Commitment Review 2025

Introductie

Begin van het project hebben we vanuit dit perspectief onze commitments opgesteld. Deze helpen om de ambities van GPT-NL te verduidelijken en om ervoor zorgen dat onze (publieke) belanghebbenden weten wat ze van ons kunnen verwachten. We doen verschillende commitments ten behoeve van (1) het ontwikkelproces, (2) het eindproduct, (3) transparantie, (4) het gebruik van data, (5) diversiteit en inclusie, en (6) communicatie en het betrekken van belanghebbenden.

Innoveren betekent echter ook dat we ons op onbekend terrein bevinden. Met name als het gaat om (generatieve) AI, kan wet- en regelgeving nog ontbreken en zijn best practices nog beperkt, maar moeten we toch verantwoorde keuzes maken over de ontwikkeling van het model en het proces van het project. De commitments zijn dan ook niet in beton gegoten. Na een jaar GPT-NL weten we nog beter waar we staan en wat er wél en wat er níet mogelijk is. Daarom hebben we na één jaar GPT-NL een review gedaan op onze commitments: zo kunnen we gestructureerd weergeven welke doelen we hebben behaald, en waar we commitments hebben moeten aanpassen omdat de kaders van wet- veren regelgeving ons helaas niet anders toelaten.

Hieronder reflecteren we op onze commitments. Wat we veranderen hebben we in de tekst ~~doorgestreept~~, en wat we toevoegen staat er in het groen. De nieuwe formuleringen van de commitments zijn aangepast op onze website.

Reflectie

1. Het proces van het project

Innoveren is een dynamisch proces, dat is ook maar weer gebleken bij GPT-NL. De workflow voor het bouwen van de dataset is dan ook niet zo lineair als we van tevoren hadden ingeschat, en daarom konden we niet aan de start van de dataverzameling aangeven hoe deze procedure eruit zag. We hebben veel geleerd bij onze manier van dataverzameling. Daarom zullen we bij oplevering de dataset niet alleen verklaren hoe we keuzes hebben gemaakt welke data we wel en niet hebben meegenomen, maar ook welke stappen we hebben genomen om de data te vergaren.

1.1 Publishing this commitments document. We will be reviewing this commitment list on a regular basis to incorporate feedback and publicly report on any changes to the commitments.

1.2 Publishing a document that describes ~~the workflow of decision-making in building~~ how we have built our dataset~~s~~. We will publish this when the dataset is finished.

2. De eindproducten van het project GPT-NL

Het project GPT-NL heeft meerdere doelen dan alleen de ontwikkeling van het taalmodel. In het eerste jaar hebben we daar de juiste invulling aan kunnen geven. Binnen de omvang van het project GPT-NL, leveren we eind 2025:

1. Een blauwdruk voor de ethische en verantwoorde ontwikkeling van LLMs,
2. Een onderzoeksfaciliteit,
3. Een eerste versie van een taalmodel genaamd GPT-NL, met een performance vergelijkbaar met Llama 2 7B and GPT-3 175B, verder uit te splitsen in
    a. De trainingsdataset;
    b. De broncodes;
    c. De modelgewichten.

Vanaf de start zijn we voornemens een succesdefinitie voor het project GPT-NL op te stellen omdat we geloven dat dit ons stuurt in het maken van de juiste beslissingen. Het project GPT-NL is een succes als we laten zien dat we bovenstaande op een verantwoorde manier kunnen ontwikkelen. Het

definiëren van succescriteria die bestaan uit technische eisen of hoeveelheden is niet voldoende, het gaat juist ook om het kunnen laten zien dat we dat op een verantwoorde manier hebben gedaan. We blijven gecommitteerd aan het publiceren van deze 'Definition of Succes' rapportage waarin we inzichtelijk maken hoe we invulling geven aan 'succes' in technische zin, en aan 'succes' vanuit het perspectief van Verantwoorde AI.
Verder hadden we ons in eerste instantie gecommitteerd tot het publiceren van *permissive licensing models* voor deze eindproducten. We hebben echter geleerd dat omwille van wet- en regelgeving en subsidievoorwaarden het niet mogelijk is om al onze eindproducten onder een open-source licentie vrij te geven. Daarnaast hebben we de juiste terminologie voor hoe open GPT-NL wordt nog niet gevonden, maar wel geleerd dat *permissive licenses* niet het type licentie is die mogelijk is voor GPT-NL. Hier lees je meer over hoe open en toegankelijk GPT-NL zal worden.

2.1 Publishing a definition of success for the GPT-NL project (describing when we see the project as a success). This should be published no later than the end of the data collection-milestone.
2.2 Publishing an overview of the end-products intended by the project; including a definition of success for each end-product, a description of intended goal, openness, and licensing. Our intention is that the end-product will be as open as possible, but as this is dependent on agreements with data providers, we cannot guarantee this yet. ~~The overview should be published no later than the end of the data collection-milestone.~~
2.3 Provide with each end-product a clearly described ~~and permissive licensing model~~ license.

