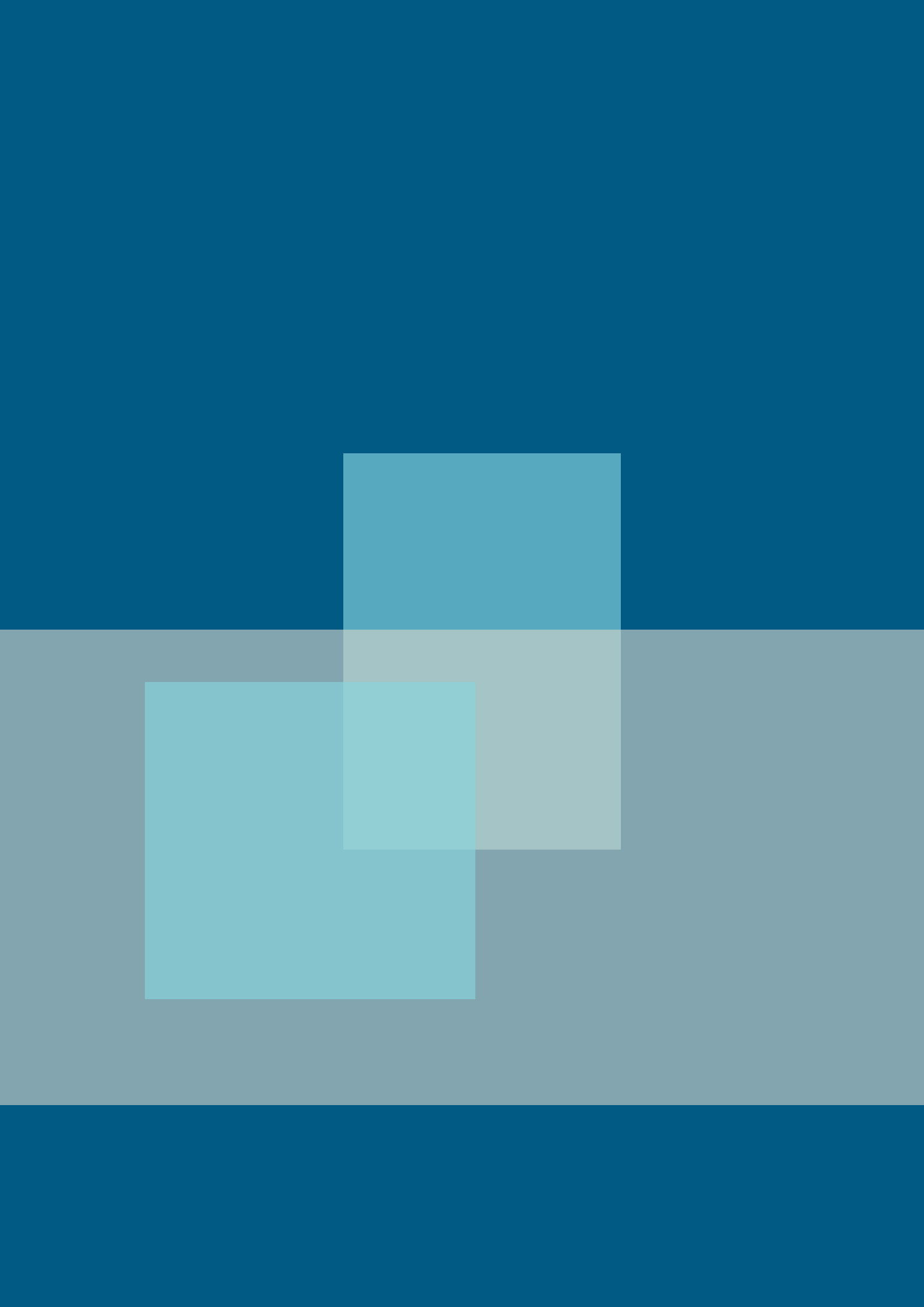




Actieplan Artificiële Intelligentie (AI) voor de Nederlandse Taal

**Digitale autonomie en weerbaarheid door te
bouwen aan een duurzaam AI-ecosysteem**



Auteurs

Tanguy Coenen, Saskia Lensink, Antal van den Bosch, Catia Cucchiari, Walter Daelemans, Martijn Kleppe, Roeland Ordeman, Vincent Vandeghinste, Hugo Van hamme, Bram Vanroy en Suzan Verberne.

Redactie

Levi Remijnse en Marten van der Meulen.

Disclaimer

Generatieve AI is gebruikt ter ondersteuning van taal en redactie. De inhoudelijke keuzes, verificatie en eindredactie zijn volledig door de auteurs uitgevoerd.

Dit is een document van het Algemeen Secretariaat van de Taalunie.

Inhoud

Samenvatting	6
1 Inleiding	9
1.1 Data: fundamentele knelpunten	10
1.2 Infrastructuur: beschikbaarheid en bruikbaarheid	10
1.3 AI-Modellen en benchmarking: doorlopende, toetsbare ontwikkeling	11
1.4 Valorisatie: van technologische capaciteit naar maatschappelijke en economische impact	12
1.5 Kennisdeling	12
1.6 Leeswijzer	12
2 - Pijler 1: Data	13
2.1 Inleiding	13
2.2 Soorten data	13
2.2.1 Tekstdata	13
2.2.2 Spraakdata	14
2.2.3 Multimodale data	16
2.3 Algemene data-aspecten	19
2.3.1 Een- of meertalige data?	19
2.3.2 Juridische aspecten	19
3 - Pijler 2: Infrastructuur	23
3.1 Inleiding	23
3.2 Rekeninfrastructuur	23
4 - Pijler 3: AI-ontwikkeling	27
4.1 Inleiding	27
4.2 Benchmarks	27
4.3 Pre-training of post-training?	29
4.4 Representativiteit, ethiek, soevereiniteit	31
5 - Pijler 4: Valorisatiepotentieel	33
5.1 Inleiding	33
5.2 Media en entertainment	33
5.3 Onderwijs	33
5.4 Publieke sector	33

5.5 Onderzoek	34
5.6 Juridische sector	34
5.7 Veiligheid	35
5.8 Zorg	35
5.9 Speerpuntclusters en topsectoren	36
6 - Pijler 5: Kennisdeling, governance en programmawerking	37
6.1 Inleiding	37
6.2 Kenniscentrum	37
6.3 Governance	38
6.4 Programmawerking	38
7 Conclusie	39
Referenties	40
Bijlage A: Het trainen van een taalmodel	42
Bijlage B: Lopende projecten	44
Colofon	48

Samenvatting

Dit document van het Algemeen Secretariaat van de Taalunie presenteert een actieplan voor de ontwikkeling van een duurzaam ecosysteem voor Nederlandstalige AI en bouwt voort op het advies van de Raad voor de Nederlandse Taal en Letteren “Digitale taalinfrastructuur en AI: aanbevelingen voor een verantwoord en duurzaam beleid”. Het actieplan bevat aanbevelingen verdeeld over vijf pijlers, **om grip te krijgen op de toekomst van Nederlandstalige AI** voor de Nederlandse taal en cultuur binnen het hele Nederlandse taalgebied, inclusief Nederland, Vlaanderen en Suriname.

AI-systemen worden in hoog tempo een onderdeel van de publieke dienstverlening, economie en informatievoorziening. Voor veel sectoren, waaronder zorg, onderwijs, media, openbaar bestuur en veiligheid, is AI die toegesneden is op de Nederlandse taal en cultuur essentieel. **Voor het Nederlands dreigt momenteel afhankelijkheid van buitenlandse modellen, datasets en infrastructuur.** Dat brengt risico's mee voor de kwaliteit van Nederlandstalige AI, maar ook voor digitale autonomie en soevereiniteit, weerbaarheid, inclusiviteit en culturele verankering.

Er is in het Nederlandse taalgebied al veel waardevolle kennis, data en basisinfrastructuur aanwezig, in universiteiten, erfgoedinstellingen, publieke organisaties en bedrijven. Daarnaast worden er in Nederland en Vlaanderen al diverse veelbelovende initiatieven ontplooid. Een recent voorbeeld is GPT-NL. Dit is een soeverein taalmodel ontwikkeld door TNO, SURF en het Nederlands Forensisch Instituut. GPT-NL is een verantwoord alternatief voor bestaande (Amerikaanse) taalmodellen, omdat het is gebouwd op rechtmatig verkregen Nederlandse data en transparant is over zijn trainingsproces. Een ander voorbeeld is het VLAM-programma in Vlaanderen, waarbij bestaande modellen worden aangepast zodat ze beter afgestemd zijn voor gebruik in bepaalde toepassingen, bijvoorbeeld in de mediasector. **GPT-NL en het VLAM-programma laten zien dat het mogelijk is om verantwoordelijkheid te nemen rond AI voor het Nederlands.**

GPT-NL, VLAM en andere bestaande initiatieven staan vooralsnog op zichzelf. Ze zijn geen AI-infrastructuur die breed inzetbaar is. Dit toont **het kernprobleem rond Nederlandstalige AI: versnippering**. Data, rechten, verwerkingsomgevingen en rekenkracht zijn er wel, maar komen niet vanzelf samen. Ook laten deze projecten zien hoe complex het is om dit soort initiatieven van de grond te krijgen en duurzaam in te richten. Een klassiek geval van lokaal marktfalen.

De kernoplossing is afstemming. Er is een modulaire en federatieve infrastructuur nodig om van losse projecten te evolueren naar een duurzame AI-voorziening voor het Nederlands. **Die leidt tot een zichzelf versterkend ecosysteem voor Nederlandstalige AI**, waarin kennis, data en infrastructuur één samenhangend geheel vormen. Waarin elke ontwikkeling, project of stap voortkomt uit een gedeelde visie. Waarin iedere investering de volgende mogelijk maakt. **De Taalunie kan zorgen voor afstemming** en een platform bieden voor structureel overleg tussen overheid, kennisinstellingen en veldspelers.

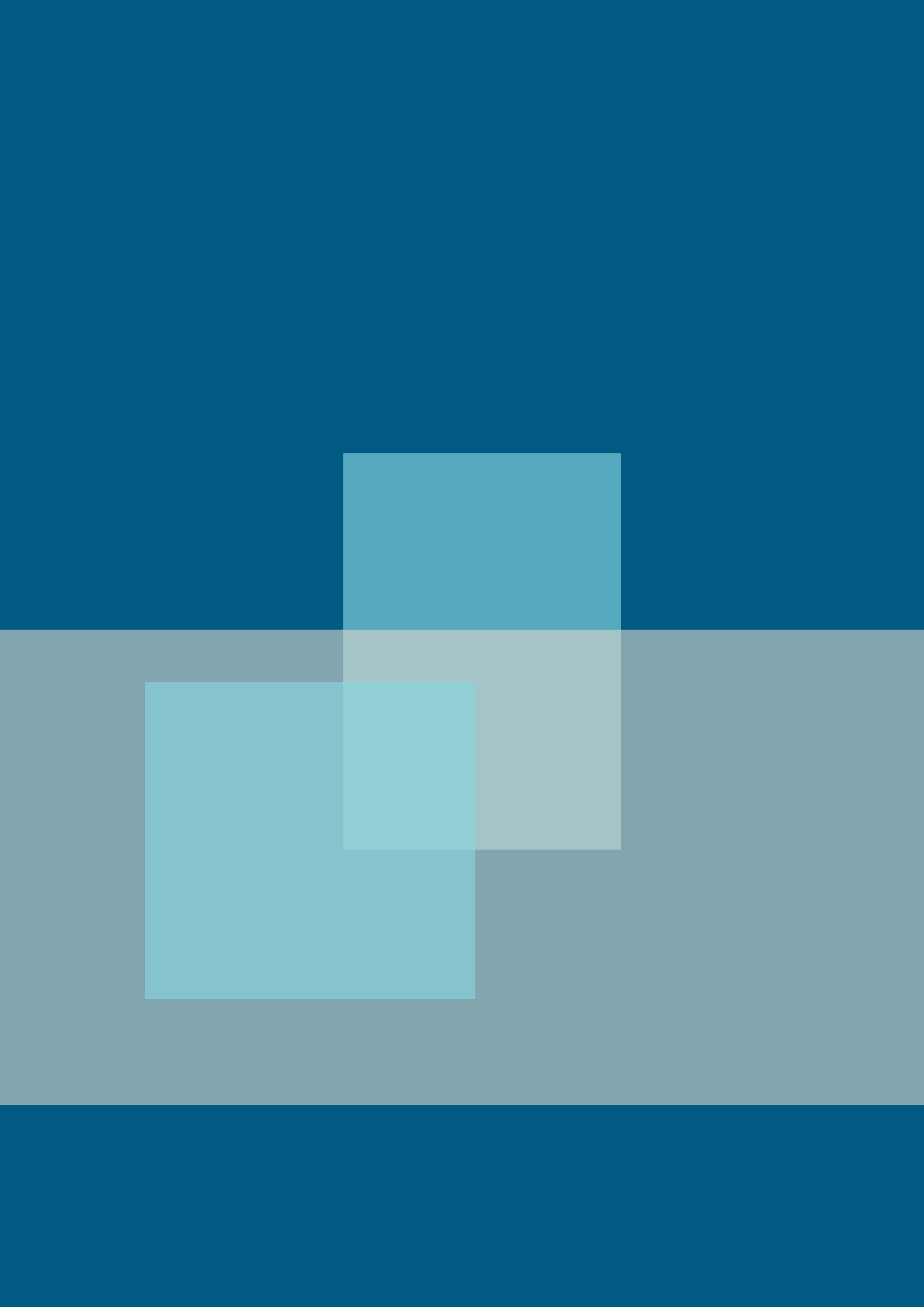
Om dit te bereiken stellen wij in dit actieplan een aanpak voor, die rust op vijf pijlers:

1. **Data.** Bestaande datacollecties voor het Nederlands zijn zelden grootschalig inzetbaar voor AI-training. Rechtenbeperkingen en privacyregels zorgen ervoor dat waardevolle digitale collecties, zoals die van de Koninklijke Bibliotheek in Nederland of Meemoo in Vlaanderen, wel doorzoekbaar zijn, maar niet bruikbaar voor de training van modellen. Met gerichte ontsluiting van deze data kunnen AI-modellen worden ontwikkeld die beter aansluiten bij de taalvariatie, cultuur en diverse gebruikersgroepen van het Nederlandse taalgebied.

2. **Infrastructuur.** Data, toegangsrechten, verwerkingsomgevingen en rekenkracht zijn verspreid over instellingen en systemen. Doordat gemeenschappelijke afspraken en toegangsmechanismen ontbreken, moet elk nieuw AI-project onderhandelen over toegang en capaciteit. Een federatieve infrastructuur maakt samenwerking mogelijk zonder data centraal op te slaan: organisaties behouden de controle over hun eigen voorzieningen, terwijl gedeelde standaarden en afspraken AI-ontwikkeling schaalbaar, duurzaam en uitbreidbaar maken.
3. **Modelontwikkeling en kwaliteitsborging.** De ontwikkeling van AI voor het Nederlands vraagt meer dan technische implementatie. Ze vereist mobilisatie van taalkundige en culturele expertise, AI-onderzoek, juridische kennis rond datagebruik en sectorale domeinkennis. Daarmee ontstaat systematische verankering in de volle breedte van de Nederlandse taal- en cultuurcontext. AI-modellen worden niet langer incidenteel ontwikkeld, maar systematisch geëvalueerd en vergeleken aan de hand van gezamenlijke kwaliteitskaders (benchmarking). Een doorlopende ontwikkel- en evaluatiecyclus maakt de prestaties van modellen beter toetsbaar, stimuleert continue verbetering en vergroot de publieke verantwoording.
4. **Valorisatie.** De maatschappelijke en economische nood aan AI voor het Nederlands is groot in vele sectoren, waaronder media, publieke dienstverlening, zorg, onderwijs en veiligheid. Toch blijft de toepassing van en kennis over de mogelijkheden rond AI versnipperd. Onzekerheid over kwaliteit, juridische risico's en afhankelijkheid van buitenlandse aanbieders remt adoptie. Domeinspecifieke valorisatie zou vorm moeten krijgen in een verbindend kader voor de structurele koppeling van ontwikkeling en gebruik, om daarmee de kans op economische en maatschappelijke impact van AI te vergroten.
5. **Kennisdeling.** Een expertisecentrum dat taalkundige, technologische, juridische en sectorale kennis bundelt, is onmisbaar voor beleidsmakers, bedrijven, onderzoekers en het brede publiek. Dit kenniscentrum kan onder andere kennisdeling op zich nemen die nodig is voor effectieve valorisatie. Een goede kandidaat hiervoor is het INT (Instituut voor de Nederlandse Taal). Dit wordt al door de Taalunie gefinancierd om verdragstaken voor de digitale taalinfrastructuur voor het Nederlands uit te voeren, heeft expertise op het gebied van AI en werkt samen met nationale en internationale partners uit kennisinstellingen, bedrijven en overheid.

Kernboodschap

Hoogwaardige AI voor de Nederlandse taal is geen luxe, maar een strategische noodzaak. Het vraagt geen eenmalige modelontwikkeling, maar een duurzaam ecosysteem. Alleen zo kunnen data verantwoord worden ontsloten, infrastructuur samenhangend worden georganiseerd, model-ontwikkeling doorlopend en toetsbaar plaatsvinden, en toepassingen systematisch worden verankerd in en over sectoren. Dit actieplan is een vertrekpunt. De volgende stap is het uitwerken van een gezamenlijk uitvoeringsplan. Dit plan bevat een gedeeld kompas voor kennisinstellingen, bedrijven, overheden en maatschappelijke organisaties in het hele Nederlandse taalgebied. Dit kompas kan uitgevoerd worden door een portfolio aan projecten, die kunnen worden gefinancierd via verschillende instrumenten. Dat vraagt gerichte investering – publiek én privaat – als strategische keuze. Hiermee kunnen we samen grip krijgen op de digitale toekomst van het Nederlands in het digitale tijdperk.



