GPT-NL

Het eerste jaar GPT-NL

Progress report #1

Voorwoord

Selmar Smit

Het eerste jaar van GPT-NL is voorbij. We hebben misschien nog niet alle doelen bereikt die we onszelf vorig jaar rond deze tijd hadden gesteld, maar zoals Chris Denbigh-White zo mooi wist te verwoorden:

“There is no reason AI can’t be compliant with GDPR, but companies need to take the time to get it right.. Organisations need to prioritise legality over speed. After all, the backlash over a legal issue is much more significant than that of the potential complaints over the timeline”

En dat hebben we dan ook gedaan. Als wij een verschil moeten maken, dan gaan we dat ook niet redden met meer data of meer computerkracht, en zelfs op technische kennis kunnen we ons maar beperkt onderscheiden van de gigantische innovatiekracht van BigTech. We hebben dan ook een andere route gekozen, een route die ons naast soevereiniteit en onafhankelijkheid, ook een verantwoorde totstandkoming brengt als ‘unique selling point’. Daar zie je dan ook onze kracht op volle toeren; het verbinden van technische kennis met domeinkennis, ethiek, legal en business development. Is dat de slotgracht die GPT-NL gaat beschermen tegen aanvallen van buiten? Dat gaat de toekomst uitwijzen, maar als je ziet wat je daarvoor nodig hebt – en dat beschrijft dit rapport uitvoerig – dan laat het wel zien dat het “goed doen” veel complexer is dan gewoon het halve internet door een neuraal netwerk trekken.

We bevinden ons nu op een kritisch moment. De komende jaren zal Big Tech waarschijnlijk verder uitlopen. Als niet snel duidelijk wordt dat de ontwikkeling van taalmodellen zich kan aanpassen aan regelgeving, zal de regelgeving zich aanpassen aan de ontwikkeling. En hoeveel ruimte is er dan nog voor ons, als klein kikkerlandje?

Inhoudsopgave

“Het kan ook best goed zijn te leren rijden in een Fiat in plaats van in een Ferrari

Waar staan we met GPT-NL?

Product Owner Saskia Lensink maakt de balans op.

Nu het najaar inmiddels volop z'n intrede heeft gedaan, is het tijd om de balans op te maken. Waar staat GPT-NL nu? Wat zijn belangrijke mijlpalen geweest en wat bleek een grote uitdaging? Een aantal vragen aan Saskia Lensink, als *product owner* één van de kartrekkers van GPT-NL. Saskia is een bevlogen en gemotiveerde *data scientist* bij TNO, die ook eerlijk toegeeft dat de complexiteit van het project iets is wat ze niet volledig had voorzien.

Saskia, eerst nog even terug naar het hoe en waarom. En dan vooral: waarom zijn TNO, SURF en het NFI gestart met de ontwikkeling van GPT-NL?
'De ontwikkeling van dit Nederlandse taalmodel is in de basis ontstaan vanuit een ongemak in Nederland en Europa rondom een gebrek aan digitale onafhankelijkheid en soevereiniteit. Oftewel: we zijn te afhankelijk van niet-Europese krachten als het gaat om bijvoorbeeld onze energievoorziening, data of cloud computing. Ook is er weerstand over de manier waarop grote techreuzen de dingen aanpakken. Dus zo spoedig mogelijk innoveren en gaandeweg maar ontdekken welke wettelijke en ethische kaders er zijn. Die kaders zijn in de VS toch anders dan bij ons, en in China al helemaal. Dan speelt er ook nog een onderliggend gevoel: we willen als Europa mee kunnen doen op het wereldtoneel, of in ieder geval niet nog verder achterop raken. We willen zelf generatieve AI kunnen ontwikkelen die past bij onze normen, waarden en behoeften. En als het even kan ook de infrastructuur ontwikkelen die daarbij hoort. Dit gezegd hebbende, loop je natuurlijk wel meteen al tegen grote uitdagingen.'

Noem is zo'n groot uitdaging?
'De grootste is misschien wel de wet- en regelgeving. Daarvan weten we nog niet zo goed hoe dat rijmt met de ontwikkeling van generatieve AI, bijvoorbeeld rondom copyrights of privacy. Zo hebben we te maken met de Europese AI Act, een wet die nog heel nieuw is. Terwijl die AI Act in de maak was, kwam generatieve AI net opzetten. Dan is het zaak om iets wat razendsnel gaat, wettelijk in te kaderen. Je raadt het al: de grote vraag is hoe je dat doet. Er is nog geen jurisprudentie. Als iets onduidelijk is, nemen we in Nederland vaak de voorzichtige afslag en doen we het nog niet, vooral als het over privacy gaat. Een goede zaak, zou je kunnen betogen. Het maakt tegelijkertijd wel dat innovatie daardoor langzaam kan gaan.'