3. Transparantie en toegankelijkheid

Zoals we in het begin stelden, zijn we gecommitteerd aan het publiceren van alle trainingsdata die is gebruikt om GPT-NL mee te trainen, al wisten we ook dat we beperkt zouden zijn in het publiceren van de data die auteursrechtelijk beschermd is: het is uiteindelijk aan de auteursrechthebbenden om te bepalen wat er met hun data gebeurt. We begrijpen echter het belang van transparantie met name voor onderzoekers. We willen daarom in de toekomst samen met auteursrechthebbenden onderzoeken of het mogelijk is om deze doelgroep een manier van beveiligde toegang te verlenen voor een bepaalde termijn (commitment 3.4).

Tot slot hebben we ook geleerd dat we door wet- en regelgevingen en subsidievoorwaarden beperkt zijn in de mate waarin we toegang kunnen geven tot GPT-NL. Met andere woorden, het model wordt niet zo 'open source' als we in het begin hadden gedacht of gehoopt. Hoe dat precies zit, leggen we verder uit in 'Hoe open wordt GPT-NL'. In het kort komt het erop neer dat we de modelgewichten beschikbaar stellen op aanvraag (en niet publiekelijk publiceren). Dit gaat wel gepaard met een nieuwe commitment, namelijk dat we licenties creëren zodat we aan de behoeftes van elke belanghebbende van GPT-NL voldoen.

3.1 Publishing all code publicly under an open-source license.
3.2 Publishing datasheets and model-cards for all datasets and models ~~(end-products)~~ according to industry best practices.
3.3 The ambition to release and publish the datasets used to train GPT-NL by default. However, some datasets might be ~~licensed~~ protected by copyright law and thereby limiting us in publication. For those datasets we will give explicit attention to creating other mechanisms of transparency.
3.4 Find out, in cooperation with the Content Board, if there are ways to give researchers and/or auditors secured access to the training dataset of GPT-NL.
3.5 Making GPT-NL as accessible as possible, meaning that
1) we try to give free access to researchers and/or auditors, and;
2) create licenses that suit all stakeholders of GPT-NL

4. Het gebruik van data

Met betrekking tot het gebruik van de training data zitten we op de goede weg. De uitdagingen van de dataverzameling hebben ertoe geleid dat we voor nieuwe dilemma's kwamen te staan,

bijvoorbeeld over de keuze welk type synthetische data we wel en niet zouden meenemen. Hoewel de dataverzameling een grote klus is geweest, hebben we vastgehouden aan onze ambitie om alleen data te gebruiken waarvoor we de juiste licentierechten hebben.

4.1 We will only use content for training GPT-NL if the data provider has the appropriate rights to provide a license to us to do so. This means that the data provider needs to be either the owner of any copyright or database rights ~~in your dataset~~ or has valid license rights granted to it by the third party owner.
4.2 We do not train GPT-NL on any information subject to regulatory or contractual confidentiality requirements (such as info under patient confidentiality, business confidential data).
4.3 We have dedicated focus on detecting, filtering, and removing personal information from the training data. We support data providers with removing personal information from the dataset.
4.4 We have dedicated focus on detecting, filtering and removing harmful content - such as, violent or criminal content, discriminatory content or hate speech– from our training data.

5. Diversiteit en inclusie

Met betrekking tot diversiteit en inclusie willen we eerlijk toegeven dat we de commitments (nog) niet hebben behaald zoals we een jaar geleden hadden gehoopt. We lopen beide commitments even langs:

Het mitigeren van bias in het model door het creëren van een dataset die zoveel mogelijk groepen representeert.
We ondernemen verschillende stappen voor het mitigeren van bias. Om kennis op te halen hebben we in september 2024 twee expert sessies georganiseerd over 'representatiebias'. Hier hebben we onder andere besproken wat een 'goede' representatie van Nederland is, welke methodes er zijn om representatiebias in een dataset te meten, en welk resultaat van het project 'goed genoeg' is als het gaat om representatie. Het volledige verslag van de expert sessies kan je hier vinden.

Echter zit de uitdaging van deze commitment in het stuk 'creëren van een dataset die zoveel mogelijk groepen representeert'. Dit heeft te maken met de uitdagingen die we zijn tegengekomen bij de dataverzameling. De dataverzameling is een tijdrovend proces geweest wat we hebben onderschat, en dit is ten koste is gegaan van het betrekken van zoveel mogelijk groepen om diverse data op te halen. We zijn ons ervan bewust dat tijd- en capaciteit een verklaring zijn, maar geen rechtvaardiging.

Verder, zo bleek ook uit de expert sessie, is het lastig om te stellen wanneer representatie 'goed genoeg' is. Een jaar geleden gaven we in deze commitment aan dat we bias willen mitigeren door zoveel mogelijk groepen mee te nemen in de dataset. Het is moeilijk om dit vooraf te kwantificeren, en tegelijkertijd is het achteraf lastig te beargumenteren waarom we wel of niet 'zoveel mogelijk' groepen hebben geïncludeerd. Op basis van de expert sessies, kiezen we ervoor de commitment aan te passen om het meetbaar te maken. Wanneer GPT-NL wordt getraind, kunnen we opnieuw reflecteren op hoeveel en hoe goed we deze groepen betrokken hebben.