1 Inleiding

AI-systemen worden in hoog tempo een onderdeel van publieke dienstverlening, informatievoorziening en economische concurrentiekracht. De dominante technologieën worden vandaag grotendeels ontwikkeld buiten het Nederlandse taalgebied volgens normen, datasets en prioriteiten die niet noodzakelijk aansluiten bij onze taal, cultuur en publieke waarden. Zonder gerichte actie dreigt structurele afhankelijkheid van buitenlandse modellen en infrastructuur, met gevolgen voor digitale autonomie, democratische weerbaarheid en de economische positie van Nederland, Vlaanderen en Suriname. Investeren in AI voor taal en cultuur in het Nederlandse taalgebied is daarom geen luxe of prestigeproject, maar een strategische noodzaak. Het gaat niet om het eenmalig bouwen van een groot model, maar om het ontwikkelen van een duurzaam ecosysteem dat technologische evoluties kan volgen, publieke waarden kan borgen en economische innovatie kan ondersteunen.

In dit document stellen we een actieplan voor om de competentie rond AI voor het Nederlands in het Nederlandse taalgebied te versterken. Onder het Nederlandse taalgebied verstaan we in dit actieplan Nederland, Vlaanderen, Suriname en het Caribisch deel van het Koninkrijk der Nederlanden. Hoewel de huidige kennisbasis vooral uit Nederland en Vlaanderen komt, dient de verdere uitwerking van deze agenda expliciet aandacht te besteden aan de taalvariatie, maatschappelijke contexten en institutionele betrokkenheid van al deze onderdelen van het taalgebied. Dit actieplan is in de eerste plaats bedoeld voor beleidsmakers, bestuurders, financiers en organisaties die betrokken zijn bij de ontwikkeling van AI voor het Nederlands. Het document biedt een strategisch kader voor gezamenlijke beleidsvorming en samenwerking binnen het Nederlandse taalgebied.

Het actieplan vertrekt vanuit het in kaart brengen van de noden en kansen binnen vijf samenhangende pijlers:

1. Data: het systematisch ontsluiten en juridisch borgen van representatieve tekst-, spraak- en multimodale data;
2. Infrastructuur: het opzetten en onderhouden van veilige en schaalbare voorzieningen waarin data op een verantwoorde manier toegankelijk zijn en modellen reproduceerbaar kunnen worden ontwikkeld en gebruikt;
3. AI-ontwikkeling: het toetsen, beheren en doorlopend verbeteren van modellen met aandacht voor kwaliteit, bias en publieke waarden uit het Nederlandse taalgebied;
4. Valorisatie: het vertalen naar concrete toepassingen in publieke en private sectoren die economische innovatie en maatschappelijke participatie versterken;
5. Kennisdeling: het delen van kennis over AI voor het Nederlands door middel van een expertisecentrum, waarbij zowel beleidsmakers, bedrijven als het brede publiek te rade kunnen gaan.

Binnen deze pijlers wordt onderzocht waar samenwerking binnen het Nederlandse taalgebied meerwaarde creëert, waar hiaten bestaan en welke randvoorwaarden nodig zijn om het ecosysteem duurzaam te versterken.

Technische beschrijvingen en voorbeelden zijn opgenomen ter onderbouwing van de voorgestelde beleidskeuzes. De nadruk ligt daarbij expliciet op het formuleren van inhoudelijke bouwstenen voor een gezamenlijke agenda en het richting geven aan een toekomstige uitvoeringsagenda. Het actieplan benoemt strategische keuzes, randvoorwaarden en mogelijke interventies die nodig zijn om een duurzaam AI-ecosysteem voor het Nederlands te realiseren. Het document bevat geen uitgewerkte planning, prioritering of financieringsvoorstellen; deze worden in een volgende fase gezamenlijk uitgewerkt. Hiervoor fungeert het actieplan als een inhoudelijk en strategisch fundament. De toekomstige uitvoeringsagenda kan gerealiseerd worden door een ecosysteem binnen het Nederlandse taalgebied en rust op de vijf samenhangende pijlers.

1.1 Data: fundamentele knelpunten

De ontwikkeling van AI voor het Nederlands wordt niet primair geremd door algoritmen, maar door structurele knelpunten in de beschikbaarheid en inzetbaarheid van data. Data bestaan in verschillende typen (tekst, spraak en multimodaal) en zijn te vinden bij verschillende commerciële en ideële organisaties. Over alle datatypen heen speelt één fundamenteel probleem: er zijn wél waardevolle en representatieve data aanwezig in het taalgebied, maar deze zijn niet georganiseerd op een manier die grootschalige, juridisch verantwoorde AI-ontwikkeling mogelijk maakt. Het oplossen van dit probleem is een kernvoorwaarde voor digitale autonomie. Meer in het bijzonder spelen specifieke problemen voor respectievelijk tekst-, spraak- en multimodale data:

- **Tekstdata:** Voor robuuste taalmodellen zijn grote, diverse en actuele corpora nodig. Juist de meest representatieve collecties (boeken, krantenarchieven, webarchieven) zijn vaak contractueel of juridisch niet inzetbaar voor AI-training. Daarnaast is taalvariatie (uit Nederland, Vlaanderen, Suriname; formeel en informeel taalgebruik) onvoldoende systematisch geborgd. Het fundamentele probleem is dus niet dat er te weinig tekst bestaat, maar dat de meest relevante tekst niet verantwoord en grootschalig gebruikt kan worden.
- **Spraakdata:** Gesproken interactie wordt steeds vaker de toegangspoort tot digitale dienstverlening. Toch zijn beschikbare spraakdatasets beperkt in omvang en niet altijd representatief voor accenten, dialecten en spontane spraak. Bovendien zijn spraakdata juridisch gevoeliger vanwege privacywetgeving. Het kernprobleem hier is dat zonder gerichte ontsluiting en juridische duidelijkheid spraakmodellen bepaalde bevolkingsgroepen systematisch slechter bedienen.
- **Multimodale data:** Toepassingen in media, onderwijs en publieke dienstverlening vragen steeds vaker om combinaties van tekst, beeld en audio. Nederlandstalige multimodale datasets zijn echter beperkt beschikbaar en kennen complexe rechtenstructuren. Daardoor worden modellen vaak getraind op niet-Nederlandstalige contexten, met als gevolg dat gegenereerde beelden, audio of video onvoldoende aansluiten bij lokale cultuur en maatschappelijke realiteit.

1.2 Infrastructuur: beschikbaarheid en bruikbaarheid

Voor AI-ontwikkeling volstaat het niet dat data ergens beschikbaar of zoekbaar zijn. Ze moeten ook op een veilige, juridische en reproduceerbare manier kunnen worden gebruikt om modellen te trainen en te verbeteren. Precies daar wringt het vandaag.

Veel waardevolle datasets zijn verspreid over instellingen en systemen met uiteenlopende toegangsvoorwaarden en rechtenbeperkingen. Daardoor moet voor elk project opnieuw worden uitgezocht wat mag, wat technisch kan en onder welke voorwaarden. Dit leidt tot vertraging, hoge kosten en beperkte schaal.

Daarnaast is passende rekenkracht een knelpunt. Het trainen, verbeteren en gebruiken van modellen vraagt gespecialiseerde hardware, zoals Graphics Processing Units (GPU's). Publieke voorzieningen zijn schaars en worden vaak competitief toegewezen, terwijl het gebruik van commerciële clouddiensten snel en gebruiksvriendelijk, maar structureel duur is. Zonder gezamenlijke coördinatie ontstaat afhankelijkheid van buitenlandse infrastructuur of remt de innovatie af.

Het kernprobleem is versnippering: data, rechten, verwerkingsomgevingen en rekenkracht zijn er wel, maar niet geïntegreerd in één samenhangende keten. Een modulaire en federatieve (gedistribueerde, zonder centrale opslag werkende) infrastructuur - met duidelijke dataregisters, gestandaardiseerde afspraken en gecontroleerde verwerkingsomgevingen - is nodig om van losse projecten te evolueren naar een duurzame AI-voorziening voor het Nederlands.

1.3 AI-Modellen en benchmarking: doorlopende, toetsbare ontwikkeling

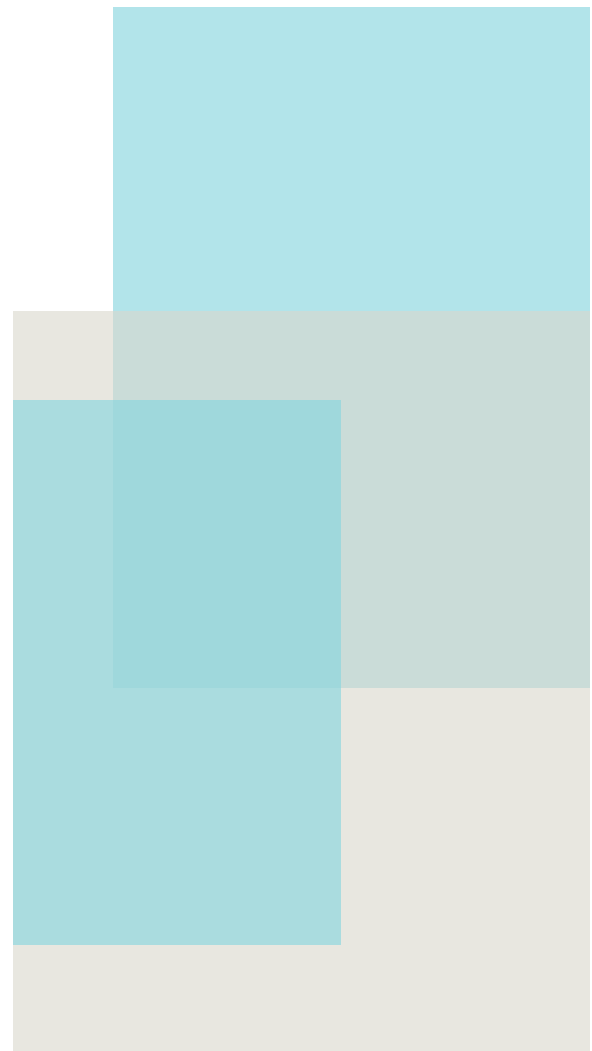
In deze pijler vindt de feitelijke ontwikkeling en verbetering van AI-modellen voor het Nederlands plaats. Vandaag gebeurt die ontwikkeling vaak projectmatig en versnipperd, terwijl de behoefte structureel en groeiend is. Zonder gezamenlijke aanpak blijven verbeteringen tijdelijk en ontbreekt een duurzame ontwikkelcyclus.

Een fundamenteel knelpunt is dat modellen wel functioneren in het Nederlands, maar niet systematisch zijn afgestemd op de volle breedte van de Nederlandstalige culturele context. Taalvariatie tussen Nederland, Vlaanderen en Suriname, domeinspecifiek taalgebruik uit bijvoorbeeld zorg of recht, informele registers en culturele referenties worden niet vanzelfsprekend of consistent meegenomen in de modellen. Daardoor blijven die soms oppervlakkig of onvolledige representatief.

Bovendien ontbreekt vaak een gedeeld en robuust benchmarkingkader: een kader waarin met een gestructureerde aanpak de prestaties van taalmodellen objectief worden gemeten, vergeleken en geëvalueerd op specifieke taken (benchmarking).

Een benchmark is gebaseerd op een uitgebreide en handmatig gecureerde verzameling vragen en antwoorden over specifieke domeinen of vaardigheden. Door de score van verschillende taalmodellen te vergelijken op die benchmarks kan op een eerlijke manier worden vergeleken welke modellen het best geschikt zijn voor een specifieke taak. Verder kunnen benchmarks ook worden gebruikt om te bepalen of bepaalde aanpassingen aan een model ook een positieve meerwaarde bieden ten opzichte van een bepaalde toepassing. Een benchmark voor evaluatie van het model op zijn kennis van de taal en de cultuur van de Nederlandstalige regio's is noodzakelijk om bestaande taalmodellen te evalueren en nieuwe te ontwikkelen. Er is ook nood aan lokale benchmarks voor spraak en andere modaliteiten, zoals afbeeldingen en video. Zonder structurele toetsing is gerichte verbetering moeilijk.

Het kernprobleem hier is dus het ontbreken van een doorlopende ontwikkel- en verbetercyclus met systematische benchmarking, transparante documentatie en periodieke updates. AI-ontwikkeling moet worden georganiseerd als een duurzame voorziening die continu leert, bijstuurt en expliciet verankerd is in de Nederlandse taal- en cultuurcontext. Alleen zo ontstaat betrouwbare, toetsbare en toekomstbestendige AI voor het Nederlands die economisch en maatschappelijk valoriseerbaar is door organisaties uit het Nederlandse taalgebied.



1.4 Valorisatie: van technologische capaciteit naar maatschappelijke en economische impact

Zelfs wanneer data, infrastructuur en modelontwikkeling geregeld zijn, ontstaat maatschappelijke en economische waarde niet automatisch. Een fundamenteel knelpunt is dat AI-kennis en -modellen vaak in onderzoeks- of experimentele sfeer blijven hangen, zonder structurele doorvertaling naar sectoren waar de impact groot kan zijn. In sectoren zoals media, publieke dienstverlening, onderwijs en zorg groeit de behoefte aan betrouwbare taal- en spraaktechnologie. Tegelijk zijn organisaties terughoudend door onzekerheid over kwaliteit, juridische risico's en afhankelijkheid van buitenlandse aanbieders. Zonder duidelijke kwaliteitskaders, sectorale samenwerking en gedeelde infrastructuur blijft toepassing versnipperd en ad hoc.

Daarnaast dreigt economische afhankelijkheid wanneer bedrijven en kennisinstellingen uit het Nederlandse taalgebied voor kerntechnologie volledig aangewezen zijn op internationale platformen. Dat heeft niet alleen gevolgen voor de concurrentiekracht, maar ook voor democratische weerbaarheid en culturele autonomie, zeker in sectoren zoals media, journalistiek of onderwijs.

Het kernprobleem is dus niet een gebrek aan toepassingsmogelijkheden, maar het ontbreken van een domeinspecifiek verbindend kader voor de structurele koppeling tussen ontwikkeling en gebruik. Valorisatie vraagt om gerichte sectorale samenwerking, gedeelde kwaliteitsnormen en een ecosysteem waarin publieke en private partijen samen innoveren. Alleen dan vertaalt AI voor het Nederlands zich in inclusieve publieke dienstverlening, economische slagkracht en versterkte digitale autonomie.

1.5 Kennisdeling

Het is belangrijk dat er een expertisecentrum komt waar zowel de bedrijfswereld, beleidsmakers, onderzoekers, als het brede publiek kunnen aankloppen voor informatie over de huidige stand van zaken met betrekking tot AI voor het Nederlands. Dit expertisecentrum deelt taalkundige expertise over taalvariatie en cultuur, technologische knowhow in AI en infrastructuur, en juridische en ethische deskundigheid rond datagebruik. Alleen door deze expertise structureel te bundelen kan het Nederlandse taalgebied grip houden op zijn digitale toekomst en voorkomen dat taaltechnologie een externe afhankelijkheid wordt in plaats van een strategische kracht.

1.6 Leeswijzer

In de rest van dit actieplan worden de vijf pijlers van dit advies nader uitgewerkt. Dit gebeurt langs de volgende samenhangende invalshoeken: 1) data (tekst, spraak, multimodale data en juridische aspecten), 2) infrastructuur (data- en rekenvoorzieningen), 3) AI-ontwikkeling (benchmarking en training), 4) valorisatie (juridische borging en concrete toepassingsdomeinen) en 5) kennisdeling. In de eerste drie pijlers wordt systematisch ingegaan op de huidige situatie, de geïdentificeerde knelpunten en noodzakelijke verbeteringen, de meerwaarde van samenwerking binnen het Nederlandse taalgebied, en ook op aanbevolen acties binnen het Nederlandse taalgebied. Pijler 4 geeft enkele voorbeelden van sectoren waarin structurele doorvertaling tot impact kan leiden. Pijler 5 behandelt puntsgewijs samenwerking, kennisdeling en governance. Het advies wordt afgesloten met een beschouwing van de randvoorwaarden voor duurzame implementatie, de aansluiting bij lopende nationale en Europese initiatieven, en een integrale samenvatting van de aanbevelingen. Een algemene inleiding tot het trainen van AI-modellen is te vinden in [Bijlage A](#). Een overzicht van lopende projecten omtrent Europese en specifiek Nederlandse AI-modellen is te vinden in [Bijlage B](#).

2 – Pijler 1: Data

2.1 Inleiding

Deze pijler beschrijft welke data nodig zijn om hoogwaardige AI voor het Nederlands mogelijk te maken en welke beleidsmatige randvoorwaarden daarvoor moeten worden vervuld. De technische beschrijvingen van tekst-, spraak- en multimodale data dienen ter illustratie van de belangrijkste knelpunten en investeringsnaden. De nadruk ligt op het creëren van voorwaarden voor duurzame toegang tot kwalitatieve, representatieve en juridisch bruikbare data binnen het Nederlandse taalgebied.

Deze pijler bestaat uit twee delen. Eerst worden de verschillende soorten data beschreven die relevant zijn voor AI-ontwikkeling voor het Nederlands. Vervolgens worden de belangrijkste aandachtspunten behandeld die van invloed zijn op de beschikbaarheid, kwaliteit en het verantwoord gebruik van deze data. Elk onderdeel wordt afgesloten met aanbevelingen voor vervolgstappen.

2.2 Soorten data

2.2.1 Tekstdata

Huidige situatie

Taaltechnologie voor tekst vormt een fundamenteel onderdeel van de digitale taalinfrastructuur voor het Nederlands. Technologieën zoals machinevertaling, informatie-extractie en automatisch samenvatten worden steeds vaker ingezet in digitale dienstverlening. In veel toepassingen fungeert tekst als interface tussen mens en AI-systemen.

Grote taalmodellen (LLM's) zijn de drijvende kracht achter de snelle vooruitgang in dit domein. Deze modellen worden getraind op zeer grote tekstcollecties en kunnen uiteenlopende taken uitvoeren. Vaak zijn ze echter primair ontwikkeld voor het Engels en enkele andere grote talen, waardoor de kwaliteit van het Nederlands afhankelijk blijft van de hoeveelheid beschikbare trainingsdata en de representativiteit daarvan.

Binnen het Nederlandse taalgebied bestaat een sterke onderzoekstraditie op het gebied van taaltechnologie. Verschillende Europese en nationale initiatieven richten zich momenteel op de ontwikkeling van open taalmodellen en datasets voor verschillende Europese talen, waaronder het Nederlands. Hierbij wordt gestreefd naar meer transparantie in trainingsdata en modelarchitecturen als alternatief voor de gesloten commerciële systemen.