Voordat we naar het positieve toekomstbeeld gaan, zijn er nog andere zaken waar je mee worstelt?
'Synthetische data is ook een goed voorbeeld. Gelukkig zijn er bestaande datasets die voldoen aan onze normen en die we

rechtmatig kunnen verkrijgen. Maar je kan ook bestaande data gebruiken om nieuwe data te ontwikkelen, oftewel synthetiseren. Dat klinkt goed, maar is best ingewikkeld. Je kan dat synthetiseren op verschillende manieren doen. Je kan bijvoorbeeld machinevertalingen inzetten, maar ook gebruikmaken van bestaande LLMs om nieuwe data mee te maken. Dan staan we voor de keuze: laten we een deel van ons productieproces uitvoeren door een 'niet zo schoon product', ofwel een AI-model waar we vraagtekens bij hebben. En zo ja, waar trek je dan de grens? Het is een voorbeeld van hoe we tegen de complexiteit van de realiteit aanlopen. Verder zijn we met 70 partijen in gesprek over hun data, maar ze hebben vaak tegenstrijdige belangen. Ook dat is uitdagend.'

Wat heeft je in dit hele proces het meest verrast?
'Vooral de zoektocht naar de balans tussen ethiek, techniek en einddoel. Kijk, je kan er voor kiezen dingen zo netjes mogelijk te doen. Dat kan dan betekenen dat je een zo schoon mogelijk model maakt. Maar als een model vervolgens niet bruikbaar zou zijn in de praktijk, dan is dat ook onethisch. Alle middelen die dan gebruikt zijn, alle CO2-uitstoot... Dat wil ik ook naar mezelf kunnen verantwoorden. Het heeft me ook verrast hoe complex het kan zijn iets te bouwen waar bij wijze van spreken heel Nederland iets aan heeft. Dat is echt iets anders dan een theoretisch, academisch product.'

Wat heb je geleerd van deze realiteit?
'Dat heldere communicatie cruciaal is. Iedereen wil betrokken zijn en op hoogte gehouden worden, maar je kan simpelweg niet alle mogelijke eindgebruikers, projectleden, maatschappelijke organisaties en tal van anderen tevreden houden. Er is niet één goede keuze. Er zullen verwachtingen zijn waar we niet aan voldoen. Zo transparant mogelijk zijn en open communiceren is dan enorm belangrijk.'

Als al deze hordes zijn genomen, wat hoop je dan dat over een jaar gerealiseerd is?
'Ik hoop dat we er dan in geslaagd zijn een schoon taalmodel te ontwikkelen binnen de beperkingen die we nu kennen. En ik hoop dat we daar een vruchtbaar ecosysteem voor hebben kunnen bouwen. We zijn momenteel met veel partijen in gesprek. Wat verwacht je, wat wil je. Daar rollen allerlei connecties en protocollen uit. Die kunnen gaan over hoe je data netjes verwerkt, welke contracten je met elkaar sluit en hoe je omgaat met ethische vraagstukken. Daar maken we een soort blauwdruk van, een onderzoeksfaciliteit over taalmodellen. Dat zou uiteindelijk supergaaf zijn en ik heb goede hoop dat het lukt.'

Hoe ver zijn we richting die mijlpaal, waar staat GPT-NL nu?
'Ons team heeft ontzettend veel technische hobbels genomen en heeft veel ontwikkeld en getest. We hebben met elkaar veel en lang nagedacht over hoe we zo goed mogelijk kunnen voldoen aan de wettelijke kaders. We hebben daar concrete stappen in gezet, zoals het vertalen naar protocollen. We blijven elkaar uitdagen om de ethische aspecten te blijven overwegen en we zetten responsible AI framworks in. We zijn met veel verschillende partijen in gesprek, zowel mogelijke eindgebruikers alsook mogelijke dataleveranciers. Dit alles bij elkaar is echt iets om heel trots op zijn.'

En zou je uiteindelijk ook niet gewoon willen zeggen: over een jaar moet er een bruikbaar alternatief voor de modellen van de grote techreuzen liggen?
‘Ja, zeker. Het is natuurlijk belangrijk dat er een schoon model komt dat kan samenvatten, vereenvoudigen en zoekmachines kan versterken. Tegelijkertijd moeten we goed beseffen dat dit een eerste iteratie is van iets wat nooit ‘af’ kan zijn. Het verder verzamelen van opt-in data en het verder verfijnen van het model wordt een continue proces.’

Tot slot, welke vragen krijg je het meest over GPT-NL?
‘Ik hoor wel eens geluiden als: ‘Denk je nou echt dat je iets kan doen met 13,5 miljoen euro? Dat gaat toch nooit werken?’ Dat gaat dan over de subsidies van het Rijk. Of mensen vragen natuurlijk wanneer GPT-NL inzetbaar is. Ik probeer dan uit te leggen dat het ook draait om kennisopbouw. En dat het ook best goed kan zijn te leren rijden in een Fiat in plaats van in een Ferrari. Ik vind die vragen niet erg. Dit project is ontzettend gaaf en ik geloof in een mooie uitkomst.’

