuratie.
   e) Maak duidelijk voor welke toepassingen het taalmodel geschikt is.
4) *Werk samen met bestaande initiatieven*
   a) Zoek (verdere) samenwerking op met al bestaande initiatieven op het gebied van (representatie-)bias in taalmodellen.
   b) Neem contact op met mensenrechtenorganisaties en betrek ze in het opstellen van lijsten van gemarginaliseerde groepen, en het verzamelen van extra data'
   c) Draag met de inzichten bij aan het opstellen best practices.
5) *Kijk verder dan de data-acquisitie en -curatiefase*

a) Zorg dat diversiteit een prioriteit vanaf het begin van de ontwikkeling, al in de data-acquisitiefase. Niet alles is in een latere fase te corrigeren.
b) Kijk ook naar latere fases in de ontwikkeling om op (representatie)bias te mitigeren.
c) Voer experimenten uit met data-augmentatie en finetuning om op representatie te optimaliseren.

Gevolgen voor GPT-NL

De expertsessies hebben geholpen het gesprek te voeren over representatiebias en inclusiviteit in AI-systemen. De gesprekken hebben ons veel inzichten gegeven over representatiebias en AI, echter blijft het een thema waar verder onderzoek naar moet worden gedaan. Niet alle aanbevelingen kunnen al worden doorvertaald naar concrete stappen. In deze afsluiting van de verslaglegging van de expertsessies, willen we reflecteren op en verwijzen naar de dingen die we doen om GPT-NL zo divers mogelijk te maken.

Als het gaat om representatie, geldt de vuistregel: hoe diverser de data, hoe beter. Dit is ook meerdere malen in de expertsessies naar voren gekomen. De intentie om tijdens de dataverzameling (januari 2024 tot 1 mei 2025) extra aandacht te schenken aan ondergerepresenteerde groepen was zeker aanwezig. Zo zijn er contacten gelegd met verschillende organisaties om data te verkrijgen van en over minderheidsgroepen, maar was het niet mogelijk om deze data te ontsluiten of was er simpelweg niet genoeg beschikbaar.[19] Het is in deze eerste ontwikkelfase moeilijk gebleken om te GPT-NL te optimaliseren op inclusiviteit. Een andere reden hiervoor is dat we *van nul af aan* met de dataset zijn begonnen. Daardoor wisten we bij de start van de ontwikkeling nog niet welke datasets we zouden ontvangen, en welke groepen er dus in mindere mate gerepresenteerd zouden worden. Voor een vervolg op GPT-NL, kunnen we gerichter kijken welke groepen minder gerepresenteerd zijn in onze huidige dataset en op basis daarvan prioriteiten stellen.

Het mitigeren van representatiebias kan, zoals ook besproken in de expertsessies, ook plaatsvinden in andere ontwikkelfases dan alleen de creatie van de dataset. Er is gekeken of er bij de finetuningsfase extra aandacht kan worden aan diversiteit en representatie. Er is een survey uitgezet om diverse input op te halen voor de instructiedataset. We merkten hierbij dat het moeilijk is om vanuit onze organisatie een diverse groep te bereiken. Bij een volgende activiteit, zullen we het bereiken van een diverse doelgroep dan ook anders moeten insteken.

Ook in de laatste fase van de ontwikkeling, de evaluatiefase (ook wel ‘benchmarking’), kan er aandacht besteed worden aan bias en representativiteit. Binnen het GPT-NL project gaan we aan de slag met het ontwikkelen van een benchmark voor bias en gaan we het model ook met stakeholders evalueren. Op die manier kunnen we de representatie van GPT-NL evalueren en de rapportage hierover publiceren. Het doel is om ook in deze fase relevante stakeholders te betrekken.

Tot slot, in de expertsessies is het belang van transparantie over de technologie en keuzes die zijn gemaakt tijdens de ontwikkeling is sterk naar voren gekomen. Transparantie is één van de kernwaarden van GPT-NL, en we willen hier dan ook zoveel mogelijk gehoor aan geven.

---

[19] Veel organisaties die wij benaderden beheerden niet zelf de auteursrechten voor de publicaties, deze waren gefragmenteerd over verschillende (soms niet meer bestaande) organisaties. Daarnaast betreft het vaak kleine datasets. Voor de ontwikkeling van GPT-NL 1.0 maakten we gebruikt van een drempel van 10 miljoen woorden; gezien de doelstelling van een minimum van 300 miljard teksttokens voor de eerste ontwikkeling was het praktisch niet mogelijk om kleinere datasets te includeren.