Het Nederlandse taalgebied beschikt over een rijk, maar versnipperd landschap van corpora, lexicaal databanken en taalkundig geannoteerde datasets, die via onderzoeksinfrastructuren zoals [CLARIN](#) beschikbaar gemaakt worden voor wetenschappelijk gebruik. Vaak zijn zij ontwikkeld binnen specifieke projecten en onder uiteenlopende licentievoorzettingen, waardoor grootschalig hergebruik niet vanzelfsprekend is.

Een bijkomende uitdaging betreft de dekking van verschillende domeinen van taalgebruik. Beschikbare corpora bevatten relatief veel formele tekst, terwijl hedendaagse taaltechnologie ook moet kunnen omgaan met informele tekstuele communicatie en domeinspecifieke terminologie. Bovendien verandert taalgebruik voortdurend, wat regelmatige actualisering van datasets noodzakelijk maakt.

Een centraal onderdeel binnen het [GPT-NL](#)-project is het grootschalig en juridisch verantwoord verzamelen van Nederlandstalige tekstdata. Er is ingezet op het bijeenbrengen van uiteenlopende tekstbronnen uit journalistiek, overheid, wetenschap, juridische domeinen en professionele communicatie, via een zorgvuldig proces van rechteninventarisatie, metadata-harmonisatie en governance-afspraken met data-aanbieders. Daarbij zijn methoden ontwikkeld voor transparante documentatie van herkomst, licenties, toegestane verwerkingsdoeleinden en kwaliteitscontrole; aspecten die cruciaal zijn voor verantwoorde modeltraining onder zowel auteursrecht als privacywetgeving. Een belangrijk resultaat van dit project is dat [een deel](#) van de verzamelde datasets ook inzetbaar is voor andere onderzoeks- en innovatieactiviteiten.

Voorgestelde verbeteringen

Voor een duurzame verankering van taaltechnologie voor tekst in het Nederlandstalige ecosysteem is een meer systematische aanpak nodig voor het beheer van datasets, evaluatie en modellen.

Ten eerste is het belangrijk om prestaties van tools en modellen met gedeelde kaders te evalueren over verschillende domeinen, tekstgenres en toepassingen. Dit in aanvulling op algemene evaluaties. Er is nood aan benchmarks. We komen hierop terug in pijl 3.

Ten tweede is er nood aan een beter juridisch kader voor de ontsluiting van datasets, omdat veel taaldata onder auteursrecht vallen of persoonsgegevens bevatten. Binnen een dergelijk kader leidt het combineren van bestaande corpora en het verzamelen van nieuwe datasets, samen met het harmoniseren van metadata en licentie-informatie mogelijk tot betere modellen.

Tot slot is structurele aandacht nodig voor de documentatie en het beheer van taalmodellen zelf. Transparantie over trainingsdata, evaluatieprocedures en modelupdates is een belangrijke voorwaarde voor reproduceerbaarheid, uitlegbaarheid en publieke verantwoording. Integratie met bestaande onderzoeksinfrastructuren kan hierbij helpen.

Samenwerking binnen het Nederlandse taalgebied

Een gezamenlijke aanpak voor taaltechnologie voor tekst ligt voor de hand. Door datasets, expertise en infrastructuur te combineren kan een schaalniveau worden bereikt dat nodig is voor de ontwikkeling van competitieve Natural Language Processing (NLP)-systemen (systemen die natuurlijke taal kunnen verwerken en produceren) en taalmodellen. Dit kan via gezamenlijke inventarisaties van corpora, afstemming van metadata- en licentiestandaarden, en de ontwikkeling van gedeelde benchmarksets. Op infrastructuurniveau bestaan duidelijke synergieën. Afstemming met onderzoeksinfrastructuren, zoals CLARIN, draagt bij aan een gedeelde toegang tot datasets en tools. Daarnaast kan samenwerking rond rekeninfrastructuur en modelontwikkeling ook schaalvoordelen opleveren. Ten slotte is juridische afstemming wenselijk, met name rond auteursrecht, databankrecht en privacy. Heldere en waar mogelijk geharmoniseerde kaders kunnen het verantwoord hergebruik van datasets en modellen binnen het Nederlandse taalgebied aanzienlijk vergemakkelijken.

2.2.2 Spraakdata

Huidige situatie

Spraakherkenning (Automatic Speech Recognition, ASR) vormt een essentiële bouwsteen van de digitale taalinfrastructuur voor het Nederlands. In een groeiend aantal toepassingen fungeert gesproken taal als primaire toegangsvorm tot digitale diensten, waaronder publieke en maatschappelijke dienstverlening. Spraakherkenning verschilt daarbij fundamenteel van generatieve AI-systemen, gezien zij bestaande gesproken taal omzet naar tekst zonder nieuwe inhoud te creëren.

Voor het Nederlands zijn verschillende open-source modellen voor spraakherkenning beschikbaar. Internationaal bekende modellen, zoals de meertalige Whisper-architecturen, bieden ondersteuning voor het Nederlands, maar hun prestaties blijven merkbaar achter bij die voor het Engels. Met name regionale uitspraak, taalvariatie, spontane spraak en spraak van specifieke groepen zoals kinderen en tweedetaallearders vormen een uitdaging. Bovendien ontbreekt transparantie over de gebruikte trainingsdata, wat vragen oproept over reproduceerbaarheid en juridische borging. Vanuit de academische wereld is daarom ingezet op alternatieven zoals het [Open Whisper-style Speech Model \(OWSM\)](#), een meertalig spraakherkenningsmodel dat uitsluitend is getraind op publiek toegankelijke data. Hoewel deze aanpak bijdraagt aan transparantie, blijven de prestaties voor het Nederlands vooralsnog onder de maat.

Het Nederlandse taalgebied beschikt over waardevolle spraakcorpora die bij het [Instituut voor de Nederlandse Taal \(INT\)](#) beschikbaar zijn, zoals het [Corpus Gesproken Nederlands](#), [JASMIN](#) en [CHOREC](#). Deze bronnen zijn echter onvoldoende omvangrijk en niet representatief genoeg om robuuste benchmarking en grootschalige modellering van hedendaagse spraakherkenningssystemen te ondersteunen. Andere spraakdatasets die zijn verzameld in het kader van onderzoek rond tweedetaalverwerving en leesonderwijs zijn beperkt bruikbaar voor modelontwikkeling door strikte ethische en juridische voorwaarden.

Naast spraakherkenning vormt spraaksynthese (Text To Speech Synthesis, TTS) een belangrijke AI-bouwsteen. Moderne systemen implementeren TTS als een conditioneel generatieprobleem: een generatief model voorspelt sequenties van discrete audio-tokens. De combinatie van de uit te spreken tekst met een korte stemprompt laat het model toe spraak te genereren in een specifieke stem (*voice cloning*). Variaties in prosodie en emotie worden gestuurd via bijkomende conditionering, zoals stijlemboddingen of expliciete controle over toonhoogte en energie. Door de noodzaak aan grote hoeveelheden gevarieerde data zijn het vooral commerciële spelers die goede kwaliteit kunnen leveren. Bovendien zijn er voor langere teksten nog steeds problemen met het genereren van een zinsmelodie die past bij de betekenis van de tekst. Ook voor spraaksynthese blijven opensourcemonellen voor het Nederlands achter, zeker als het gaat om controleerbaarheid en aanpasbaarheid van stemmen, emoties, stijl en consistentie van taalvariëteit.

Voorgestelde verbeteringen

In het kader van een toekomstbestendige digitale taalinfrastructuur volstaat het niet om spraakherkenning en -synthese uitsluitend te benaderen als een technische uitdaging. Er is behoefte aan structurele borging van spraakherkenning en -synthese binnen het bredere AI-beleid voor het Nederlands. Een eerste prioriteit is het expliciet opnemen van spraakherkenning in benchmarking- en evaluatiekaders voor AI-systemen. Daarbij moet niet alleen worden gekeken naar gemiddelde nauwkeurigheid, maar ook naar prestaties over verschillende vormen van taalvariatie, zoals regionale accenten, spontane spraak en tempo. Dit is essentieel om risico's op systematische uitsluiting of ongelijke behandeling tijdig te identificeren en af te zwakken.

De beschikbaarheid van voldoende kwalitatieve, representatieve en juridisch goed ontsloten spraakdata vormt een essentiële voorwaarde voor zowel benchmarking als modellering. Dit betreft zowel hedendaagse spraakdata, bijvoorbeeld uit de media en de publieke sector, als onderzoek naar spraakvariatie. Taalgebruik verandert namelijk voortdurend, waardoor nieuwe vormen van variatie ontstaan die in bestaande datasets nog onvoldoende zijn vastgelegd. Deze ontsluiting dient

te gebeuren volgens de Algemene Verordening Gegevensbescherming (AVG), met een duidelijke afbakening van verwerkingsdoeleinden, rolverdeling tussen verwerkingsverantwoordelijke en verwerker, en – waar nodig – een gezamenlijke Data Protection Impact Assessment (DPIA) voor modeltraining.

Het uitbreiden, actualiseren en duurzaam inzetten van spraakdata vraagt om een intensievere samenwerking met lokale en regionale instellingen, zoals omroepen, archieven en andere gespecialiseerde organisaties die beschikken over omvangrijke spraakdatabestanden uit specifieke maatschappelijke domeinen. Cruciaal daarbij is dat deze data systematisch worden ontsloten via datacatalogi, waarin informatie wordt vastgelegd over de aard en samenstelling van de datasets, hun representativiteit, en de voorwaarden waaronder zij kunnen worden gebruikt. Dergelijke catalogi vormen een schakel tussen data-eigenaren, onderzoekers en modelontwikkelaars.

Daarnaast is expliciete aandacht nodig voor de juridische ontsluiting van spraakdata. Merk hierbij op dat spraakherkenningsmodellen – in tegenstelling tot LLM's en spraaksynthese – niet generatief (kunnen) worden ingezet. Dit gegeven maakt de auteursrechtelijke discussie minder complex omdat de identiteit en de inhoud die de bronsprekers aanleveren niet herkenbaar zijn. Aan de andere kant worden spraakdata als “persoonlijke data” bestempeld, wat leidt tot juridische beperkingen. Een verfijnd wettelijk kader ontbreekt hier. Dit is een noodzakelijke voorwaarde voor publieke coördinatie, uitlegbaarheid en verantwoord gebruik van spraaktechnologie. De juridische borging moet integraal deel uitmaken van de model-lifecycle (intake, training, validatie, publicatie, monitoring en hertraining), zodat auditbaarheid en herleidbaarheid structureel geborgd zijn.

Samenwerking binnen het Nederlandse taalgebied

De verdere ontwikkeling en verankering van spraakherkenning en –synthese als onderdeel van de digitale taalinfrastructuur voor het Nederlands vraagt om nauwe samenwerking binnen het Nederlandse taalgebied. Een gezamenlijke aanpak kan bestaan uit het bundelen en ontsluiten van bestaande spraakcorpora en het ontwikkelen van gedeelde evaluatie- en benchmarkkaders voor spraakherkenning. Daarbij is het van belang dat datasets niet alleen variatie tussen Nederland en Vlaanderen omvatten, maar ook taalgebruik uit Suriname en het Caribisch deel van het Koninkrijk, zodat modellen representatief zijn voor de volledige Nederlandstalige gemeenschap.

Tot slot is afstemming op het niveau van juridische kaders van belang, met name waar het gaat om de toepassing en interpretatie van Europese regelgeving rond data, auteursrecht en privacy. Door deze inspanningen expliciet te positioneren binnen het bredere traject rond digitale taalinfrastructuur en AI, kan spraaktechnologie duurzaam worden verankerd als onderdeel van de digitale taalinfrastructuur voor het Nederlands.

2.2.3 Multimodale data

Multimodale data zijn data die bestaan uit verschillende soorten informatiebronnen of modaliteiten die samen worden gebruikt om een situatie, object of gebeurtenis te beschrijven. In plaats van één enkel type data, zoals alleen tekst of alleen audio, combineren multimodale data bijvoorbeeld tekst, afbeeldingen, audio en video. Elke van deze modaliteiten legt andere aspecten van de werkelijkheid vast en draagt bij aan een rijkere en genuanceerdere betekenisrepresentatie.

Vele toepassingen die we vandaag online zien in het AI-domein zijn gebaseerd op tekst als modaliteit. In toenemende mate zien we ook toepassingen die audio, afbeeldingen of video genereren. Dit is bijvoorbeeld bijzonder relevant in de mediasector, waarbij deze modaliteiten moeten worden aangemaakt op een manier die verankerd is in de lokale cultuur van het Nederlandse taalgebied.

Een andere interessante vorm van multimodale data zijn gebarentaalcorpora, bijvoorbeeld voor NGT (Nederlandse Gebarentaal) en VGT (Vlaamse Gebarentaal). Deze kunnen louter uit video bestaan, of bevatten soms een transcriptie of een vertaling naar het Nederlands. In het kader van automatische analyse, vertaling en synthese van gebarentaal zijn zulke vertaalde of geannoteerde corpora uiterst belangrijk. Daarnaast bestaan er ook sensoriële datasets van hand- en lichaamsbewegingen voor gebarentalen, waarbij de beweging in de driedimensionale ruimte geregistreerd wordt.

Huidige situatie

Momenteel ligt de focus vooral op tekst als dominante modaliteit. Videoverwerking vraagt nog steeds veel rekenkracht en videogeneratie is doorgaans beperkt tot kortstondige fragmenten. Daarnaast worden video-, audio- en afbeeldingsdata uit Vlaanderen en Nederland nauwelijks ontsloten. Bovendien zijn trainingsdata voor video, audio en beeld voornamelijk afkomstig uit niet-Nederlandstalige contexten, wat ertoe leidt dat de gegenereerde multimodale output onvoldoende aansluit bij de cultuur en leefwereld van het Nederlandse taalgebied.

Dit kan bijvoorbeeld ongunstig zijn bij gebruiksscenario's in de media, waarbij afbeeldingen of video moeten worden gegenereerd voor een Nederlandstalig publiek. Denk hierbij aan het genereren van straatbeelden: zaken zoals bewegwijzering zien er anders uit in de Verenigde Staten of China dan in Vlaanderen of Nederland. Als we ooit video willen genereren die verankerd is in Nederlandstalige culturele context, is het dus nodig om ook video- en afbeeldingsdata ter beschikking te stellen. Maar deze data huizen vooral bij mediabedrijven en in audiovisuele archieven, die in eerste instantie niet werden beheerd met generatieve-AI-gebruik in het achterhoofd.

Voor automatische gebarentaalherkenning, -synthese of -vertaling zijn geen goed functionerende modellen beschikbaar. Er zijn een aantal datasets, maar vaak ontbreekt vertaling of worden de gebaren uitgevoerd door horende tolken in plaats van doof geboren tolken (zoals meestal bij televisieprogramma's), wat volgens de dovengemeenschap leidt tot minder authentieke gebarentaal. Voor synthese van

gebarentalen worden vaak avatars gebruikt, waarbij technologie uit de gaming-industrie nuttig blijkt, mits voorzien van de juiste invoerparameters. Bij het INT zijn een aantal [gebarentaaldatasets](#) beschikbaar als output van het [SignON](#)-project.

Voorgestelde verbeteringen

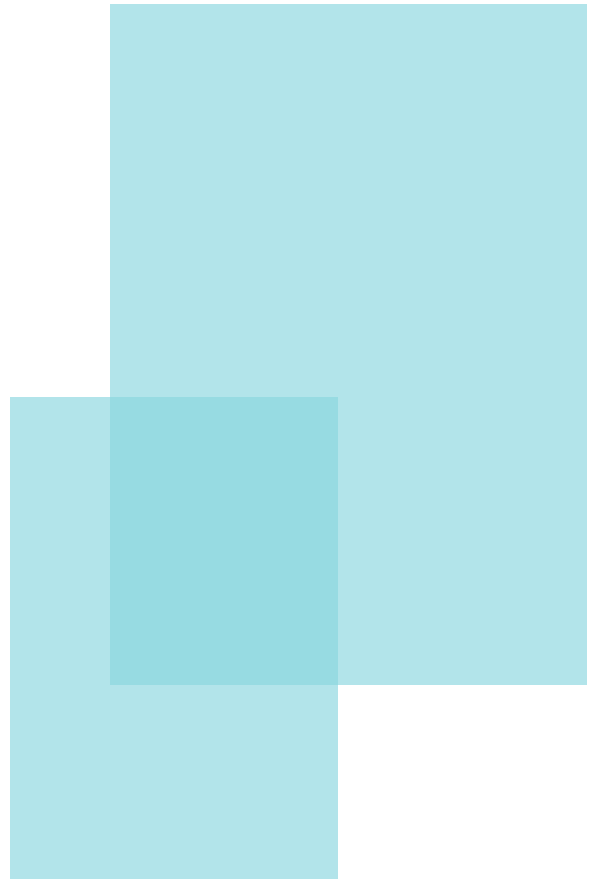
Dit alles benadrukt de nood aan initiatieven die het mogelijk maken om multimodale data systematisch te ontsluiten binnen de Nederlandstalige context. Een eerste noodzakelijke stap daarbij is het in kaart brengen van de multimodale data die al bestaan in het Nederlandse taalgebied. Daarbij moet duidelijk worden aangegeven wat de aard van de data is, waar deze data te vinden zijn, hoe ze kunnen worden gebruikt en welke juridische stappen nodig zijn om dat gebruik mogelijk te maken.

Mechanismen voor het delen van data vormen echter een bijzonder complex vraagstuk, zoals ook blijkt uit de dataspace die de voorbije jaren door de Europese Unie werden ondersteund. Een dataspace heeft als doel het delen van data tussen verschillende organisaties te faciliteren. De grootste uitdagingen liggen hierbij niet zozeer op het technisch vlak, maar vooral op juridisch gebied en vanuit het perspectief van businessmodellen. Vooral dat laatste is een heikel punt, omdat de motivatie meestal ontbreekt om data, waar de kerncompetentie van een organisatie vaak op berust, te delen met andere organisaties of onderzoeksinstituten. Zware investeringen in technologie voor het delen van data zijn dan vaak ook voorbarig en leveren niet altijd de gehoopte return op.