Over Saskia Lensink
Saskia heeft een achtergrond in de taalwetenschap. Ze promoveerde in Leiden en werkt nu 4,5 jaar bij TNO. Als linguïst kijkt ze wat voor structuren er in taal zitten. Ze meet hoe het menselijk brein, oren en spraakapparaat werken. Daarbij onderzoekt ze ook hoe je dat in computermodellen kunt vatten (natural language processing en computational linguistics). Als product owner binnen GPT-NL brengt Saskia de verschillende subteams en de buitenwereld samen. Denk aan data acquisitie, systeemarchitectuur en datacuratie, maar ook de vereisten en belangen van externe stakeholders.

Milestones

We hebben misschien nog niet alle doelen bereikt die we onszelf vorig jaar rond deze tijd hadden gesteld; maar we hebben zeker niet stilgezeten! Met het team van GPT-NL hebben we dit jaar onontgonnen terrein verkend, zoals het beschermen van auteursrechten in een wereld van dataverzameling, en ons ingezet om digitale soevereiniteit te versterken.

We zijn trots op iedereen die helpt GPT-NL neer te zetten en zetten daarom een aantal milestones van 2024 op een rij.

1. Data Acquisitie
2. Data Curatie
3. Architecture & Framework
4. Auteursrecht en GPT-NL
5. Onze commitments en kernwaarden
6. Licenties GPT-NL Model

Data Acquisition

Het vergaren van de data is uiteraard de eerste stap voor de ontwikkeling van GPT-NL. Tegelijk is dit ook één van de meest intensieve fases in het proces, waarin strategische keuzes moeten worden gemaakt. Vanaf de start van GPT-NL zijn we in gesprek gegaan met potentiële data providers over het aanleveren van hun data voor GPT-NL. In de eerste maanden hebben we intensief uitgezocht hoeveel data er in Nederland beschikbaar is om GPT-NL op te trainen. Helaas moesten we een aantal maanden verder toch concluderen dat er simpelweg niet genoeg beschikbaar is. Daarom ontstond er, zo halverwege 2024, de noodzaak om een nieuwe strategie te formuleren. Inmiddels hebben we een ambitie voor een dataset gedefinieerd. Daarnaast hebben we een aanpak kunnen ontwikkelen om met verschillende data providers samen te werken.

Om een inschatting te maken van hoeveel data nodig is om een model vanaf nul te trainen, kijken we naar eerder ontwikkelde taalmodellen. Alle betere taalmodellen van de afgelopen jaren zitten tussen de 300 en 15.000 miljard tokens aan trainingsdata, dus 300 miljard teksttokens is zo ongeveer het minimum voor GPT-NL. Dat is een gigantische hoeveelheid tekst, ongeveer zo veel Nederlandse tekst als nu op het internet is te vinden. Als je uitgaat van 300 miljard tokens, dan zijn dat 3 miljoen Harry Potter boeken of een stapel papier van 10 km hoog - dat is nog hoger dan de Mount Everest.

Content Board

Die data moet ergens vandaan komen. Ons team rondom de dataverzameling zorgt voor een juiste strategie. Er zijn zo'n zeventig Nederlandse partijen geïnteresseerd in het delen van data voor GPT-NL, maar we merken dat de belangen en ambities van deze partijen uiteenlopen. Om ervoor te zorgen dat we alle belangen zoveel mogelijk dienen, en hierin transparant te werk gaan, hebben we een zogenaamde Content Board opgezet. De Content Board is een brede vertegenwoordiging van dataproviders die zelf data aanleveren voor GPT-NL. Er is veel interesse om zitting te nemen in de board. Vanuit deze coöperatieve wijze, laten we zien dat we dataverzameling bij GPT-NL anders aanpakken dan huidige LLM-aanbieders. Midden december heeft de eerste meeting plaatsgevonden tussen deelnemers van de Content Board. Dit betrof een groep data providers die vanaf het begin van GPT-NL nauw betrokken waren bij de ontwikkeling. Meer informatie over de Content Board volgt begin 2025.

Naast de private data, aangeleverd door de content board, moeten we de dataset voor GPT-NL met andere bronnen aanvullen. Hiervoor gebruiken we ook Engelstalige data, Germaanse data en code, verkregen via publiek beschikbare bronnen of synthetisch gegenereerd.

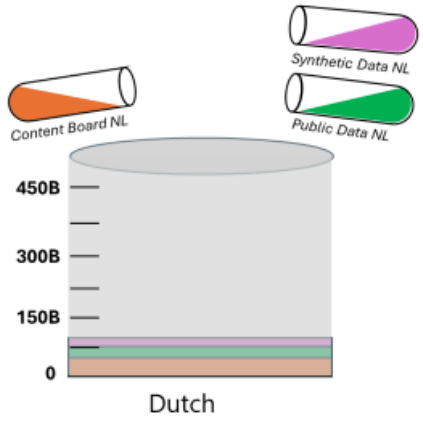
Data Acquisition

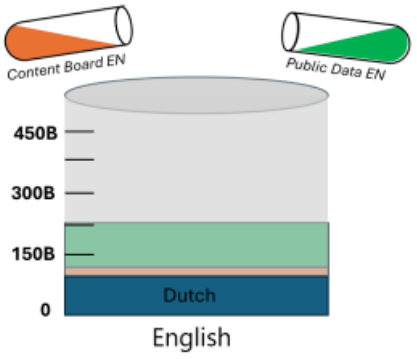
Private data
De dataproviders die zelf data aanleveren nemen deel in de Content Board.