- Aan het eind van de training van GPT-NL zullen we datasheets publiceren over de verschillende datasets, evenals de keuzes die zijn gemaakt om tot deze datasets te komen. We hopen dat deze documentatie dient als een gespreksstarter over representatie en onderzoekers helpt in het doen van onderzoek.
- We publiceren aan het eind van het project de Decision Sheets die gemaakt zijn tijdens het project. Hierop hebben we de ontwerpkeuzes van GPT-NL geregistreerd: zowel die van de dataset, zoals hierboven benoemd, als in andere ontwikkelfases.
- We gaan in gesprek met onze dataproviders over de mogelijkheden om onderzoekers toegang te geven tot de datasets[20] en het model voor onderzoeksdoeleinden.

Uiteindelijk zal ook GPT-NL een 'biased' taalmodel opleveren. In het project hebben we binnen onze handelingsruimte geprobeerd bewuste keuzes te maken. We hopen dat de publicatie van deze inzichten uit onze expertsessies helpen het gesprek over representatiebias te voeren. Zo hopen we dat er meer aandacht ontstaat voor biasmitigatie, vooral op dataniveau bij taalmodellen.

[20] Delen van de GPT-NL dataset omvat auteursrechtelijk materiaal, hierdoor is het niet vanzelfsprekend om de dataset breed te delen.

Appendix: Recommendations (ENG)

Based on the expert sessions, we have compiled an overview of recommendations. This is not exhaustive but may be helpful for the development of language models and the discussion on more inclusive AI.

Recommendations from the expert sessions:

1) *Prioritize transparency*
    a) Make explicit choices, document them, and share them publicly.
    b) Show how the model compares to alternatives.
    c) Store all data, including deleted or filtered data, so that it can be researched.
2) *Full representation is not possible; be cautious with "solutions" for representation bias*
    a) Focus on representing the Netherlands as it is today and avoid projecting a future or desired Netherlands.
    b) Working from lists and categories carries risks of undesirable categorization, exclusion of groups, or insufficient consideration of intersectionality.
    c) Start with traditional NLP techniques before embarking on large-scale use of more experimental methods.
3) *Also pay attention to the use and the user*
    a) Clearly define the intended applications of the model.
    b) Raise awareness and literacy among GPT-NL users about bias in the user interface.
    c) Create a user manual for applying the language model, clearly stating the context in which the model can best be applied and the type of data on which the model has been trained.
    d) Consider the potentially harmful consequences of choices made in the model from the perspective of future users. Based on this, work backwards to make informed decisions about data curation.
    e) Make clear for what end products the language model can be used.
4) *Collaborate with existing initiatives*
    a) Seek collaboration with existing initiatives in the field of representation bias in language models.

b) Contact human rights organizations and involve them in drawing up lists of marginalized groups and collecting additional data.

c) Contribute your insights to the development of best practices.

5) *Look beyond the data acquisition and curation phase*

a) Ensure that diversity is a priority from the start of development, i.e., the data acquisition phase. Not everything can be corrected at a later stage.

b) Also consider later stages of development to mitigate representational bias.

c) Conduct experiments with data augmentation and fine-tuning to optimize representation.

Colofon

Dit is een publicatie vanuit het GPT-NL project. Meer informatie is te vinden op www.gpt-nl.nl en vragen zijn welkom op info@gpt-nl.nl.

Auteurs (op alfabetische volgorde):

- Duuk Baten, SURF
- Dominique Blok, TNO
- Lieke Dom, TNO
- Eliza Hobo, TNO
- Merel van Steenis, SURF

Deelnemende experts:

- Sebastiaan Berendsen, Deloitte
- Katrien Depuydt, Instituut voor de Nederlandse Taal
- Maaike Harbers, Hogeschool Rotterdam
- Bo Hijstek, Rathenau Instituut
- Dong Nguyen, Universiteit Utrecht
- Sander van der Waal, Waag Futurelab
- Oskar van der Wal, Universiteit van Amsterdam

Experts hebben deelgenomen op basis van hun persoonlijke expertise: opvattingen en meningen zijn uitsluitend van henzelf en niet uit naam van hun werkgever(s).