Daarom pleiten we ervoor om aan te sluiten bij de dataspace-aanpak voor het delen van multimodale data, maar tegelijk te focussen op de meest impactvolle en wendbare elementen van de onderliggende architectuur, met name de datacatalogus en het gebruik van modelcontracten. Via een datacatalogus kan in kaart worden gebracht wat beschikbaar is in het Nederlandse taalgebied. Via een modelcontract kunnen vervolgens snel de juridische maatregelen worden genomen om het delen van data mogelijk te maken.

Samenwerking binnen het Nederlandse taalgebied

Lopende projecten in Vlaanderen en Nederland hebben al veel inzichten opgeleverd met betrekking tot het delen van multimodale data. Hierbij denken we bijvoorbeeld aan het [VLAM](#)-programma in Vlaanderen of [GPT-NL](#) in Nederland. We stellen dan ook voor om deze kennis te verankeren door een multimodale datacatalogus op te stellen over de landsgrenzen en om aan deze catalogus ook modelcontracten te koppelen. Deze modelcontracten voor datadeling kunnen dan per dataset aangeven welke contracten moeten worden opgesteld en met welke juridische diensten moet worden gesproken over welke onderwerpen. Nederland en Vlaanderen kunnen ook samenwerken bij het opzetten van een groot gebarentaalcorpus voor beide gebarentalen (VGT en NGT) met bijhorende annotaties en vertalingen.



Aanbevolen acties binnen het Nederlandse taalgebied

Voor alle datatypen bevelen we het volgende aan:

- Inventarisatie van bestaande data voor tekst, audio, beeld, video en gebarentaal in Nederland en Vlaanderen met metadata zoals over aard, kwaliteit, locatie, gebruiksmogelijkheden en juridische randvoorwaarden;
- Samenwerking met media, archieven en omroepen om aanvullende multimodale datasets te verzamelen en te genereren;
- Opbouw en uitbreiding van hoogwaardige gebarentaalcorpora voor Vlaamse en Nederlandse Gebarentaal (VGT en NGT), inclusief annotaties en vertalingen, als noodzakelijke basis voor herkenning, synthese en vertaling;
- Opstellen van modelcontracten die per dataset aangeven welke juridische overeenkomsten nodig zijn en welke rollen en verantwoordelijkheden gelden tussen data-eigenaren, verwerkers en modelontwikkelaars;
- Ontwikkeling van gestandaardiseerde toegangsmechanismen en beveiligde processingomgevingen, zodat gevoelige of auteursrechtelijk beschermde data binnen een federatieve data-infrastructuur veilig kunnen worden gebruikt zonder centrale opslag. Zie ook pijler 2;
- Ontwikkeling van gedeelde benchmark- en evaluatiekaders voor taal- en AI-systemen, inclusief benchmarksets voor tekst- en spraaktechnologie, met expliciete aandacht voor domeinen, genres, taalvariatie, maatschappelijke contexten en inclusiviteit. Zie ook pijler 3;
- Implementatie van gedeeld *model-lifecyclebeheer* voor taal-, spraak- en multimodale modellen, inclusief logging, versiebeheer, validatie, monitoring, documentatie en periodieke hertraining, ter ondersteuning van reproduceerbaarheid, uitlegbaarheid en publieke verantwoording. Zie ook pijler 3;
- Duurzame opslag en gestructureerde documentatie van datasets en modellen, gekoppeld aan onderzoeks- en rekeninfrastructuren zoals CLARIN en nationale HPC-voorzieningen. Zie ook pijler 2;
- Structurele afstemming en koppeling van Vlaamse en Nederlandse onderzoeks- en data-infrastructuren, inclusief uitwisseling van expertise rond dataverzameling, governance, evaluatie en infrastructuurbeheer;
- Gezamenlijke afstemming van juridische kaders en interpretaties rond Europese regelgeving (auteursrecht, databankrecht, privacy), zodat landgrensoverschrijdend hergebruik van datasets en modellen binnen het Nederlandse taalgebied wordt vereenvoudigd;
- Verankering van deze acties binnen een duurzaam publiek overleg op digitale taalinfrastructuur en AI voor het Nederlands, waarbij schaal, transparantie en maatschappelijke verantwoording centraal staan.

2.3 Algemene data-aspecten

2.3.1 Een- of meertalige data?

Huidige situatie

Hoewel dit document gaat over AI voor het Nederlands, dient opgemerkt te worden dat het bij datasets, zowel voor tekst als voor spraak, kan gaan om zowel eentalige als meertalige datasets. Zulke datasets zorgen voor een grotere schaal en meer diversiteit, waardoor de modellen robuuster worden en beter omgaan met verschillende stijlen en domeinen. Dit leidt tot generalisatie over talen heen, waarbij talen met veel data de prestaties kunnen verbeteren voor talen met minder data, zoals het Nederlands. In de praktijk betekent dit dat het Nederlands wel deel uitmaakt van de meeste meertalige corpora, maar vaak als middelgrote taal, dus met minder data dan bijvoorbeeld Engels, Duits en Frans. Veel van die corpora zijn gebaseerd op [Common Crawl](#).

Voorgestelde verbeteringen

Op basis van recente resultaten rond meertalige pre-training en *instruction-tuning* (verfijning op instructie-antwoordparen) wordt een aantal gerichte ingrepen voorgesteld om de prestaties voor het Nederlands te verbeteren. Een eerste belangrijke maatregel is het hanteren van een meertalige mix van trainingsdata waarin het Nederlands een gegarandeerd en voldoende tokenbudget (capaciteit voor een specifieke taak) krijgt, aangezien kwaliteitsdegradatie vooral optreedt wanneer talen onvoldoende representatie hebben in de trainingsdata. Daarnaast wordt voorgesteld om een compacte maar strategische injectie van Nederlandstalige *instruction-tuning* voorbeelden toe te voegen, omdat onderzoek aantoont dat zelfs beperkte meertalige toevoegingen kunnen leiden tot sterk verbeterde generalisatie. Tot slot is de opbouw van een natuurlijke en diverse Nederlandstalige instructiedataset cruciaal, aangezien vertaalde of template-gebaseerde prompts doorgaans minder effectief zijn dan authentieke prompts die verschillende registers, taalvariatie en natuurlijke formuleringen weerspiegelen.

Samenwerking binnen het Nederlandse taalgebied

Het is belangrijk dat in de meertalige trainingsdata voldoende gevarieerd Nederlands zit, dus uit het volledige Nederlandse taalgebied. Dit geldt ook voor de instructiedataset.

Aanbevolen acties binnen het Nederlandse taalgebied

- **Zorgen voor een meertalige trainingsmix waarin het Nederlands een gegarandeerd en voldoende aandeel (tokenbudget) krijgt, zodat kwaliteitsverlies door ondervertegenwoordiging wordt vermeden;**
- **Opbouwen van een natuurlijke, diverse Nederlandstalige instructiedataset die verschillende registers, variatie en authentieke formuleringen weerspiegelt, in plaats van vertaalde of template-gebaseerde prompts.**

2.3.2 Juridische aspecten

Huidige situatie

In lopende projecten waarin basismodellen voor het Nederlandse taalgebied werden ontwikkeld wordt ofwel gebruikgemaakt van data uit het publieke domein, ofwel van data uit het private domein.

Daarbij kan worden aangenomen dat veel modellen die in het buitenland zijn ontwikkeld eveneens steunen op publiekelijk beschikbare data. In die context vormen vooral data uit het Nederlandse taalgebied die zich in het private domein bevinden een belangrijk, onderscheidend en waardevol element.

Tegelijk tonen verschillende projecten in Vlaanderen en Nederland aan dat het verkrijgen van toegang tot dergelijke datasets gepaard gaat met aanzienlijke juridische inspanningen en lange doorlooptijden. Dit is deels het gevolg van het ontbreken van een algemeen gekende en gestandaardiseerde werkwijze, en deels van de vele afstemmings- en onderhandelingsprocessen die nodig zijn met elke afzonderlijke organisatie die data wil aanleveren. De ervaringen uit GPT-NL, het VLAM-programma en CLARIN bevestigen deze observaties en laten zien dat de juridische uitdagingen rondom datatoegang en -gebruik niet alleen groot zijn, maar structureel vertragend werken voor de ontwikkeling van een taalmodel voor het Nederlands of de aanpassing (zogenaamde post-training of finetuning) ervan.

Dit geldt niet uitsluitend voor het private eigendom van data; ook binnen het publieke domein blijkt ontsluiting vaak problematisch. Publieke organisaties beschikken weliswaar over grote hoeveelheden waardevolle Nederlandstalige data, maar discussies over grondslag, proportionaliteit, doelbinding en hergebruik onder de AVG leiden tot terughoudendheid en langdurige interne afwegingen, waardoor datasets niet of slechts beperkt beschikbaar worden. Daarnaast is binnen GPT-NL en VLAM gebleken dat het ontbreken van gestandaardiseerde juridische instrumenten leidt tot een gefragmenteerd en tijdrovend onderhandelingsproces met iedere dataleverancier.

Iedere dataset vraagt een afzonderlijke juridische toets op privacy, auteursrecht, intellectueel eigendom, vertrouwelijkheid en rechten van derde partijen. Deze versnippering werkt dus sterk vertragend, vooral wanneer moet worden vastgesteld onder welke voorwaarden een partij zich mag terugtrekken en wat de implicaties daarvan zijn voor het model (bijvoorbeeld hertraining of dataverwijdering).

Voorgestelde verbeteringen

De ervaringen uit projecten waarin AI-modellen gebruik maakten van data afkomstig uit verschillende organisaties kunnen systematisch worden geïnventariseerd en ontsloten. Op deze manier kunnen ze voor andere partijen als referentie dienen bij het opstellen van contracten die toegang tot data mogelijk maken. Deze kennis kan worden aangeboden in de vorm van modelcontracten, uniforme bijdrageovereenkomsten met afspraken over het gebruik van de data, of sectorbrede richtlijnen, aangevuld met een duidelijke methode voor het opstellen en juridisch valideren ervan. Op die manier kan de doorlooptijd om toegang te krijgen tot relevante multimodale datasets aanzienlijk worden teruggebracht, evenals de bijbehorende juridische kosten. Daarnaast tonen de ervaringen uit GPT-NL en VLAM aan dat een duidelijke en gezamenlijk gedragen governancestructuur — zoals een contentraad waarin dataleveranciers inspraak hebben in gebruiksvoorwaarden, hergebruik en terugtrekkingsscenario's — cruciaal is om vertrouwen te waarborgen en duurzame datadeling mogelijk te maken.

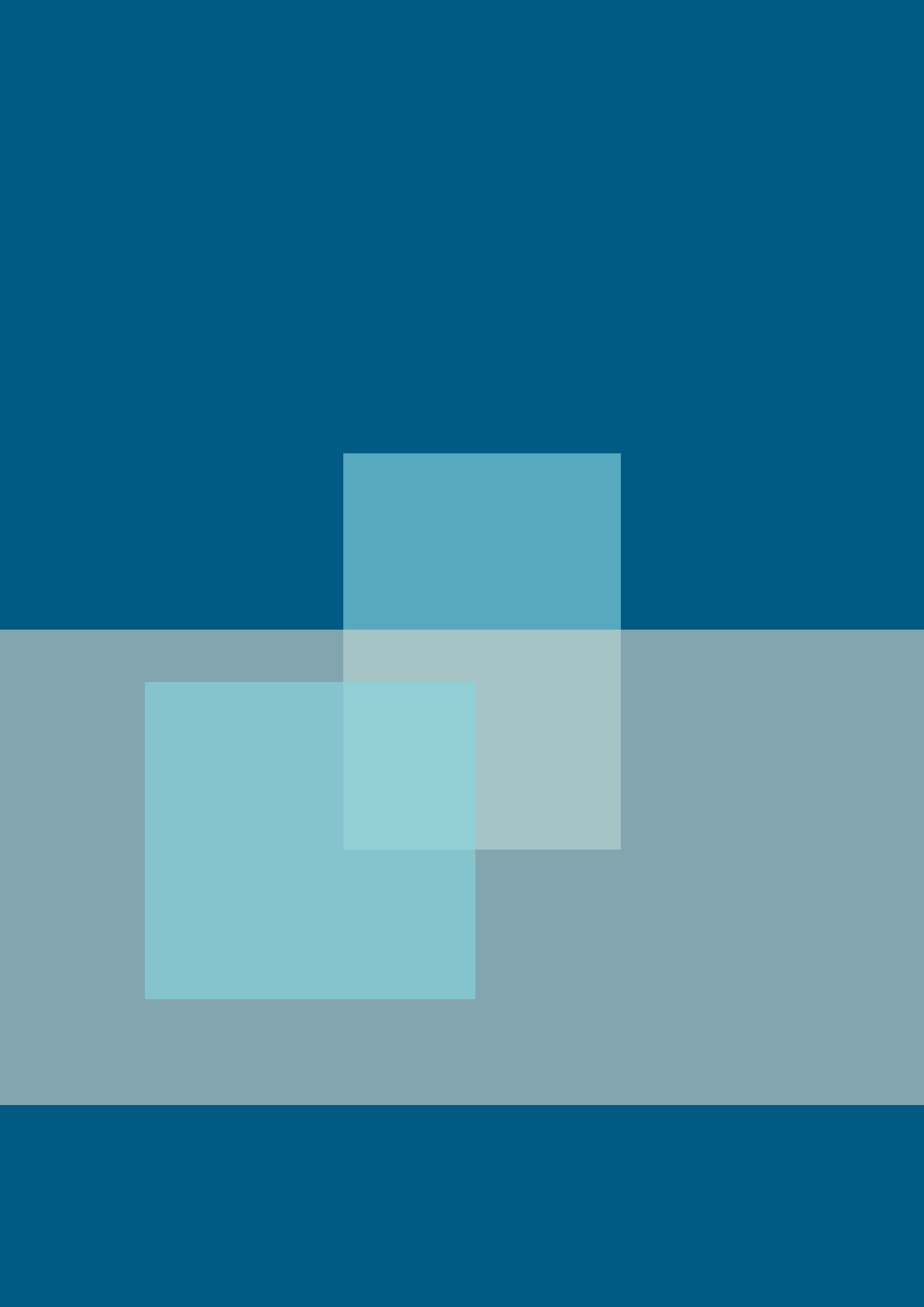
Samenwerking binnen het Nederlandse taalgebied

Deze initiatieven kunnen worden opgezet voor datasets die relevant zijn voor het Nederlandse taalgebied en die afkomstig zijn van organisaties binnen datzelfde taalgebied. Het bijbehorende werk kan worden uitgevoerd door één organisatie of door een consortium van organisaties dat over de landsgrenzen heen opereert. De modelcontracten kunnen vervolgens online worden aangeboden en actueel worden gehouden op een centraal platform dat geleidelijk naamsbekendheid en vertrouwen opbouwt binnen het Nederlandse taalgebied.

Het voordeel van toepassingen die gebruik maken van soevereine en onafhankelijke AI voor het Nederlands is ook dat het mogelijk wordt om via deze toepassingen grote hoeveelheden data te verzamelen. Vervolgens kunnen deze data weer gebruikt worden om de toepassingen te verbeteren. Op deze manier kan opwaartse spiraalwerking in gang worden gezet die het hele AI-ecosysteem ten goede komt.

Aanbevolen acties binnen het Nederlandse taalgebied

- In kaart brengen van de juridische ervaringen uit lopende en afgeronde AI- en dataprojecten (zoals GPT-NL en VLAM), met aandacht voor privacy, auteursrecht, intellectueel eigendom, vertrouwelijkheid, rechten van derde partijen en verschillen in wettelijke kaders tussen Nederland, Vlaanderen en Suriname (opgenomen in metadata);
- Ontwikkeling van uniforme modelcontracten en *content-contributor-agreements* die duidelijk vastleggen onder welke voorwaarden datasets kunnen worden aangeleverd, gebruikt, hergebruikt en eventueel teruggetrokken;
- Formulering van sectorbrede richtlijnen voor datadeling binnen het Nederlandse taalgebied, zodat publieke en private organisaties kunnen terugvallen op een herkenbaar en vertrouwd juridisch kader;
- Uitwerking van een duidelijke en herbruikbare methode voor het opstellen, toetsen en juridisch valideren van datacontracten, inclusief AVG-grondslagen, doelbinding en proportionaliteit;
- Oprichting van een gezamenlijke governancestructuur, zoals een contentraad, waarin dataleveranciers inspraak hebben in gebruiksvoorwaarden, hergebruik, toezicht en terugtrekkingsscenario's;
- Organisatie van de ontwikkeling en het beheer van juridische instrumenten via één organisatie of een consortium dat over de landsgrenzen heen opereert, zodat schaalvoordelen ontstaan en doorlooptijden voor datatoegang worden verkort;
- Investeren in naamsbekendheid, transparantie en communicatie rond het centrale platform en de governance-aanpak om vertrouwen bij dataleveranciers en modelontwikkelaars te versterken;
- Ontwikkeling van een gezamenlijk kader voor verantwoorde dataverzameling, waarin juridische, ethische en privacyvereisten vanaf het ontwerp worden geïntegreerd en dat richting geeft aan nieuwe dataverzamelingen binnen het Nederlandse taalgebied.



3 – Pijler 2: Infrastructuur

3.1 Inleiding

Deze pijler behandelt de infrastructuur die nodig is om data veilig, efficiënt en duurzaam te kunnen gebruiken voor AI-ontwikkeling. De technische passage over data in sectie 2.3.2 en die over rekeninfrastructuur hieronder benadrukken de beleidsmatige noodzaak van gezamenlijke investeringen en coördinatie. Het gaat niet om een technisch ontwerp, maar om de strategische keuzes die nodig zijn om versnippering tegen te gaan.

Over federatie en coördinatie

In dit actieplan wordt onderscheid gemaakt tussen federatieve infrastructuur en centrale coördinatie. Met een federatieve infrastructuur bedoelen we dat data, modellen en rekenvoorzieningen niet noodzakelijk centraal worden opgeslagen of beheerd, maar bij verschillende organisaties kunnen blijven. Samenwerking wordt mogelijk gemaakt via gedeelde standaarden, afspraken en toegangsmechanismen. Tegelijk vraagt een dergelijke federatieve aanpak om gezamenlijke afstemming en centrale coördinatie. Die centrale coördinatie betreft niet het eigenaarschap van data of infrastructuur, maar het afstemmen van prioriteiten, het ontwikkelen van gemeenschappelijke kaders, het faciliteren van samenwerking en het bewaken van publieke belangen. Centrale coördinatie en federatieve infrastructuur zijn daarmee complementaire onderdelen van hetzelfde ecosysteem.