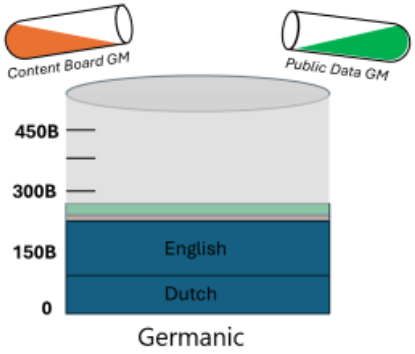
Publiek beschikbare data
Er wordt een copyright compliant extract gemaakt van CommonCrawl (CC-BY en CC-0 licenties) in samenwerking met het Instituut voor de Nederlandse Taal. Open data waar we expliciet toestemming voor hebben gekregen wordt ontsloten in samenwerking met Openstate.eu. De verwachte datum is 31 Maart 2025.

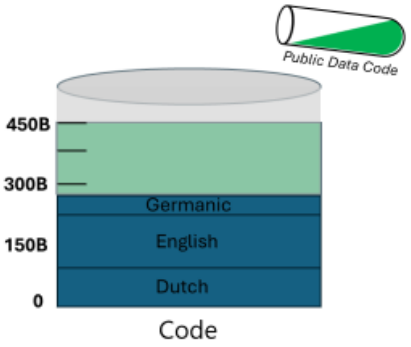
Synthetische data
We creëren synthetische data omdat er te weinig private, beschikbare data is om GPT-NL op te trainen. Daarnaast creëren we ook synthetische data om de kwaliteit van onze dataset te verhogen. We hebben samengewerkt met het Utrechts Archief (HUA) om hun scans (afbeeldingen) om te zetten in tekst. Daarnaast zijn we bezig met het vertalen van grote datasets. Het genereren van tekst uit gestructureerde data zoals tabellen en kennisgraven volgt in Q1 2025.

Code
Naast de 300 miljard teksttokens vullen we de dataset aan met 150 miljard tokens aan code. Het is aangetoond dat dit redenering van het model verbetert.









Data Curation

Waar het team van dataverzameling kijkt naar hoe de databronnen ontsloten kunnen worden, richt het datacuratieteam zicht op de uiteindelijke de dataset om GPT-NL mee te trainen. Allereerst zorgt het datacuratieteam dat de databronnen worden opgeschoond zodat ze samen een schone en kwalitatieve dataset kunnen vormen. De Data Curation pipeline bestaat uit zes stappen (hiernaast omschreven).

Momenteel zijn al deze modules geïmplementeerd en is het team bezig met het afmaken van de documentatie. Daarnaast moet nog een keuze worden gemaakt tussen onze eigen PII-module of de module van een extern bedrijf. We denken de code begin 2025 open source te kunnen maken.

Ook heeft het team onderzocht wat het betekent om een representatieve dataset te creëren. Hiervoor zijn in september 2024 twee sessies georganiseerd met externe experts om input op te halen over Representatie Bias. Het verslag wordt begin 2025 gepubliceerd.

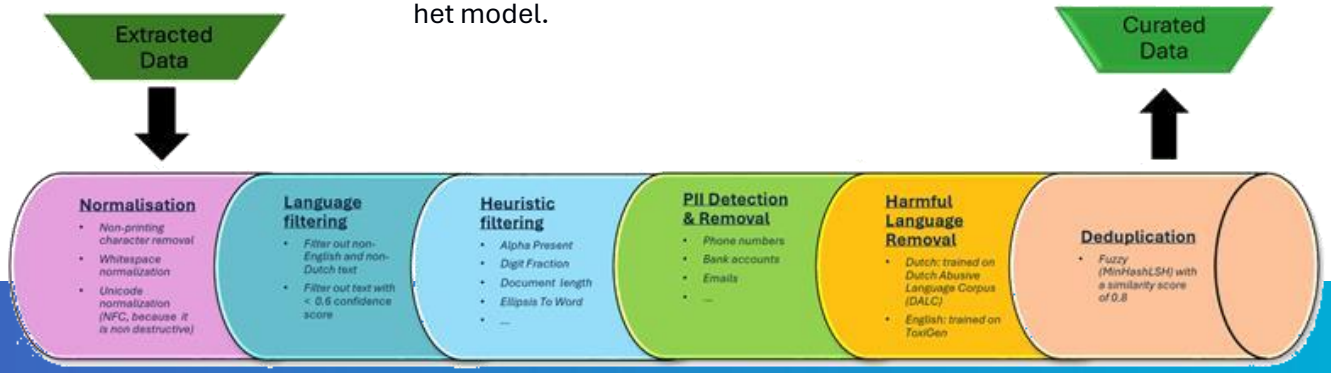
Toepassen van tekstnormalisatie en heuristische filters
Kwalitatief slechte data uit de dataset filteren, bijvoorbeeld data die bijna alleen bestaat uit cijfers en speciale tekens.