Het betrekken van ondergepresenteerde groepen in de fine-tuning fase
We hebben een onderhandse tender uitgezet voor de fine-tuning van GPT-NL. Eén van de factoren waar we een annotatiepartij op selecteren is de diversiteit van de groep annotators. Hoewel we dit dus op de radar hebben staan, hebben we nog geen plan opgesteld waarmee we nu kunnen laten zien hoe we ondergepresenteerde groepen actief te betrekken. We kunnen daarom niet stellen dat we het tot nu toe op deze commitment zo goed hebben gedaan als we hadden gehoopt. Wel organiseren we een focus groep om van externe experts inzichten te krijgen voor het vergroten van diversiteit tijdens de fine-tuningsfase.

5.1 Mitigating bias in the model to the best of our ability by creating a diverse foundation dataset that represents as many groups as possible. We focus on minority groups in the categories below.
   a) Gender identification
   b) Age
   c) Ethnicity
   d) Religion
   e) Sexual orientation
   f) Disabilities
   g) Socio-economic status

5.2 Involving underrepresented groups to help us improve the model in the fine-tuning stage. We look at the same underrepresented groups as listed in 5.1.

6. Stakeholders en communicatie

Onze commitment tot communicatie en het betrekken van stakeholders is anders uitgepakt dan van tevoren verwacht. We hebben ons eerste progress report gepubliceerd en zijn gecommitteerd deze voortgangsupdates vol te houden, al publiceren we geen communicatieplan meer. Dit halen we daarom weg uit de commitment. Communicatie heeft namelijk een andere vorm gekregen dan vooraf gedacht.
Zo bleek bijvoorbeeld de vorm van een kwartaalupdate niet te passen bij de voortgang van het project maar geven we veel updates in andere vorm, zoals in presentaties, interviews en externe media. Ook zijn we gecommitteerd om GPT-NL begrijpelijk te maken voor een publiek voor wie (generatieve) AI nieuw is, bijvoorbeeld via externe media zoals NOSop3 en eigen artikelen. Dit hebben we extra toegevoegd aan onze commitments.

Met betrekking tot het publiceren over juridische en ethische dilemma's hebben we nog geen gestructureerde aanpak gevonden, al adresseren we deze uitdagingen in andere vormen van communicatie. Vooralsnog zijn we voornemens hier voor het einde van dit jaar meer over te delen in de vorm van een rapportage of reflectie.

Het betrekken van stakeholders is ook anders gelopen dat gedacht. Zo hebben we rondom de data contributeurs een zogeheten Content Board opgezet. We zijn voornemens eenzelfde soort board op te zetten voor gebruikers van GPT-NL om co-creatie te faciliteren (User Board). We hebben deze stakeholderbijeenkomsten toegevoegd aan onze commitments.
Echter hebben we niet zoveel publieke stakeholder sessies georganiseerd als gehoopt. We hebben twee stakeholdersessies gehad over het onderwerp representatiebias in de dataset (zie ook commitment 5.1). Echter hebben we deze in ons netwerk doorgestuurd, in plaats van op onze website aangekondigd zoals onze commitment stelt. Hier hebben we ons dus niet goed aan gehouden.

6.1 ~~Publishing a communication plan and communicating towards the public.~~ We proactively communicate and regularly share our progress with the general public. :
   a) with explanatory articles
   b) via external media
   c) with progress reports ~~on a quarterly basis (every three months).~~

6.2 Publishing regular (public) reports on decisions made within the project. Including reporting on legal and ethical dilemmas and decisions.

6.3 Reporting on our conclusions of stakeholder consultations (below).

6.4 We will organise public stakeholder consultations for at least the following:

6.4a Involvement in preparation of the fine-tuning stage.

6.4b Consultation on methods of evaluating model performance (on technical and societal benchmarks).

~~6.5~~ Stakeholder consultations will be announced publicly on the GPT-NL website and social media

channels.

6.5 We organise influence of stakeholders that are data providers (Content Board) and/or users (User Board) to facilitate co-creation of GPT-NL.

Tot slot

Bij de start van GPT-NL hebben we gesteld dat we gecommitteerd het bouwen van een betrouwbaar model dat in lijn is met de Ethics Guidelines for Trustworthy AI van de EU. In deze guideline staat dat betrouwbare AI-systemen rechtmatig, ethisch en robuust moeten zijn. Om een betrouwbaar taalmodel te bouwen, hebben we de zes thema's van onze commitments geformuleerd. In 2024 hebben we in een paper uitgelegd hoe onze commitments zich tot de Ethics Guidelines verhouden. Dit hebben we gebundeld en gepresenteerd op een conferentie over 'AI for the public'. De presentatie is hier te bekijken, de publicatie komt dit jaar online.