3.2 Rekeninfrastructuur

Huidige situatie

Het trainen van basismodellen vergt een aanzienlijke hoeveelheid rekenkracht, waarbij vooral de pre-trainingsfase bijzonder veeleisend is. De post-trainingsfase is veel minder intensief. Cloudinfrastructuur biedt het voordeel dat ze snel en eenvoudig toegankelijk is, maar gaat in essentie gepaard met een verloren kost. Door de overheid ter beschikking gestelde infrastructuur

is toegankelijk via een competitief systeem, waardoor het niet gegarandeerd is dat men voor een specifieke trainingstaak effectief toegang krijgt. Door administratieve procedures is de doorlooptijd bovendien erg lang, gezien de snelheid waarmee het veld wereldwijd evolueert.

Op Europees niveau is infrastructuur momenteel nog in opbouw. In Vlaanderen werd recent een investering goedgekeurd voor een [AI-Antenna](#), waarvan de bijhorende infrastructuur eveneens nog in ontwikkeling is. Daarnaast heeft Vlaanderen toegang tot de [LUMI-supercomputer](#) en kreeg het [Vlaams Supercomputer Centrum](#) recent een uitbreiding van de rekencapaciteit. In Nederland is de beschikbaarheid van grootschalige AI-rekeninfrastructuur eveneens beperkt en gefragmenteerd. [SURF](#) beheert nationale HPC-voorzieningen (zoals [Snellius](#)), maar deze zijn sterk concurrerend, voornamelijk academisch georiënteerd en niet structureel ingericht op grootschalige LLM-training. Toegang verloopt via calls en allocatieprocedures, wat de inzet voor tijdkritische ontwikkelingen bemoeilijkt. Commerciële niet-Europese cloudpartijen worden daarom in de praktijk vaak ingezet, met bijbehorende kosten en afhankelijkheden van niet-Europese aanbieders als gevolg. Initiatieven zoals de Nederlandse AI-fabriek zijn veelbelovend, maar bevinden zich grotendeels nog in de plan- en opbouwfase.

Naast het eenmalig trainen van een AI-model is ook de dagelijkse inzet ervan sterk afhankelijk van beschikbare rekenkracht. Elke keer dat een gebruiker een vraag stelt aan een taalmodel, moet het model die vraag in realtime verwerken en een antwoord genereren. Dit vraagt continu om krachtige en gespecialiseerde computerinfrastructuur, zeker bij grotere modellen en bij breed gebruik door publieke of private organisaties. In tegenstelling tot training is deze rekenvraag permanent en schaalafhankelijk: hoe meer gebruikers en toepassingen, hoe groter de

benodigde capaciteit en kosten. Zonder structurele toegang tot voldoende, betrouwbare en betaalbare rekeninfrastructuur komt het grootschalig en soeverein inzetten van AI-modellen in het gedrang, met risico op afhankelijkheid van buitenlandse commerciële cloudproviders.

In Nederland en Vlaanderen is op dit moment slechts in beperkte mate structurele capaciteit beschikbaar voor het hosten en uitvoeren van grote AI-modellen. De bestaande nationale onderzoeksinfrastructuur en HPC-voorzieningen zijn primair ingericht op wetenschappelijk rekenen en incidenteel gebruik, en minder geschikt voor continue, schaalbare en bedrijfsmatige AI-toepassingen met veel gelijktijdige gebruikers. Ook rijkscloudvoorzieningen, zoals VLAM in Nederland (dat niets te maken heeft met het Vlaamse VLAM-programma), de rijksbrede digitale werkomgeving voor ambtenaren, zijn vooral bedoeld voor veilige kantoor- en applicatiediensten en niet voor grootschalige AI-rekenkracht. Daarnaast is VLAM ook niet voor de gehele overheid beschikbaar: de ambitie is rijksbreed, maar vooralsnog voorlopig beperkt tot de ministeries. De gemeenten vallen dus buiten scope. In de praktijk wordt hosting en uitvoering van grote taalmodellen daarom vaak ondergebracht bij internationale, veelal niet-Europese cloudproviders. Hoewel in Europa en Nederland en Vlaanderen wordt geïnvesteerd in soevereine cloud- en AI-infrastructuur (onder meer via [EuroHPC](#) en nationale initiatieven), bevindt deze capaciteit zich grotendeels nog in een opbouwfase en is zij nog niet breed beschikbaar voor langdurige, productiegerichte uitvoering op grote schaal. Er zijn echter wel degelijk aanbieders: [De Dutch Cloud Community \(DCC\)](#) stelt dat Nederland beschikt over een sterke cloud- en datacenterbasis en tot de best verbonden digitale infrastructures van Europa behoort met hoogwaardige datacenters, ervaren cloud- en hostingbedrijven en ruime kennis van grootschalige IT-operaties. Tegelijkertijd constateert zij dat deze capaciteit onvoldoende wordt benut voor strategische doeleinden zoals AI-rekeninfrastructuur. Hoewel Nederland deze infrastructuur in principe samen met de eigen cloudsector zou kunnen opzetten, gebeurt dit volgens de DCC te beperkt. Dit komt doordat het beleid versnipperd is, investeringen vooral ad hoc plaatsvinden en AI-infrastructuur vaak wordt benaderd als een operationele cloudkeuze in plaats van als strategische, publieke infrastructuur.

Voorgestelde verbeteringen

Evoluties in de Verenigde Staten en China tonen aan dat investeringen in rekeninfrastructuur noodzakelijk zijn om competitief te worden in de AI-economie van de nabije toekomst. Naar Europese normen zijn deze investeringen ronduit gigantisch: het gaat om bedragen van honderden tot zelfs duizenden miljarden euro. In de Verenigde Staten leeft de vrees dat deze investeringen mogelijk te hoog oplopen en wordt er gesproken over een AI-infrastructuurbubbel. Tegelijkertijd benadrukt dit dat de wereldeconomie ervan overtuigd is dat substantiële investeringen in AI onontbeerlijk zijn om een competitieve positie uit te bouwen. Europa dreigt echter achtergesteld te raken in deze investeringen wat betreft het beschikbaar stellen van de nodige infrastructuur. Aangezien het niet duidelijk is wat de juiste en optimale hoeveelheid investering hierin is, lijkt het wel nuttig om een analyse te maken van de nodige rekeninfrastructuur voor de komende jaren. In dit licht vormen Europese AI-fabrieken een belangrijk strategisch instrument om investeringen in rekenkracht te bundelen, versnippering tegen te gaan en gerichte toegang te bieden tot rekenkracht voor strategische usecases, zoals de ontwikkeling van soevereine taalmodellen. Een expliciete koppeling tussen nationale en regionale initiatieven en deze AI-fabrieken is daarbij essentieel.

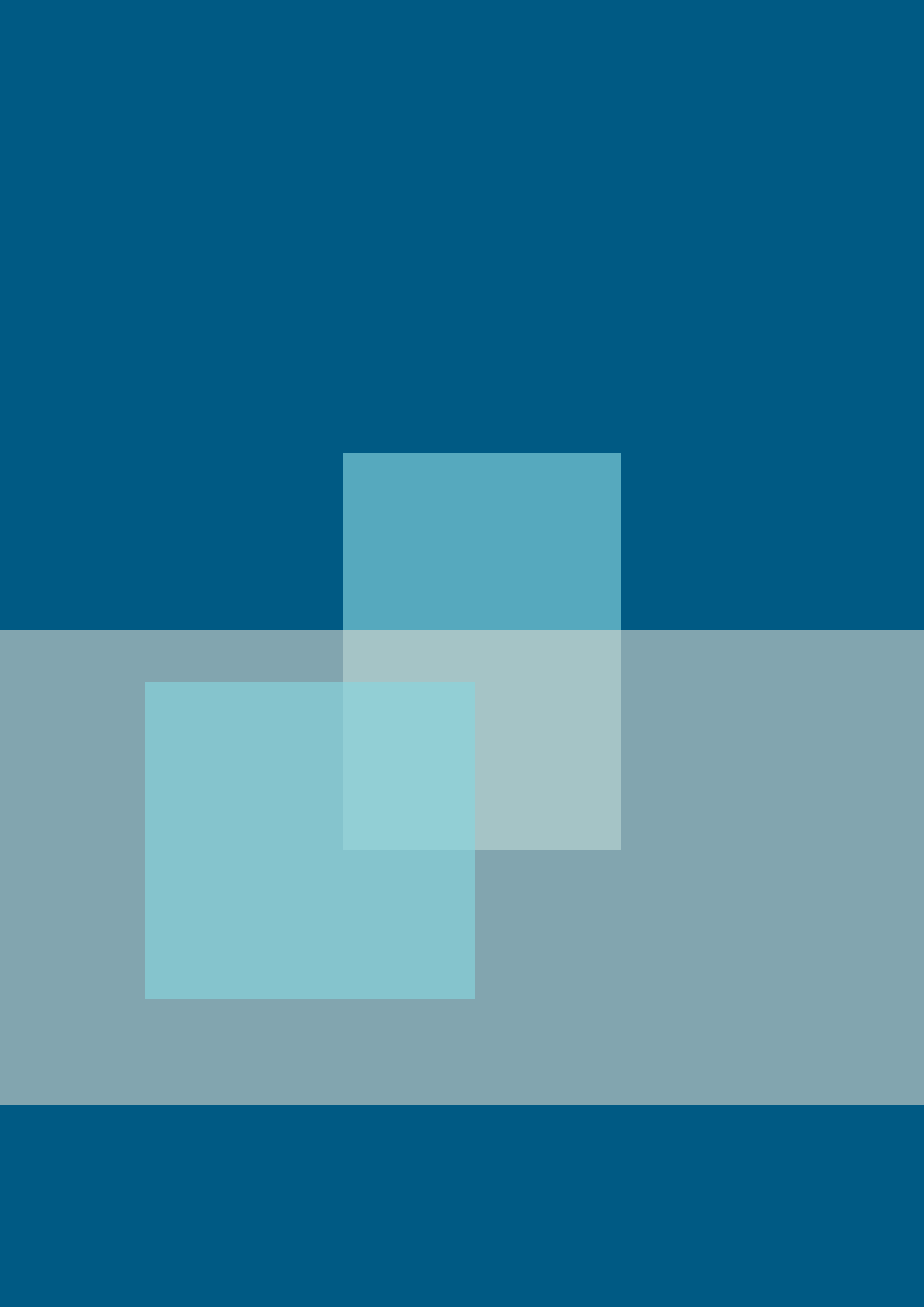
Samenwerking binnen het Nederlandse taalgebied

Omdat de investeringen in rekenkracht hoog zijn, lijkt het zinvol dat Vlaanderen en Nederland samenwerken om te onderzoeken hoe gezamenlijke investeringen kunnen worden opgezet waarbij budgetten worden gebundeld om infrastructuur uit te bouwen die gezamenlijk kan worden gebruikt. Een analyse die dit in kaart brengt en periodiek opvolgt hoe de infrastructuurnoden evolueren en hoe het aanbod daarop is afgestemd of moet worden bijgestuurd, lijkt daarbij bijzonder waardevol. Het zou dan ook aangewezen zijn om zo'n oefening gezamenlijk te ondernemen.

Deze samenwerking kan bovendien worden gepositioneerd in relatie tot Europese AI-fabrieken zoals de AI-fabriek in Groningen, waarbij gezamenlijk kan worden aangeklopt om noden, usecases (zoals AI voor het Nederlands) en investeringen te laten aansluiten bij bestaande of toekomstige AI-fabrieken. Dit vergroot de kans op structurele toegang tot Europese infrastructuur en cofinanciering.

Aanbevolen acties binnen het Nederlandse taalgebied

- Uitvoering van een gezamenlijke analyse van de rekeninfrastructuurnoden voor AI in Vlaanderen en Nederland voor de komende jaren, met specifieke aandacht voor training en grootschalige, continue uitvoering van taalmodellen;
- Opzetten van gezamenlijke investeringen in rekeninfrastructuur door het bundelen van Vlaamse en Nederlandse budgetten, om schaalvoordelen te realiseren en gedeelde infrastructuur mogelijk te maken;
- Structurele monitoring en bijsturing van vraag en aanbod van rekenkracht, zodat infrastructuurontwikkeling blijft aansluiten bij evoluerende noden en usecases;
- Afstemming en koppeling van nationale en regionale initiatieven rond rekeninfrastructuur aan Europese AI-fabrieken om versnippering te vermijden en strategische investeringen te versterken;
- Gezamenlijke positionering van AI-toepassingen voor het Nederlands (zoals soevereine taalmodellen) als expliciete usecases binnen Europese AI-fabrieken onder meer richting de AI-fabriek in Groningen;
- Vergroting van de kans op structurele toegang tot Europese infrastructuur en cofinanciering door Vlaanderen en Nederland gezamenlijk te laten optreden in Europese trajecten;
- Vermindering van afhankelijkheid van niet-Europese commerciële cloudproviders door inzet op gezamenlijke, soevereine en publiek aangestuurde rekeninfrastructuur voor AI-toepassingen.



4 – Pijler 3: AI-ontwikkeling

4.1 Inleiding

Deze pijler gaat in op de ontwikkeling, evaluatie en verbetering van AI-modellen voor het Nederlands. De meer technische onderdelen over benchmarking, training en representativiteit zijn bedoeld als onderbouwing van de beleidskeuzes die nodig zijn om kwaliteit, transparantie en maatschappelijke verankering te waarborgen. De pijler formuleert geen onderzoeksprogramma, maar benoemt de voorwaarden voor een duurzame ontwikkelcyclus.

4.2 Benchmarks

Huidige situatie

Bestaande benchmarks, bijvoorbeeld van [EuroEval](#), lijken aan te tonen dat grote commerciële en opensourcemodelen (zoals [GPT-5](#), [Claude-3](#), [Gemini-3](#), enz.) al een redelijk hoog niveau van taalbegrip en culturele competentie voor het Nederlands bereiken. Dit staat in contrast met de huidige specifiek Nederlandstalige modellen (zoals [Fietje](#), [GEITje](#), en [ChocoLlama](#)) die minder accuraat zijn. Dit kan komen door de kleinere omvang van die modellen, maar de goede benchmarkprestaties van de commerciële modellen lijken vooral een gevolg te zijn van het feit dat de gebruikte benchmarks niet voldoen: ze zijn automatisch vertaald uit het Engels waardoor het een oppervlakkige vertaling van Amerikaanse taalkundige en culturele contexten blijft in plaats van authentiek Nederlands. Een deel

ervan kan bovendien gelekt zijn in het materiaal voor de training van grote modellen, of de benchmarks zijn inhoudelijk beperkt en onwetenschappelijk. Hierdoor meten ze vaak niet wat ze zouden moeten meten: authentiek taalgebruik en culturele kennis. Automatische vertaling van benchmarkdata vanuit het Engels kan bovendien leiden tot culturele bias, grotere verengelsing en taalerosie. Bij de inzet van taalmodellen in onderwijs, overheid en bedrijfsleven kan dit leiden tot een verhoogd risico op miscommunicatie en foutieve of juridisch/maatschappelijk onbetrouwbare output. Een bestaande Nederlandstalige benchmark, de Dutch Model Benchmark ([DuMB](#)), bestaat uit een aantal NLP-taken waar wel Nederlandstalige data voor beschikbaar zijn. Maar deze data komen voor een groot deel uit oudere bronnen, zoals de manuele annotaties van het Corpus Gesproken Nederlands en het [SoNaR](#)-corpus.

Voorgestelde verbeteringen

Een robuuste Nederlandstalige benchmark moet minstens drie aspecten testen: (1) taalkundige competentie (woordenschat, syntaxis, idiomatische expressies, etc.); (2) culturele kennis, ook in contrast met andere culturen (vakanties, feesten, tv en film, etc.); (3) en spontaan taalgebruik in interactie met correcte interpretatie van context (dialoog). Zo'n benchmark kan niet alleen dienen voor betrouwbare evaluatie, maar (daarna) ook voor hoogwaardige finetuning van bestaande modellen. We stellen voor de vooruitgang in bestaande initiatieven rond benchmarks (bijvoorbeeld [Flembench](#) en de benchmarking voor het Belgisch-Nederlands in project [Narrate](#)) te combineren in een ruimer project voor benchmarking van taalvariatie en cultuur. We willen waar mogelijk aansluiten bij bestaande, veelgebruikte benchmarks, zoals [BLiMP](#) voor formeel-taalkundige competentie en [SuperGLUE](#) voor functioneel-taalkundige competentie. Van BLiMP bestaat een Nederlandse versie, maar voor zover wij weten nog niet van SuperGLUE. Hier liggen kansen voor de ontwikkeling van een benchmark van hoge kwaliteit (niet via automatische vertaling).

Met een dergelijk benchmarking-initiatief laten we toe AI-modellen systematisch te testen op taalbegrip, maar ook op hun vermogen om Nederlandstalige contexten te herkennen. We bieden hiermee ook basismateriaal aan dat bruikbaar is voor bedrijven, onderzoekers, onderwijsinstellingen en de overheid, en dat bijdraagt aan technologische onafhankelijkheid van Engelstalige AI-standaarden.

Samenwerking binnen het Nederlandse taalgebied

Pas door taalmodellen te pre-trainen of te post-trainen met tekstverzamelingen uit verschillende delen van het Nederlandse taalgebied, kunnen de modellen historisch, taalkundig en cultureel leren differentiëren. De datasets die nu bijvoorbeeld in DuMB gebruikt worden, bestaan voor een groot deel uit projecten die in het kader van het gezamenlijke Nederlands-Vlaamse programma Spraak- en Taaltechnologische Essentiële Voorzieningen in het Nederlands ([STEVIN, 2005-2011](#)) gemaakt zijn. Een update en uitbreiding hiervan is na 20 jaar noodzakelijk. We adviseren ook dat onze benchmarkacties bij voorkeur aansluiting zoeken bij internationale initiatieven, zoals [ALT-EDIC](#) / [CLARIN](#) / [OpenEuroLLM](#), zodat de data voor het Nederlands ook bij die projecten lokaal worden aangemaakt en duplicatie wordt vermeden.