Taaldetectie
Gezien we zowel Engelstalige als Nederlandstalige data gebruiken, is het nodig om te detecteren in welke taal een document is geschreven. Deze informatie wordt in de volgende stappen gebruikt.

PII Detectie
Het detecteren en verwijderen van persoonlijke informatie, zoals namen, adressen en burgerservicenummers.

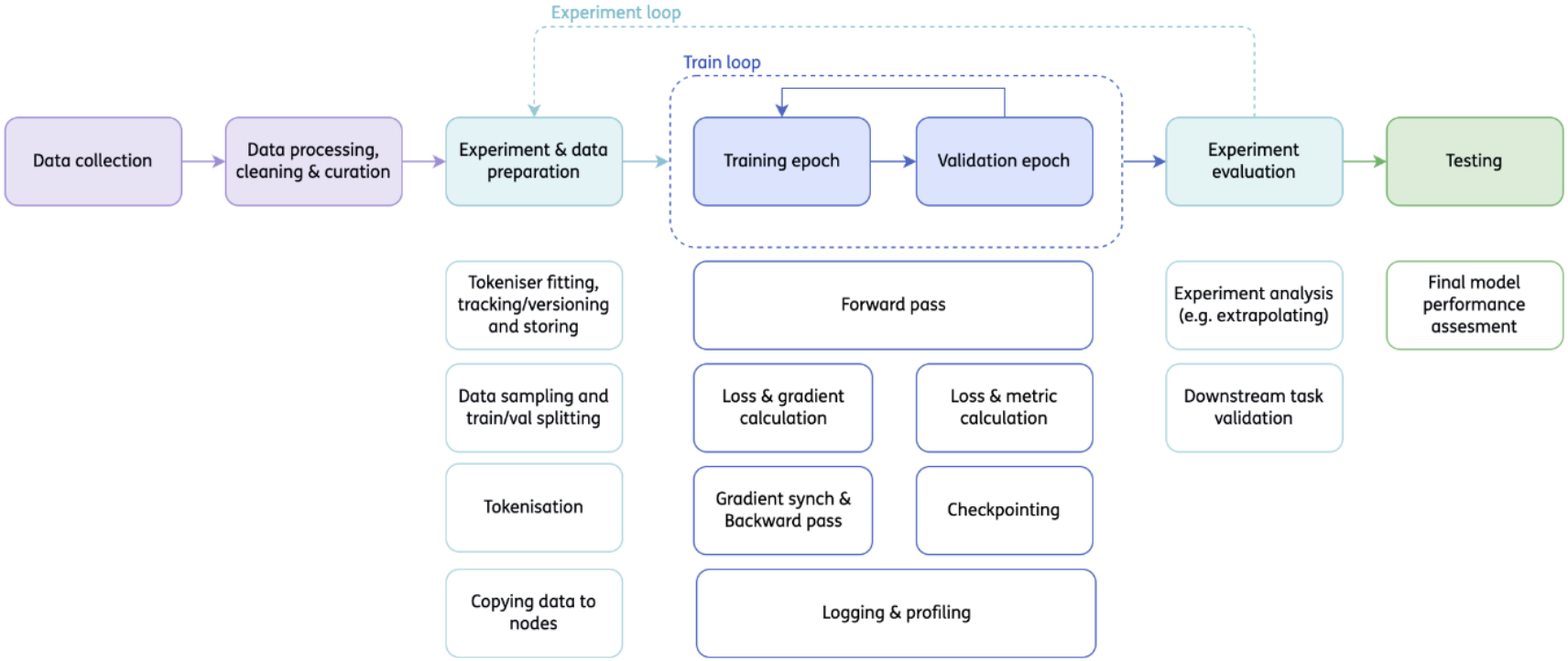
Harmful language
Detectie en verwijderen van schadelijke taal, zoals doodsbedreigingen en racistische uitlatingen.

Deduplicatie
Het verwijderen van duplicaten voor kwalitatievere output van het model.



Model architecture and framework

Het Model Architecture Team en Model Framework team werken samen om de efficiëntste manier te vinden om het model te trainen. Omdat er onvoorstelbaar veel rekenkracht gebruikt gaat worden om het model te trainen telt elke procent efficiëntiewinst. We hebben een vergelijking gemaakt tussen twee trainingframeworks. Deze resultaten zijn in december 2024 gepresenteerd op de Advance Computing User Day en komen in Q1 2025 op de site van GPT-NL.



Auteursrecht en GPT-NL: onontgonnen terrein

De New York Times die een rechtszaak aanspant tegen OpenAI en Microsoft wegens auteursrechtinbreuk, omdat de content van de krant zonder hun toestemming is gebruikt voor het trainen van ChatGPT. Stichting BREIN die een grote Nederlandstalige dataset offline haalt met illegale kopieën van boeken en ondertitels, bedoeld om LLMs op te trainen. Zomaar twee voorbeelden die laten zien dat voor de ontwikkeling van huidige LLMs vaak data is gebruikt zonder toestemming van de auteursrechthebbenden. Met GPT-NL haken we hierop in en streven we naar een schone dataketen. Voor het trainen wordt gezocht naar data waarvoor een geldige licentie is gegeven of in het geheel geen licentie nodig is. Bijvoorbeeld omdat de auteur al lang geleden is overleden. We scrapen niet zomaar zonder toestemming data van het internet.