Aanbevolen acties binnen het Nederlandse taalgebied

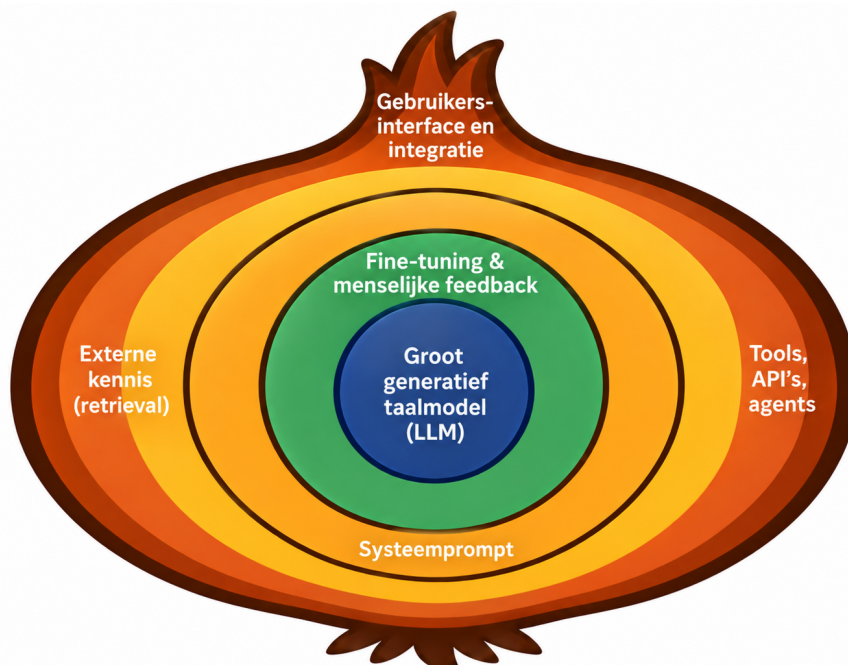
- **In kaart brengen van bestaande benchmarks om duplicatie te vermijden en bouwstenen te identificeren;**
- **Ontwikkeling van testcategorieën voor taalkundige competentie, culturele kennis en spontaan taalgebruik in dialoog;**
- **Creatie van authentieke, niet-automatisch vertaalde datasets, inclusief een Nederlandse SuperGLUE-versie;**
- **Validatie van de benchmarks door experts uit het Nederlandse taalgebied om kwaliteit en representativiteit te waarborgen;**
- **Publicatie van de benchmarks en integratie bij CLARIN en ALT-EDIC voor internationale aansluiting.**

4.3 Pre-training of post-training?

Huidige situatie

Figuur 1 geeft aan dat de kern van een tekstuele AI-toepassing altijd een groot taalmodel is dat teksten kan genereren in één of meerdere talen. Dit model kan verder getraind worden met meer data (bijvoorbeeld in een specifiek domein of taalvariëteit), en gefinetuned voor een taak. De meeste AI-assistenten die we kennen, zoals [ChatGPT](#) en [Claude](#), zijn gefinetuned voor het voeren van conversaties en het opvolgen van instructies. In dit proces wordt ook menselijke feedback ingezet. Daarnaast zijn ze uitgebreid met allerlei tooling, bijvoorbeeld om externe informatiebronnen aan te roepen.

Bijlage A bevat een overzicht van bestaande taalmodellen voor het Nederlands met informatie over hun training. Allemaal hebben ze een meertalig/Engelstalig model als startpunt. Het enige bestaande model uit onze regio dat voor het Nederlands is gepretraind is GPT-NL (zie hieronder). Dit model bevat naast een meerderheid van Engelstalige teksten voornamelijk teksten uit Nederland en deels uit Vlaanderen. Dat het model ook op Engels is getraind heeft te maken met volume: zonder het toevoegen van Engelstalig materiaal zou er niet voldoende materiaal zijn geweest om een bruikbaar LLM te trainen. Er is tot nu toe slechts beperkte finetuning op GPT-NL uitgevoerd, en het is nog niet uitgebreid met de andere lagen van de figuur.



Figuur 1: Een AI-systeem bestaande uit verschillende lagen.

Voorgestelde verbeteringen

Het pre-traineren van grote taalmodellen is een grote onderneming. GPT-NL is hiervan het bewijs en ook andere Europese initiatieven in Europa (zoals [TrustLLM](#)) worstelen met de pre-trainingfase. De uitdaging zit in drie aspecten: (1) de enorme hoeveelheden data van hoge kwaliteit die nodig zijn voor een goed taalmodel; (2) de rekenkracht die nodig is voor langdurige tijd om het model op grote hoeveelheden data te trainen; (3) de ethische en juridische aspecten van het verzamelen van data met goede kwaliteit. Waar uitdaging (2) nog wel praktisch overwonnen kan worden, is (3) een principieel probleem waar niet eenvoudig overheen is te stappen.

We denken daarom dat het een wijze keuze is om voort te bouwen op al bestaande basismodellen, initiatieven te volgen die zich richten op pre-training (zoals GPT-NL), en meteen de stap te zetten naar de volgende fasen van taalmodeltraining. Deze fasen bestaan uit verder pre-traineren (bijvoorbeeld het toevoegen van meer domeinspecifieke Nederlandstalige data aan een meertalig model) en finetuning (bijvoorbeeld het trainen op Nederlandstalige dialogen). Vooral in dit verder pre-traineren en finetunen zit het potentieel tot het opwaarderen van basismodellen en het afstemmen op bepaalde taken.

Samenwerking binnen het Nederlandse taalgebied

Een eerste stap zou kunnen zijn om in kaart te brengen hoeveel taalvariatie er in de bestaande Nederlandstalige modellen zit en dit verder in balans te brengen via verder pre-traineren. Ook in de finetuningfase kunnen we hier aandacht aan besteden via specifiek trainen op taalvariëteiten die ondervertegenwoordigd zijn. Buiten het inzetten op het behandelen van modellen die tekst genereren, is het ook relevant om te kijken naar modellen die in staat zijn om andere modaliteiten dan tekst te genereren, bijvoorbeeld afbeeldingen, audio en video. Zoals eerder vermeld zijn deze ook bijzonder relevant voor het Nederlandse taalgebied.

Aanbevolen acties binnen het Nederlandse taalgebied

- **In kaart brengen hoeveel geografische, etnische en sociale taalvariatie aanwezig is in bestaande Nederlandstalige modellen en identificeren van ondervertegenwoordiging;**
- **Pre-training uitvoeren met Nederlandstalige data die geografische variatie dekken, inclusief informele domeinen en diverse taalregisters;**
- **Ontwikkeling van instructiemodellen via supervised finetuning die etnische, sociale en situationele taalvariatie integreren;**
- **Toepassing van DPO/RLHF-alternatieven om modellen te post-traineren op ondervertegenwoordigde variëteiten en bias te voorkomen. Continued pre-training (kosteneffectief bestaande modellen verder trainen op nieuwe data zonder volledige pre-training te herhalen);**
- **Post-training uitvoeren op bestaande basismodellen om deze beter af te stemmen op taken die waarde toevoegen voor het Nederlandse taalgebied;**
- **Testen van de ontwikkelde modellen met de benchmarks uit pijler 3 om representativiteit voor de hele taalgemeenschap te waarborgen.**

4.4 Representativiteit, ethiek, soevereiniteit

Huidige situatie

Veel huidige AI-modellen zijn getraind op grootschalige, maar vaak ongebalanceerde datasets, die veelal automatisch van het web worden verzameld. Dergelijke data weerspiegelen slechts gedeeltelijk de taalkundige, sociale en culturele diversiteit van de taalgemeenschappen waarin deze modellen worden ingezet. Voor het ontwikkelen van robuuste en betrouwbare AI-systemen is het nochtans essentieel dat trainingsdata voldoende variatie bevatten en representatief zijn voor de volledige bevolking die de taal spreekt, inclusief verschillen in register, regio en context. Vanuit de Taalunie wordt het belang van taalvariatie ook benadrukt, zie bijvoorbeeld het rapport van de Adviescommissie Taalvariatie uit 2019. Ook in het [meerjarenbeleidsplan](#) van de Taalunie en in het [advies van de Raad voor de Nederlandse Taal en Letteren](#) over digitale taalinfrastructuur en AI wordt aangegeven dat rekening houden met taalvariatie noodzakelijk is om modellen te kunnen ontwikkelen die representatief zijn voor de hele taalgemeenschap.

Daarnaast zijn ethische overwegingen van fundamenteel belang. AI-technologie kan een grote impact hebben op de samenleving en de democratische processen. Daarom moeten AI-systemen worden ontworpen, ingezet en geëvalueerd in overeenstemming met ethische normen en publieke waarden die relevant zijn binnen het Nederlandse taalgebied. Dit omvat onder meer aandacht voor representativiteit, transparantie, verantwoording en het vermijden van systematische vertekeningen of uitsluiting.

In het kader van AI speelt ook digitale soevereiniteit een cruciale rol, met name binnen Europa. Tegen de achtergrond van geopolitieke spanningen en strategische afhankelijkheden groeit de nood om kerncompetenties rond AI veilig te stellen en autonoom te ontwikkelen. Dit vergt niet alleen technologische capaciteiten, maar ook zeggenschap over data, modellen en infrastructuur.

Projecten als GPT-NL in Nederland en VLAM in Vlaanderen spelen een belangrijke rol in het

aankaarten van deze bekommernissen in het Nederlandse taalgebied. OpenEuroLLM en [LLMs4EU](#) zijn dan weer projecten die op AI-basismodellen en de evaluatie ervan werken binnen de EU. Meer over deze projecten staat beschreven in [bijlage B](#).

Samenwerking binnen het Nederlandse taalgebied

De uitdagingen rond representativiteit, ethiek en digitale soevereiniteit kunnen niet op een geïsoleerde manier worden aangepakt, maar vragen om nauwe en structurele samenwerking binnen het volledige Nederlandse taalgebied. Taalvariatie, maatschappelijke diversiteit en culturele context overstijgen nationale grenzen, waardoor een gezamenlijke Vlaamse en Nederlandse aanpak noodzakelijk is om AI-systemen te ontwikkelen die werkelijk representatief zijn voor de hele taalgemeenschap. Door data, expertise en infrastructuur te bundelen kan vermeden worden dat modellen eenzijdig worden gevoed met beperkte of scheefgetrokken datasets, wat het risico op bias en uitsluiting vergroot.

Initiatieven zoals GPT-NL in Nederland en het VLAM-programma in Vlaanderen tonen aan dat er aan beide zijden van de grens sterke aandacht bestaat voor publieke waarden zoals transparantie, inclusiviteit en publieke coördinatie. Deze projecten vervullen een voorbeeldfunctie door representativiteit, ethische reflectie en soevereiniteit expliciet centraal te stellen in de ontwikkeling van taalmodellen. Door deze inspanningen beter op elkaar af te stemmen en structureel te verbinden, kan een gezamenlijke leercurve worden gerealiseerd waarin datasets, methodologieën en governance-modellen elkaar wederzijds versterken.

Samenwerking binnen het Nederlandse taalgebied biedt bovendien de mogelijkheid om gezamenlijke keuzes te maken over ethische kaders en ontwerpprincipes voor AI-systemen. Een gedeeld debat over ethiek, publieke waarden en maatschappelijke impact helpt om te komen tot breed gedragen normen voor *ethics-by-design* (verankering van ethische aspecten in het ontwerpproces van de AI-systemen). Deze normen kunnen vervolgens consequent worden toegepast in dataverzameling, modelontwikkeling en ingebruikname. Dit vergroot niet alleen het maatschappelijk vertrouwen, maar versterkt ook de democratische legitimiteit van AI-toepassingen.

Tot slot is samenwerking cruciaal voor het bevorderen van digitale soevereiniteit. Door gezamenlijk op te treden richting Europese initiatieven en projecten rond basismodellen en infrastructuur, zoals OpenEuroLLM en LLMs4EU, kunnen Vlaanderen en Nederland hun belangen beter verankeren binnen het Europese AI-ecosysteem. Een gecoördineerde aanpak maakt het mogelijk om via publieke en Europese samenspraak te investeren in data, modellen en infrastructuur die zijn afgestemd op het Nederlands en op de waarden van de Nederlandstalige samenleving, om zo duurzame alternatieven te bieden voor niet-Europese, commerciële AI-oplossingen.

Aanbevolen acties binnen het Nederlandse taalgebied

- **Opname van voldoende variatie en representativiteit in de data om ervoor te zorgen dat bepaalde groepen taalgebruikers niet uitgesloten worden van gebruik van de resulterende modellen en de applicaties die daaruit voortvloeien;**
- **Verzamelen en ontsluiten van geografisch diverse data (tekst, spraak en non-verbale communicatie) uit Nederland, Vlaanderen, Suriname en het Caribisch deel van het Koninkrijk, zodat regionale variatie, meertalige contexten en culturele referenties uit het volledige Nederlandse taalgebied systematisch worden gerepresenteerd;**
- **Uitbreiding van datasets met etnische en meertalige taalvariatie, inclusief bijdragen van tweedetaalleerders en mensen met uiteenlopende taalachtergronden, om inclusieve AI-toepassingen te ondersteunen;**
- **Waarborging van sociale representativiteit in data door expliciet materiaal op te nemen van verschillende leeftijdsgroepen, genders, opleidingsniveaus, sociale contexten en van mensen met communicatieve beperkingen of gezondheidsgerelateerde taalvariatie;**
- **Integratie van situationele en functionele taalvariatie door verschillende registers en gebruikscontexten van het Nederlands op te nemen, zoals formele en informele tekst, sociale-mediacommunicatie, chat- en berichtendiensten, spontane spraak en voorbereide spreekvormen;**
- **Systematische bewaking en verbetering van representativiteit en inclusiviteit in datasets en modellen, zodat geen specifieke groepen structureel worden uitgesloten van het gebruik en de voordelen van AI-toepassingen;**
- **Uitwerking van een soevereine AI-infrastructuur voor het Nederlandse taalgebied, die maximaal gebruikmaakt van Europese hardware, software, clouddiensten en open technologieën, met als doel strategische afhankelijkheden te verminderen en publieke coördinatie, transparantie en duurzaamheid te versterken;**
- **Organisatie van een structureel ethisch debat rond AI binnen het Nederlandse taalgebied, en expliciete verankering van ethische aspecten zoals inclusiviteit, transparantie, verantwoordingsplicht en mensenrechten in het ontwerp, en de ontwikkeling en de inzet van AI-systemen (ethics-by-design).**

5 – Pijler 4: Valorisatiepotentieel

5.1 Inleiding

Deze pijler verkent waar investeringen in AI voor het Nederlands maatschappelijke en economische meerwaarde kunnen opleveren in Nederland, Vlaanderen en Suriname. De beschreven sectoren dienen als illustratie van mogelijke toepassingsdomeinen en zijn geen uitputtende prioriteitenlijst. De bedoeling is om zichtbaar te maken hoe een gedeelde AI-infrastructuur kan bijdragen aan publieke dienstverlening, innovatie en concurrentiekracht.

en soevereiniteit, zeker in het huidige geopolitieke klimaat. Het Vlaamse VLAM-programma, opgericht in 2025, ondersteunt mediabedrijven in AI-innovatie, onder meer via spraakmodellen voor het Nederlands, AI-ontsluiting van multimodale media-archieven en generatieve toepassingen voor tekst, audio en video. Door deze werven te versterken via samenwerking tussen onderzoek, overheid en industrie in het Nederlandse taalgebied, dragen we bij aan toekomstige welvaart, democratie en digitale soevereiniteit.

5.2 Media en entertainment

De media- en entertainmentsector in het Nederlandse taalgebied staat onder sterke druk door buitenlandse concurrentie. Sociale media en streamingplatformen hebben de zakenmodellen de voorbije jaren grondig hertekend, en sinds de opkomst van ChatGPT is daar een nieuwe golf van AI-veranderingen bijgekomen. Fenomenen zoals zero-clickinternet, waarbij informatie rechtstreeks aan de gebruiker wordt geleverd zonder door te klikken op websites, zetten de inkomsten van mediaspelers verder onder druk. Ook generatieve AI voor tekst, audio en video zal de komende jaren een toenemende impact hebben op media- en entertainmentbedrijven.

Deze sector leent zich bij uitstek voor samenwerking over landgrenzen heen, aangezien meerdere grote mediabedrijven actief zijn in zowel Vlaanderen als Nederland, zoals [DPG Media](#), [Roularta](#) en [Mediahuis](#). Ook de publieke omroepen [VRT](#) en [NOS](#) werken al nauw samen. Het ondersteunen van deze bedrijven in hun zoektocht naar nieuwe toegevoegde waarde is cruciaal. Verlies van marktaandeel aan buitenlandse spelers zoals [TikTok](#), [Google](#) of [OpenAI](#) zou niet alleen economische schade veroorzaken, maar ook risico's meebrengen voor democratie

5.3 Onderwijs

In het onderwijs kan AI voor verschillende doeleinden worden ingezet. Een daarvan is het verlagen van de werkdruk bij leraren. Bij het geven van instructie of het nakijken kan AI ingezet worden om diagnoses te stellen en beslissingen te nemen. Met *learning-analytics* kunnen patronen in data van leerlingen en hun context worden geanalyseerd. Het resultaat daarvan kan zichtbaar worden gemaakt in dashboards die leraren informatie bieden over de voortgang en leerresultaten van leerlingen. Deze automatische analyses kunnen meer gedetailleerd zijn dan de gangbare toetsen en kunnen meer informatie opleveren ten behoeve van diagnostiek en remediëring met minder tijdsinvestering. Leraren blijven verantwoordelijk voor het nemen van besluiten, maar worden daarbij ondersteund door AI.

Een andere manier om AI in te zetten in het onderwijs is door leerlingen zelfstandig te laten oefenen met AI-toepassingen. Deze kunnen automatisch fouten detecteren en verbeteren, en desgewenst uitleg geven. Zo kunnen leerlingen hun schrijfvaardigheid verbeteren door direct feedback te krijgen op grammatica, zinsstructuur, spelling, stijl en woordkeuze. Learning-analytics kunnen ingebed worden in systemen voor gepersonaliseerd

leren, geautomatiseerde feedback en adaptieve leermiddelen die het taalleerproces versterken en leerlingen kansen bieden om ook thuis te oefenen en begeleiding te krijgen. Dit is vooral belangrijk bij vaardigheden waarvoor niet voldoende tijd is in de klas zoals hardop lezen en spreekvaardigheid. Met AI-toepassingen zoals een leesbegeleider kunnen leerlingen op hun eigen tempo, zelfstandig en onbeperkt oefenen met hardop lezen, en ook feedback krijgen buiten de klas. Uitgevers, toetsontwikkelaars en de hele onderwijssector zouden gebaat zijn bij toegang tot een soeverein en verantwoord AI-ecosysteem voor het Nederlands. Data die nu terechtkomen in toepassingen van Big Tech zouden dan terug kunnen vloeien in het ecosysteem zodat ze verder kunnen worden gebruikt om de AI en de toepassingen te verbeteren.

5.4 Publieke sector

Zowel de Nederlandse als de Vlaamse overheid beklemtonen het belang van duidelijke en begrijpelijke taal in overheidscommunicatie. Het Nederlandse regeerakkoord 2026 is zelfs beschikbaar in een versie in begrijpelijke taal. AI-toepassingen voor het Nederlands kunnen worden ingezet om de leesbaarheid van bestaande teksten te meten. Indien nodig kunnen deze teksten automatisch worden aangepast aan een gewenst leesbaarheidsniveau. Daarnaast bestaan er schrijfhulptools die auteurs tijdens het schrijfproces ondersteunen bij het creëren van begrijpelijke teksten. Dit zijn AI-toepassingen voor het Nederlands die specifiek relevant zijn voor zowel Nederland als Vlaanderen.

Verder is er in Vlaanderen de ambitie om de dienstverlening van de overheid te vereenvoudigen via AI. De overheid biedt een groot aantal diensten en hier wegwijs in worden is geen eenvoudige taak voor de burger. Hier kan AI bij helpen, maar de gebruikte modellen moeten dan wel goed omgaan met geschreven en gesproken Nederlands en de lokale cultuur goed begrijpen.

5.5 Onderzoek

In wetenschappelijk onderzoek zijn er heel wat vakgebieden die Nederlandstalige data als studieobject hebben, zoals bijvoorbeeld geschiedenis, literatuurwetenschappen en erfgoedwetenschappen. Maar ook psychologen, logopedisten en pedagogen bestuderen casussen waarin het Nederlands vaak een belangrijke rol speelt. Het gebruik van AI-modellen en toepassingen die toegespitst zijn op het Nederlands kan voor deze wetenschappen nieuwe onderzoeksmogelijkheden openleggen. We denken hierbij bijvoorbeeld aan het automatisch samenvatten en grondig analyseren van de inhoud van grote hoeveelheden tekst- en audiobestanden op een grootschalige en toch gedetailleerde manier die anders niet mogelijk is.

5.6 Juridische sector

De juridische sector is erg tekstgeoriënteerd en bij uitstek een sector die gewend is om land- en taalspecifiek te werken. Juristen in Nederland werken met wetten, artikelen en jurisprudentie uit Nederland en juristen in Vlaanderen met diezelfde typen documenten uit België. Deze teksten zijn allemaal Nederlandstalig. Waar zoek- en classificatietechnologie nog wel redelijk eenvoudig taalafhankelijk gemaakt kan worden, of aangepast kan worden naar het Nederlands, is dit lastiger voor generatieve taken zoals samenvattings- en vraag-antwoord-systemen. Eerder onderzoek heeft aangetoond dat met name de kleinere modellen moeite hebben met domeinspecifiek taalgebruik, en dat de grote, commerciële, meertalige modellen het meest effectief zijn.

[Wolters Kluwer](#), de uitgeverij van de grootste Nederlandse juridische zoekmachine, wil graag meer gebruikmaken van generatieve modellen, maar maakt zich ook zorgen over de betrouwbaarheid van generatieve dialoogsystemen. Hier liggen kansen voor het ontwikkelen van Nederlandstalige modellen die verder getraind zijn op juridische taal, en bovendien niet te creatief zijn, maar dicht bij de oorspronkelijke juridische tekst blijven. Deze modellen zouden bijvoorbeeld voor samenvattingen niet te abstractief (nieuwe tekst genererend) zijn,

maar op cruciale taken extractief (combineert belangrijke tekstsegmenten uit het brondocument) te werk moeten gaan. Voor dialoogsystemen zouden de modellen altijd gebruik maken van Retrieval-Augmented Generation (RAG), een zoektechnologie die de antwoorden voorziet van externe, actuele databronnen.

Verder is het groeiende aantal bezwaarschriften binnen de Nederlandse overheid een urgent digitaliseringsprobleem. Burgers en advocaten gebruiken steeds vaker generatieve AI om bezwaren te produceren, waardoor deze langer, juridischer en talrijker worden. Uitvoeringsorganisaties raken hierdoor overbelast: behandeltermijnen staan onder druk en medewerkers besteden veel tijd aan het handmatig verwerken van omvangrijke documenten. Omdat ambtenaren zelf vaak geen generatieve AI mogen gebruiken, ontstaat een asymmetrie tussen burger en overheid die de werkdruk verhoogt en de uitvoeringscapaciteit aantast.

5.7 Veiligheid

De asymmetrie in de juridische sector tussen burger en overheid in het gebruik van generatieve AI verruimt zich naar de veiligheidsketen, waar organisaties zoals justitie, politie en toezichthouders werken met grote volumes complexe Nederlandstalige documenten. Huidige generatieve AI-modellen zijn doorgaans onvoldoende afgestemd op juridische en bestuurlijke teksten, wat leidt tot onnauwkeurigheden en risico's in besluitvorming. Bovendien zijn veel modellen buitenlands, niet transparant en moeilijk verenigbaar met Nederlandse eisen rond governance en gegevensbescherming.

De ontwikkeling van soevereine Nederlandstalige AI-oplossingen biedt daarom groot potentieel. Zulke modellen maken het mogelijk om documentstromen efficiënter te analyseren, juridische consistentie te borgen, en datacontrole te garanderen. Dit versterkt de uitvoeringskracht van publieke organisaties, verkleint de asymmetrie tussen burger en overheid, en vermindert de afhankelijkheid van buitenlandse technologie, terwijl het tegelijk een sterk nationaal innovatie-ecosysteem stimuleert.

5.8 Zorg

Voor de zorgsector biedt Nederlandstalige AI veel kansen, die variëren van het verlagen van de werkdruk en het efficiënt inrichten van processen tot het stellen van diagnoses en het faciliteren van therapie. AI-taalmodellen kunnen worden ingezet om administratieve taken te verlichten, dossiers te doorgronden en samenvattingen te maken. Daarbij kunnen de modellen gesproken verslagen in tekst omzetten en druk tijdig signaleren, bijvoorbeeld van de instroom op de Spoedeisende Hulp/ Spoedgevallendienst.

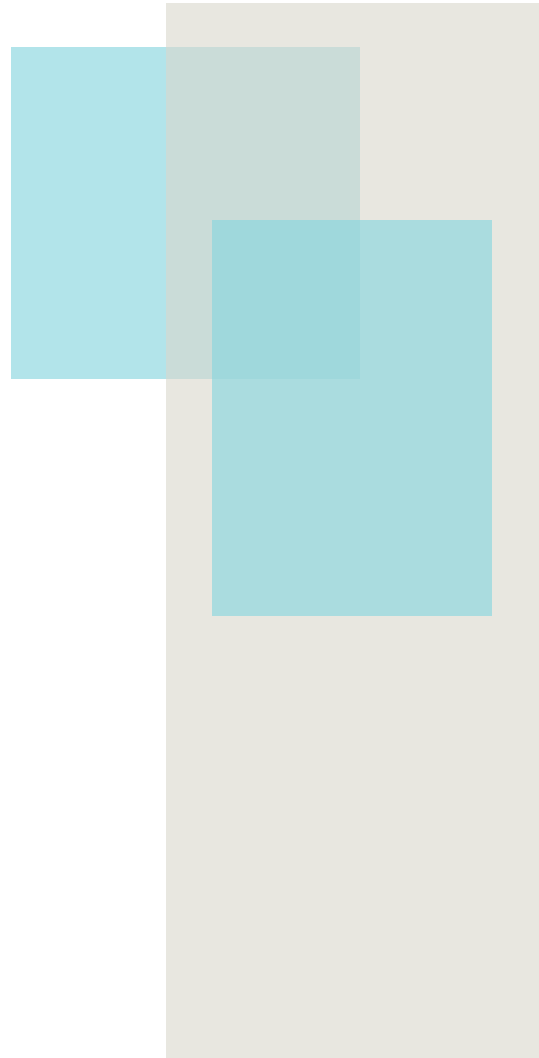
Daarnaast kan Nederlandstalige AI ondersteuning bieden bij anamnese, triage en monitoring op afstand door symptomen te checken en afwijkingen tijdig op te sporen. Dit kan bijvoorbeeld worden geïmplementeerd via AI-chatbots en zorgrobots die ondersteuning bieden aan patiënten en zorgverleners. Nederlandstalige AI kan ook complexe medische brieven of ontslagverslagen automatisch omzetten in begrijpelijk Nederlands om zo de zelfredzaamheid van patiënten te vergroten.

Verder kan Nederlandstalige AI bijdragen aan gepersonaliseerde geneeskunde, diagnostiek en therapie. Door enorme hoeveelheden data te combineren (bijvoorbeeld genetica, leefstijl en medische geschiedenis) kan AI de optimale behandeling helpen voorspellen. Ook is het mogelijk om Nederlandstalige AI in te zetten bij de diagnose van ziektes die zich vroegtijdig in spraak manifesteren, zoals Parkinson, Alzheimer en Chronic Obstructive Pulmonary Disease (COPD). In het verlengde daarvan kan Nederlandstalige AI ook worden ingezet voor het bieden van tijdrovende, individuele (taal- en spraak)therapie die anders niet mogelijk zou zijn. Voor patiënten die beperkt mobiel zijn kan dit ook op afstand.

5.9 Speerpuntclusters en topsectoren

Om de valorisatie te borgen in de verschillende relevante sectoren zijn samenwerkingen met organisaties in Vlaanderen en Nederland nodig die het innovatiebeleid in de verschillende sectoren coördineren. In Vlaanderen zijn dat de speerpuntclusters, die voornamelijk waken op het financieren en opvolgen van strategische projecten in verschillende domeinen. Een relevant speerpuntcluster die bijvoorbeeld kan aansluiten bij het AI voor het Nederlands-initiatief is de ENT.A-media-en-entertainment-speerpuntcluster. Deze werd opgericht begin 2026 en telt onder zijn ledenbestand een groot aantal mediabedrijven die zowel leverancier van relevante data kunnen zijn als afnemers van de componenten die worden ontwikkeld in het AI voor het Nederlands-actieplan.

Een gelijkaardige werking bestaat in Nederland onder de vorm van het topsectorenbeleid. Het identificeert tien sectoren waar Nederland internationaal sterk in is en waar publiek-private samenwerking wordt gestimuleerd om innovatie, valorisatie en economische impact te versnellen. Het gaat bijvoorbeeld om Agri & Food, Tuinbouw & Uitgangsmaterialen en Creatieve Industrie. Binnen elke topsector werken bedrijven, kennisinstellingen en overheid aan gezamenlijke innovatieagenda's met aandacht voor digitalisering, AI, data, menselijk kapitaal en toepassingsgerichte opschaling. Verschillende topsectoren kunnen baat hebben bij een sterkere werking rond AI voor het Nederlands. We zien er dan ook baat bij om hen nauw te betrekken.



6 – Pijler 5: Kennisdeling, governance en programmawerking

6.1 Inleiding

Deze pijler beschrijft hoe samenwerking, kennisdeling en governance structureel kunnen worden georganiseerd. De voorstellen focussen op de inrichting van een duurzaam samenwerkingsmodel en vormen een aanzet voor een toekomstige uitvoeringsagenda. De nadruk ligt op afstemming, coördinatie en continuïteit.

6.2 Kenniscentrum

Om de digitale autonomie van het Nederlandse taalgebied duurzaam te versterken, is het gezien de snelle ontwikkelingen in AI noodzakelijk om te voorzien in een structureel en permanent Vlaams-Nederlands samenwerkingsverband. Dit samenwerkingsverband fungeert als kenniscentrum dat op gecoördineerde wijze data, infrastructuur, modelontwikkeling en toepassingen van generatieve AI voor het Nederlands stimuleert en aanstuurt. Waar mogelijk zoekt dit kenniscentrum aansluiting bij Europese initiatieven, zoals ALT-EDIC, de [Language Data Space](#) of CLARIN, om internationale afstemming en schaalvergroting te realiseren. Het kenniscentrum heeft als kerndoelen het uitwerken van een langetermijnstrategie, het bepalen van prioriteiten, het vermijden van overlap tussen initiatieven en het versterken van de internationale positionering en samenwerking rond AI binnen het Nederlandse taalgebied. De werking wordt ondersteund door een brede adviesraad waarin overheden, universiteiten, erfgoedinstellingen, industrie, en juridische en ethische experts vertegenwoordigd zijn.

Het kenniscentrum neemt daarbij een aantal centrale taken op zich. Het fungeert als één loket voor informatie over standaardlicenties en -contracten, juridische vraagstukken, en compliance met onder meer AVG en de AI Act. Daarnaast staat het in voor het opzetten, coördineren en exploiteren van een federatieve data- en rekeninfrastructuur, inclusief passend toegangsbeheer en de archivering van modellen. Een bijkomende kerntaak is de wetenschappelijke benchmarking en evaluatie van AI-modellen. Dit gebeurt zowel generiek als sectorspecifiek, met aandacht voor domeinen zoals overheid, zorg, media, onderwijs en veiligheid. Op basis hiervan ondersteunt het kenniscentrum de selectie en prioritering van relevante usecases. Verder speelt kennisverspreiding een expliciete rol, onder meer via transparante publieke rapportering over gemaakte keuzes, prioriteiten en evaluatieresultaten. De financiering van het kenniscentrum is structureel van aard. Voor infrastructuur en modelontwikkeling wordt voorzien in een meerjarige (vijf jaar, hernieuwbaar) basisfinanciering. De ontwikkeling en uitrol van concrete usecases gebeurt via cofinanciering, met bijdragen van fondsen, industrie en Europese programma's.

De oprichting van dit kenniscentrum is strategisch noodzakelijk. Bij gebrek aan actie dreigt een toenemende afhankelijkheid van niet-Europese AI-modellen. Dit zijn voornamelijk Amerikaanse en Chinese modellen, met onvoldoende aandacht voor taalvariatie en culturele context binnen het Nederlandse taalgebied. Dit kan leiden tot taalerosie, verlies aan maatschappelijke en economische meerwaarde, en een structurele achterstand in innovatie en competitiviteit.

Het kenniscentrum heeft dus twee typen taken: enerzijds beleids-, organisatie-, coördinatie- en strategietaken, en anderzijds kennis-, uitvoerings-, en infrastructuurtaken. De al bestaande Nederlands-Vlaamse-Surinaamse tandem Taalunie (voor het eerste type taken) - Instituut voor de Nederlandse Taal (voor het tweede type) kan dit kenniscentrum verder uitbouwen. De coördinerende rol van de Taalunie betreft strategische afstemming, kennisdeling en governance. Data, infrastructuur en modellen kunnen daarbij onder beheer blijven van de deelnemende organisaties binnen het federatieve ecosysteem. Dit is in lijn met het advies van de Raad voor de Nederlandse Taal en Letteren “Digitale taalinfrastructuur en AI: Aanbevelingen voor een verantwoord en duurzaam beleid”. De Taalunie is een taalbeleidsorganisatie en het INT voert al taken van het Taalunieverdrag uit voor de digitale taalinfrastructuur voor het Nederlands, heeft expertise op het gebied van AI en werkt al samen met nationale en internationale partners uit overheid, academia en bedrijfsleven.

6.3 Governance

Als voorbeeld voor de governance-structuur zou het model van het Vlaams-Nederlandse stimuleringsprogramma STEVIN kunnen dienen. STEVIN werd gefinancierd door de Nederlandse overheid via [NWO](#), de Nederlandse ministeries van [EZ](#) en [OCW](#), en door de Vlaamse Overheid, vertegenwoordigd door het Departement [EWI](#), het [IWT-Vlaanderen](#) en het [FWO-Vlaanderen](#). STEVIN werd gecoördineerd en financieel beheerd door de Taalunie. Deze organisaties vormden samen met een aantal experts uit het veld het TST-bestuur dat toekeek op het STEVIN-programma. NWO en SenterNovem (nu [RVO](#)) vormden samen het STEVIN-programmabureau, dat belast was met de uitvoering. STEVIN was een Vlaams-Nederlands stimuleringsprogramma. Voor dit actieplan wordt in de governance- en adviesstructuren gestreefd naar vertegenwoordiging van organisaties en expertise uit het gehele Nederlandse taalgebied, inclusief Suriname en het Caribisch deel van het Koninkrijk.

6.4 Programmawerking

Het actieplan legt een aantal pijlers vast die van belang zijn voor de verdere uitbouw van de AI-innovatiecompetentie in het Nederlandse taalgebied. Het bevat echter nog te weinig detail om onmiddellijk geïmplementeerd te kunnen worden. Daarom is het uitwerken van een overkoepelende uitvoeringsagenda, die de vijf verschillende pijlers met elkaar verbindt, een logische volgende stap. Eens deze uitvoeringsagenda is uitgewerkt, kunnen de verschillende taken concreet worden uitgevoerd via projecten. Deze projecten kunnen een mix van financieringsstromen combineren, zowel van overheden en bedrijven uit het Nederlandse taalgebied als van Europese programma's. De uitvoering kan gebeuren door samenwerkingen tussen bedrijven, overheidsinstellingen, onderzoekscentra en organisaties uit het maatschappelijk middenveld. Ook het nader uitwerken van een duidelijke governance voor de programmawerking vormt daarbij een belangrijke vervolgstap.