Onze doelen vormen een ambitieus plan. De stappen die we hiervoor moeten nemen vormen dan ook een tijdrovend proces. In 2024 hebben we veel werk verzet om met auteursrechthebbenden in gesprek te gaan en een manier te vinden om deze data providers te vertegenwoordigen. Zo hebben we een Content Board opgezet waarin alle data providers deelnemen. Door deze samenwerking is het mogelijk een systeem op te zetten waarbij we auteursrechthebbenden kunnen compenseren vanuit de opbrengsten die onze LLM genereert.

Het project GPT-NL legt dan ook bloot dat we momenteel in een transitie zitten wat betreft auteursrechten. Door de mogelijkheden in dit onontgonnen terrein te verkennen, willen samen laten zien hoe je auteurs een eerlijke plek kunt geven in de transitie naar verantwoorde ontwikkeling van LLMs.

Onze commitments en kernwaarden

GPT-NL is een ambitieus project met idealistische inslag. Wij geloven dat technologie betrouwbaar en transparant moet zijn, een wederkerige bijdrage moet leveren, en soevereiniteit van Nederland moet versterken. Vanuit deze kernwaarden willen wij het AI innovatielandschap versterken. In onze missie-en-visie verhaal staat in meer detail uitgelegd wat ze voor ons betekenen.

Begin dit jaar hebben we vanuit dit perspectief onze commitments opgesteld (zie www.gpt-nl.nl/commitments). Deze helpen om onze ambities van het project 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.

Onze commitments helpen ons om een weg te vinden in de vele uitdagingen die we tegenkomen. Ze geven sturing wanneer we voor dilemma's staan en helpen onze kernwaarden hoog te houden. Dat betekent niet dat het uitdagingen makkelijker maakt. We hebben aan het eind van dit jaar een review gedaan en ons gerealiseerd dat we nog niet op ieder punt hebben bereikt wat we ons hebben voorgenomen. In sommige gevallen is het nog niet het juiste moment in het proces, in andere gevallen bleek een commitment toch lastiger om te realiseren dan we van tevoren hadden gedacht. Begin 2025 ronden we onze review af en publiceren we een reflectie en update op onze site.

Soevereiniteit

Wederkerigheid

Betrouwbaarheid

Transparantie

Licenties GPT-NL

Vanuit EZK is een subsidie ontvangen (Faciliteiten Toegepast Onderzoek, FTO) voor het opzetten van een onderzoeksfaciliteit rondom GPT-NL en een eerste ronde training van het model. Dat betekent dat er in 2025 niet al een afgerond product kan worden opgeleverd: er gaan meerdere iteraties nodig zijn voor de doorontwikkeling van het GPT-NL model.

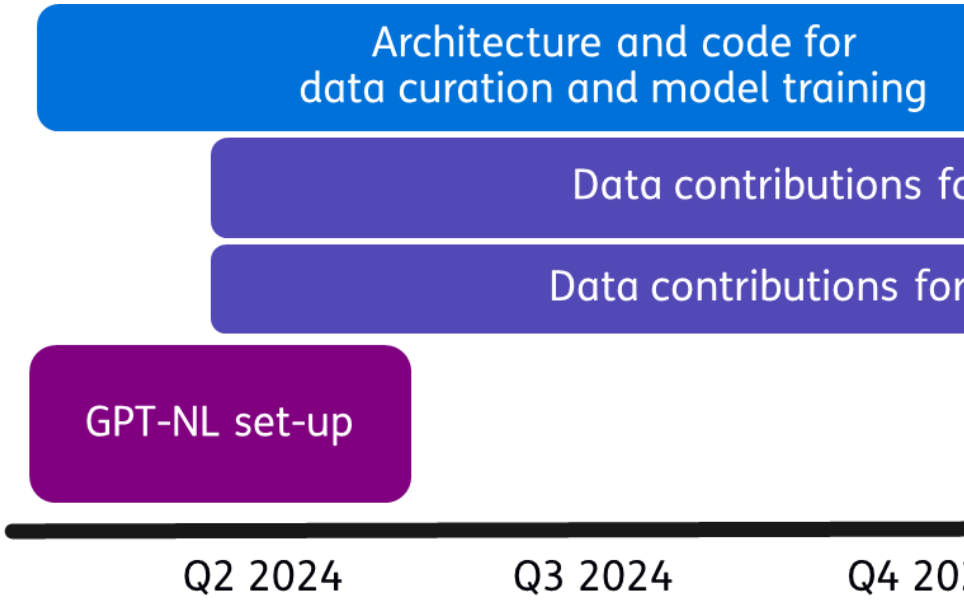
Om de kosten te dekken zal GPT-NL via licenties beschikbaar zijn. Dit jaar hebben we kunnen formuleren hoe de licenties eruit zullen zien:

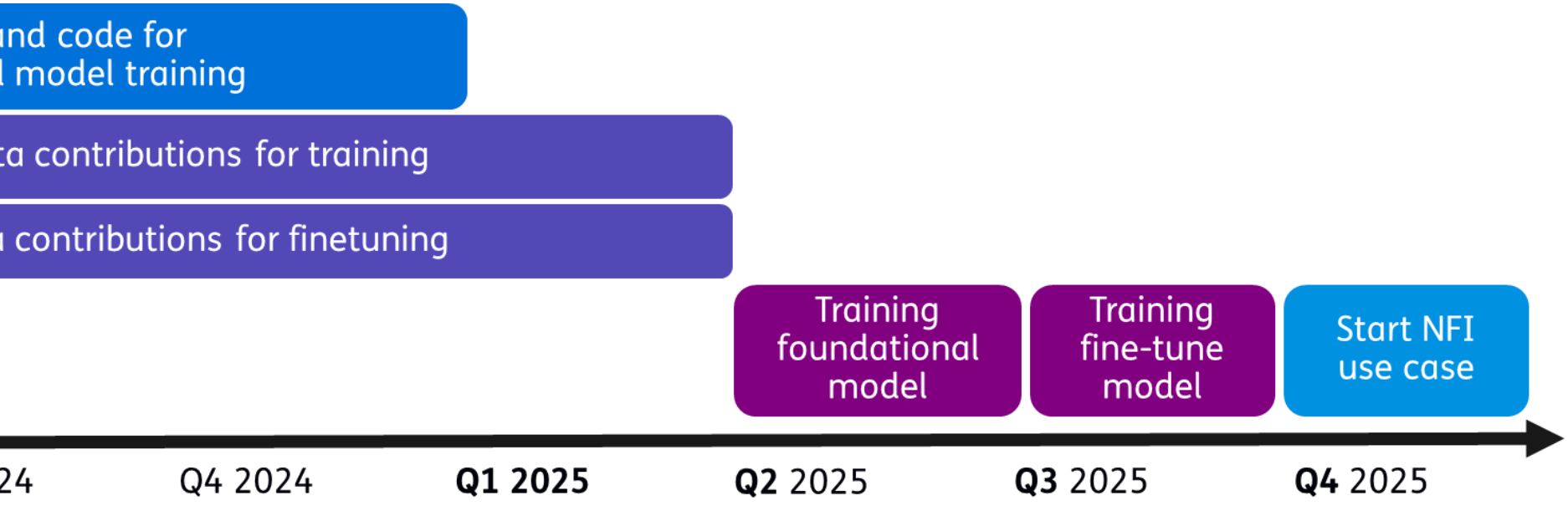
- **Een onderzoekslicentie**, beschikbaar voor non-commercieel gebruikt. Deze is goedkoop, maar het is verplicht om de onderzoeksresultaten te delen.

- **Een commerciële licentie,** op basis van *pay per use*. De helft van deze inkomsten gaat terug naar de data-aanbieders. Data-aanbieders kunnen dit ontvangen als financiële uitbetaling of het gebruiken om toegang te krijgen tot het model. De andere 50% zal enkel gebruikt worden voor verdere ontwikkeling van het taalmodel en van het GPT-NL project.

PLANNING

Hiernaast staat de planning weergeven zoals we die voor het GPT-NL model nu voor ogen hebben.




nd code for
model training
ta contributions for training
contributions for finetuning
Training foundational model
Training fine-tune model
Start NFI use case
24
Q4 2024
Q1 2025
Q2 2025
Q3 2025
Q4 2025

MISSION AND VISION
of the GPT-NL Model

Om scherp te krijgen waar GPT-NL en het team voor staat, hebben we in 2024 meerdere strategische sessies gehad om de richting en scope van het project te bepalen. Voor het GPT-NL Model hebben we dit doorvertaald in een heldere missie en visie, geschreven vanuit de why-how-what. Hierin staat kort en bondig wat we willen bereiken, hoe we dat gaan doen, en wat we dan precies in 2025 gaan opleveren.

Het missie en visie stuk is in het Engels opgemaakt.

Introduction

The GPT-NL initiative, a collaboration between TNO, SURF, and NFI, aims to support the Large Language Model community through the development of a state-of-the-art research facility and the pursuit of a sovereign, lawful, trustworthy and transparent language model. We aim to create useful and yet law compliant and sovereign large language models that provide an alternative to big tech's generative AI models. We share the knowledge gained during model development to foster a collaborative and open environment for LLM research. Furthermore, we strive at a fair and responsible use of the technology that benefits contributors and society alike.

Why

01#

Alternative

The current LLM-market does not offer sufficient transparency and cannot ensure that they comply with our laws and values.

02#

Complexity

The quality of current LLMs, that are trained on translated Dutch texts, is insufficient when it comes to understanding Dutch texts in complex cases.

03#

Believe

We believe that technology should be reciprocal, trustworthy, transparent, and ensure sovereignty of our citizens and institutions.

How

01#

A competitive model

By offering a useful alternative to existing LLMs, we'll demonstrate that it is possible to create language models that comply with our laws and values.

02#

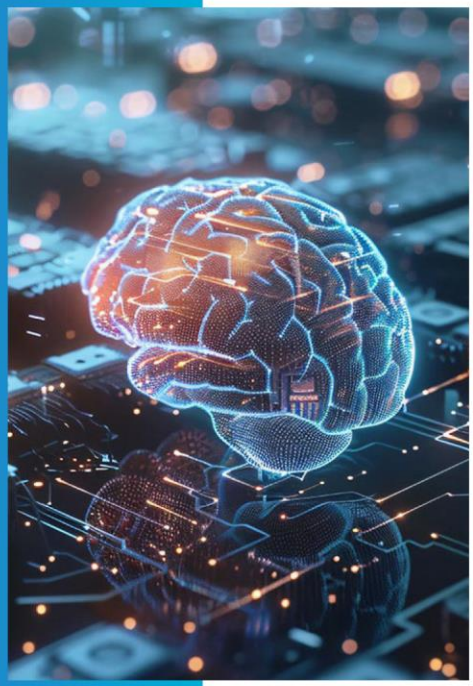
A new standard

By being open and transparent about the data sources and decisions made during the design process, we'll set a standard on how to develop LLMs in responsible way and in compliance with legislation.

03#

A strong ecosystem

By cooperating with stakeholders, we'll create a strong ecosystem that strengthens the European AI innovation landscape.



What

A RESPONSIBLE LARGE LANGUAGE MODEL BUILT FROM SCRATCH

Data

We will develop, train and fine-tune one large language model using mostly Dutch, English, German, and Code sources. This will be a combination of only:

- Opt-in data
- Data that is legally accepted for the training of LLMs
- Non-IP infringing synthetic data

Performance

The GPT-NL model will be trained on at least three hundred billion text tokens and can perform text generation, summarization, and simplification tasks at a level of performance comparable to the Llama2 7B model, GPT-3 175B models. The dataset will be finished with 1,5 hundred billion tokens of code.

Transparency

The GPT-NL Model will be supported by extensive documentation on the GPT-NL model to enable understanding and transparency, including a datasheet describing qualities of the dataset and a model card describing qualities of the model.

Core Values

Sovereignty

We believe in freedom of choice and independence from the big tech market. GPT-NL also believes in the relevance of cultural diversity language models.

Trustworthy

The GPT-NL Model is designed to be compliant with Dutch and European laws and to have a competitive, technical quality. Our organisation is reliable and committed to building Responsible AI.

Reciprocity

We believe innovation should benefit everyone and should contribute to a fair and inclusive society. Technology should be built in cooperation with important stakeholders.

Transparency

We are accountable, verifiable, and honest. We strive to be as open as possible.

Our Vision & Mission

What is our future dream?

Our future dream envisions a Europe capable of building powerful LLMs in the way that serves citizens best, wherein the technology adheres to our laws, norms, and values.

What will we achieve?

GPT-NL will strengthen the Dutch and European AI innovation landscape by supporting a fair, sovereign, and safe ecosystem of knowledge-share around performant, responsibly developed, and legally compliant LLMs.

How will we get there?

Through the development of the GPT-NL Model, comprehensive documentation, and active engagement with LLM communities, we will build a strong ecosystem where knowledge is shared.

Future **Goals**

In 1 year from now

we have launched the GPT-NL Model and developed a fair business model for data suppliers and content creators for their contributions.

In 5 years from now

our key partners have integrated customized versions of GPT-NL into their core processes, while the general public can access it through various trusted providers. This in turn has inspired the creation of multiple Dutch and European LLMs.

In 10 years from now

we will have strengthened the digital sovereignty of Europe by increasing capability in AI, by developing technologies aligned with the AI Act, and earning the trust of the European people.

Doe mee!

Voor het trainen van GPT-NL is een enorme hoeveelheid data nodig die divers genoeg is om GPT-NL breed toepasbaar te maken. Daarom is elke datadonatie van grote waarde. Met uw data wordt GPT-NL relevanter voor uw sector.

Wilt u meewerken aan de ontwikkeling van GPT-NL? Ga naar www.gpt-nl.nl/samenwerken of mail ons op info@gpt-nl.nl.

Hier kan je ons vinden

Wil je ons ontmoeten, of heb je vragen over een samenwerking? De onderstaande events staan op de planning. Online vind je altijd de laatste informatie. Ga naar onze website www.gpt-nl.nl of volg ons op LinkedIn.

- **IT Circle** – 28 januari 2025, Utrecht
- **E-Discovery Symposium** – 4 maart 2025, Leiden

Colofon

Financiering
Faciliteiten Toegepast Onderzoek (FTO)

Partners
TNO
SURF
NFI

Copywriters
Jochem Vreeman
Lieke Dom

Design
Jeroen Poots, TNO

GPT-NL

TNO innovation for life

Nederlands Forensisch Instituut
Ministerie van Justitie en Veiligheid

SURF