7 Conclusie

De snelle opkomst van artificiële intelligentie maakt duidelijk dat taal geen randvoorwaarde is, maar een strategische kerncomponent van digitale autonomie, economische weerbaarheid en democratische veerkracht. Voor het Nederlandse taalgebied brengt dit zowel kansen als risico's met zich mee. Zonder gerichte en gecoördineerde actie dreigt structurele afhankelijkheid van buitenlandse AI-modellen en infrastructuren die onvoldoende rekening houden met de taalvariatie, cultuur en publieke waarden van Nederland, Vlaanderen en Suriname. Dit actieplan toont aan dat het fundament voor een alternatief traject aanwezig is, maar dat structurele versterking noodzakelijk is.

Het actieplan maakt duidelijk dat de knelpunten rond AI voor het Nederlands niet los van elkaar staan. Beperkingen in data-ontsluiting, versnipperde infrastructuur, projectmatige modelontwikkeling, moeizame valorisatie en fragmentarische kennisdeling versterken elkaar en verhinderen de uitbouw van een duurzaam ecosysteem. Daarom pleit dit actieplan voor een samenhangende benadering langs vijf pijlers: data, infrastructuur, AI-ontwikkeling, valorisatie en kennisdeling. Niet één afzonderlijke ingreep, maar de onderlinge afstemming tussen deze pijlers bepaalt het succes.

Tegelijk kiest dit actieplan bewust voor realisme. Het formuleert de inhoudelijke bouwstenen voor een gezamenlijke routekaart, maar bevat geen uitgewerkte uitvoeringsagenda met prioritering, planning en financiering. De volgende fase bestaat uit het uitwerken van een gedragen uitvoeringsagenda die prioriteiten, timing en instrumenten vastlegt en die vertaald kan worden naar concrete projecten. Die projecten kunnen worden uitgevoerd in publiek-private samenwerkingen en gefinancierd via een combinatie van nationale, regionale en Europese middelen.

Cruciaal daarbij is het opzetten van een robuuste governancestructuur en een permanent kenniscentrum dat strategie, uitvoering en expertisebundeling op zich neemt. Alleen met structurele organisatie, transparante afspraken en publieke coördinatie kan AI voor het Nederlands uitgroeien tot een duurzame voorziening in plaats van een reeks losse initiatieven. Aansluiting bij bestaande Nederlands-Vlaamse instituties en Europese infrastructuren vergroot daarbij zowel de slagkracht als de internationale positionering.

Dit actieplan biedt daarmee een gezamenlijk vertrekpunt voor Nederland, Vlaanderen en Suriname om AI voor het Nederlands strategisch te verankeren. Door nu te investeren in samenwerking, infrastructuur en publieke waarden wordt de basis gelegd voor een toekomst waarin Nederlandstalige AI niet alleen technisch krachtig is, maar ook maatschappelijk verantwoord, cultureel verankerd en economisch waardevol.

Referenties

Adviescommissie Taalvariatie. (2019). *Visie op taalvariatie en taalvariatiebeleid*. <https://taalunie.org/publicaties/88/visie-op-taalvariatie-en-taalvariatiebeleid>

AI-Antenna. (2026). <https://www.be-aifa.eu/en/>

ALT-EDIC. (n.d.). https://www.alt-edic.eu/nl_nl/

ChatGPT. (n.d.). <https://chatgpt.com/>

CHOREC. (2009). <https://taalmaterialen.ivdnt.org/download/tstc-chorec-spraakcorpus/CLARIAH-NL>. (n.d.). <https://clariah.nl/>

CLARIAH Media Suite. (n.d.). <https://mediasuite.clariah.nl/>

CLARIN ERIC. (n.d.). <https://www.clarin.eu/>

Claude. (n.d.). <https://claude.ai/login>

Claude-3. (2024). <https://www.anthropic.com/news/claude-3-family>

Common Crawl. (n.d.). <https://commoncrawl.org/>

Corpus Gesproken Nederlands. (2018) <https://taalmaterialen.ivdnt.org/download/tstc-corpus-gesproken-nederlands/>

De Vries, W. (n.d.). *DUMB benchmark repository*. <https://github.com/wietsedv/dumb>

De Vries, W., Wieling, M., & Nissim, M. (2023). DUMB: A benchmark for smart evaluation of Dutch models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (pp. 7221–7241). Association for Computational Linguistics.

Dutch Cloud Community. (2025). *Digitale soevereiniteit*. <https://dutchcloudcommunity.nl/digitale-soevereiniteit/>

EuroEval. (n.d.). <https://euroeval.com/>

EuroHPC. (n.d.). https://www.eurohpc-ju.europa.eu/index_en

Flembench. (2025). <https://it3.ugent.be/projects/flembench-een-benchmarkinitiatief-voor-taal-en-cul/>

Gemini-3. (n.d.). <https://aistudio.google.com/models/gemini-3>

GPT-5. (n.d.). <https://openai.com/nl-NL/gpt-5/>

GPT-NL. (2026). <https://gpt-nl.nl/>

Halleweyn, E. D., Odijk, J., Teunissen, L. M., & Cucchiari, C. (2006). Dutch-Flemish HLT Programme STEVIN: Essential Speech and Language Technology Resources.

JASMIN. (2008) <https://taalmaterialen.ivdnt.org/download/tstc-jasmin-spraakcorpus/>

Language Data Space. (n.d.). https://language-data-space.ec.europa.eu/index_en

LLMs4EU. (2025). https://www.alt-edic.eu/nl_nl/projects/llms4eu/

LUMA. (n.d.). <https://lumi-supercomputer.eu/>

Meeus, M., Rathé, A., Remy, F., Delobelle, P., Decorte, J. J., & Demeester, T. (2024). Chocollama: Lessons learned from teaching llamas dutch. *arXiv preprint arXiv:2412.07633*.

Narrate. (n.d.). <https://project-narrate.eu/>

OpenEuroLLM. (n.d.). <https://openeurollm.eu/>

Peng, Y., Tian, J., Yan, B., Berrebbi, D., Chang, X., Li, X., ... & Watanabe, S. (2023, December). Reproducing whisper-style training using an open-source toolkit and publicly available data. In *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)* (pp. 1-8). IEEE.

Raad voor de Nederlandse Taal en Letteren. (2026). *Digitale taalinfrastructuur en AI: Aanbevelingen voor een verantwoord en duurzaam beleid*. <https://raadntl.org/>

SignON project. (2021). <https://signon-project.eu/>

SoNaR-corpus. (2015). <https://taalmaterialen.ivdnt.org/download/tstc-sonar-corpus/>

Suijkerbuijk, M., Prins, Z., De Heer Kloots, M., Zuidema, W., & Frank, S. L. (2025). BLiMP-NL: A corpus of Dutch minimal pairs and acceptability judgments for language model evaluation. *Computational Linguistics*, 51(4), 1267–1301. https://doi.org/10.1162/COLI_a_00559

SuperGLUE. (n.d.). <https://super.gluebenchmark.com/>

SURF. (n.d.). SANE: *Secure environment for analysing sensitive data*. <https://www.surf.nl/en/themes/research-infrastructure/sane-secureenvironment-for-analysing-sensitive-data>

Snellius. (n.d.). <https://www.surf.nl/diensten/rekenen/snellius-de-nationale-supercomputer>

Taalunie. (2025). *Meerjarenbeleidsplan 2025–2029*. <https://taalunie.org/publicaties/228/meerjarenbeleidsplan-2025-2029>

TrustLLM. (n.d.). <https://trustllm.eu/>

Vanroy, B. (2024). Geitje 7b ultra: A conversational model for dutch. *arXiv preprint arXiv:2412.04092*.

VLAM. (n.d.). <https://www.imec.be/nl/vlaamse-innovatiemotor/impactdomeinen/media-entertainment/vlam-programma>

Wang, A., Pruksachatkun, Y., Nangia, N., Singh, A., Michael, J., Hill, F., Levy, O., & Bowman, S. (2019). SuperGLUE: A stickier benchmark for general-purpose language understanding systems. In *Advances in Neural Information Processing Systems*, 32.

Warstadt, A., Parrish, A., Liu, H., Mohananey, A., Peng, W., Wang, S.-F., & Bowman, S. R. (2020). BLiMP: The benchmark of linguistic minimal pairs for English. *Transactions of the Association for Computational Linguistics*, 8, 377–392. https://doi.org/10.1162/tac1_a_00321

Bijlage A: Het trainen van een taalmodel

Ter opfrissing: het trainen van een taalmodel voor chat bestaat uit verschillende stadia. Om te beginnen is er een pre-training stadium, waarin het model aan de hand van ongestructureerde, lopende tekst leert om het volgende token te voorspellen. Zo leert het logische woordcombinaties, gangbare zinsbouw, en uiteindelijk ook coherentie over lange stukken tekst heen. Daarnaast leert het zo ook meer over de inhoud van de teksten. Dit basismodel kan dus enkel van een gegeven tekst de volgende woorden voorspellen, maar is nog niet geschikt om mee te chatten en is al zeker niet afgestemd om instructies te volgen als assistent.

Het is mogelijk om een ge-pre-trained model verder te pre-trainen (*continued pre-training*) met meer en andere ongelabelde data, bijvoorbeeld voor domeinadaptatie of het verminderen van bias. Om een chatfunctie te implementeren wordt gebruik gemaakt van zogenoemde *supervised finetuning* (SFT) waarbij de data wel minimaal gestructureerd zijn. De data in dit stadium zijn gestructureerde gesprekken (één of meerdere vraag-antwoordparen) die aan de hand van een chat-template dan tot gestructureerde tekst worden gemaakt.¹ Het model wordt dus nog steeds op lopende tekst getraind (het leert het volgende token voorspellen) maar doet dat wel op een vastgelegde structuur. Door het trainen leert het model dan dat na het eindigen van de “input-structuur” het op die input een antwoord moet voorzien. Deze fase is dus cruciaal om het model de gangbare structuur van een gesprek aan te leren.

Dan volgt het stadium van afstemming waarbij een taalmodel beter wordt afgestemd op onze voorkeuren (het wordt daarom ook wel *preference-tuning* genoemd). In het verleden werd vaak *reinforcement-learning* (het model leert beslissingen te nemen door interactie) gebruikt (zoals RLHF), maar tegenwoordig zijn er ook veel RL-vrije preference-tuningmethodes (zoals DPO).² Deze voorkeuren gaan meestal over het verbeteren van de kwaliteit van de antwoorden, maar dit kan worden opgevat onder verschillende aspecten: eerlijkheid en behulpzaamheid vergroten en mogelijkheid tot misbruik verkleinen (gaande van discriminerende taal tot het weigeren van informatie over onwettigheden). Daarnaast worden gelijkaardige algoritmes ook breder toegepast, bijvoorbeeld om betere code te genereren, sterker te worden in “tool use” waarbij een model toegang heeft tot verschillende gereedschappen zoals toegang tot het internet om het redeneervermogen te verfijnen, enzovoort.

Tot slot kan een vorm van aanpassing gedaan worden tijdens het draaien van het model (uitvoering/prompt-time). Het model wordt dan niet meer aangepast, maar krijgt mogelijk externe informatie mee (opgehaald uit een index van bijvoorbeeld recente documenten of data) en kan geïnstrueerd worden om een bepaalde taak uit te voeren of een specifiek uitvoerformaat te genereren.

Hoewel bovenstaande beschrijving lijkt te spreken over fijn afgegrensde stadia, wordt in de praktijk vaak gesproken over pre-training enerzijds (volgende token voorspellen) en post-training anderzijds (model aanpassen om beter instructies te volgen). Elke modelbouwer heeft hierbij een andere aanpak (zie bv. Llama 3³ en Qwen 3⁴), maar het algemene concept “eerst pre-trainen, daarna aanpassen aan instructies” blijft.

Voor multimodale modellen – zoals Vision Language Models (VLMs) of speech-enabled LLMs - worden in de laatste fase multimodale dataparen opgelijnd: het model wordt aangeleerd om een tekst te genereren uit een numerieke representatie van tekst, aangemaakt door een voorgetrainde beeld-of spraak-encoder.

¹ <https://huggingface.co/learn/llm-course/en/chapter11/2#common-template-formats>

² <https://www.jair.org/index.php/jair/article/view/17541/27180>

³ <https://arxiv.org/pdf/2407.21783>

⁴ <https://arxiv.org/pdf/2505.09388>

Tabel 1: Bestaande Nederlandse taalmodellen en hoe ze getraind zijn. Met ‘base’ verwijzen we naar verder pre-trainen en met ‘instruct’ naar vormen van post-trainen.

Model	Startpunt	Continued pre-training	Post-training
Tweety Dutch ⁵	Mistral v0.1	Ja	Nee
ChocoLlama ⁶	Llama 3 8B	Ja	Ja
Schaapje ⁷	IBM Granite 3.0 2B instruct	Ja	Ja
Fietje ⁸	Phi-2	Ja	Ja
GEITje ⁹	Mistral v0.1	Ja	Ja
Reynaerde	Mistral v0.3 instruct	Nee	Ja
GPT-NL	-	Nee	Deels

⁵ <https://huggingface.co/Tweeties/tweety-7b-dutch-v24a>; <https://arxiv.org/pdf/2408.04303>

⁶ <https://arxiv.org/pdf/2412.07633>; <https://huggingface.co/ChocoLlama>

⁷ <https://github.com/RobinSmits/Schaapje>

⁸ <https://github.com/BramVanroy/fietje-2>; <https://www.clinjournal.org/clinj/article/view/213>

⁹ <https://github.com/Rijgersberg/GEITje>

Bijlage B: Lopende projecten

Spraakdata

Binnen het Nederlandse taalgebied is de afgelopen jaren substantiële expertise opgebouwd op het vlak van spraakherkenning, vaak in nauwe samenwerking met de mediasector en het onderwijs. In Vlaanderen richt het NeLF-project (Next Level Flemish) zich op het ontwikkelen van inclusieve spraakherkenning die laagdrempelig inzetbaar is. Parallel hieraan bouwt het innovatieproject “Automatische Ondertiteling 2026–2027 (Auto2)” aan geautomatiseerde workflows voor ondertiteling voor de Vlaamse mediasector, waarin automatische spraakherkenning een centrale rol speelt. Binnen deze projecten is bijzondere aandacht voor dialectvariatie en niet-moedertaalsprekers. Door de inzet van grootschalige spraak- en ondertitelingsdatasets, afkomstig uit de mediasector, zijn sterke ASR-modellen ontwikkeld voor transcriptie en ondertiteling. Deze datasets kunnen echter niet publiekelijk worden gedeeld.

Innovatieve modelarchitecturen maken het mogelijk om zowel uit ruwe spraakdata als uit ondertitelingsdata te leren, wat heeft geleid tot prestaties die vaak vergelijkbaar zijn met of beter dan die van generieke meertalige modellen. Ook in Nederland vindt intensieve samenwerking plaats tussen onderzoek en mediasector. Binnen het HOSAN-project (Hoogwaardige Spraakherkenning voor al het Nederlands) wordt vanuit een publiek-maatschappelijke invalshoek gewerkt aan robuuste benchmarks voor de evaluatie van spraakherkenning in maatschappelijk relevante contexten, evenals aan modellen die adequaat omgaan met spontane spraak, taalvariatie en contextafhankelijke interpretatie. HOSAN positioneert zich expliciet als beheerder van de nationale model- en validatielaag, inclusief training, validatie, documentatie en lifecyclebeheer, terwijl implementatiepartijen toepassingen ontwikkelen bovenop deze modellen. Spraakherkenning wordt daarbij benaderd als een fundamentele bouwsteen voor bredere AI-toepassingen waarin verstaanbaarheid en toegankelijkheid centraal staan.

Binnen zowel HOSAN als NeLF is aangetoond dat hoogwaardige spraakherkenningsmodellen voor het Nederlands kunnen worden ontwikkeld op basis van beschikbare data, mits deze publiek worden beheerd. In Nederland zijn daarnaast belangrijke stappen gezet in de juridische en ethische verankering van datagebruik voor modeltraining, onder meer in samenwerking met mediapartners. Parallel hieraan wordt in samenwerking met SURF gewerkt aan een duurzame data- en rekeninfrastructuur om herbruikbaarheid, schaalbaarheid en publieke coördinatie op spraaktechnologie structureel te verankeren. In deze opzet faciliteren infrastructuurpartijen de gecontroleerde data- en rekenomgeving, terwijl modelontwikkeling en validatie onder publieke verantwoordelijkheid blijven.

GPT-NL

GPT-NL hanteert vanaf de start een waardengedreven benadering waarin representativiteit, ethiek en soevereiniteit centraal staan. Bij de samenstelling van de dataset is expliciet gestuurd op juridische zorgvuldigheid, transparantie over herkomst en aandacht voor maatschappelijke context en taalvariatie. Tegelijkertijd wordt onderkend dat de huidige dataset nog niet in alle opzichten optimaal representatief is voor de volledige diversiteit van de Nederlandse samenleving. Het verbeteren daarvan is een expliciete ontwikkelopgave voor de komende fases. Door deze keuzes bewust via Nederlandse en Europese afstemming te maken, draagt GPT-NL bij aan digitale soevereiniteit: publieke en private organisaties kunnen AI inzetten op basis van modellen die niet alleen technisch betrouwbaar zijn, maar ook aansluiten bij publieke waarden en democratische controle.

Het Vlaamse VLAM

Het VLAM-programma (Vlaamse AI-modellen) is een strategisch initiatief van imec, een organisatie ondersteund door de Vlaamse overheid dat zich richt op het versterken van de positie van Vlaanderen in het snel evoluerende landschap van artificiële intelligentie binnen de mediasector. Het programma vertrekt vanuit de vaststelling dat Vlaanderen beschikt over sterke troeven, zoals rijke media-archieven, een uitgesproken culturele identiteit en een

stevige technologische kennisbasis, maar dat deze onder druk komen te staan door de dominantie van internationale AI-spelers. VLAM wil daarom onderzoeken hoe Vlaanderen zelf meer regie kan nemen over de ontwikkeling en toepassing van AI, met aandacht voor economische valorisatie, maatschappelijke meerwaarde en digitale autonomie. Binnen het programma staat een verkennende en voorbereidende aanpak centraal. In plaats van onmiddellijk grootschalige investeringen te doen in nieuwe AI-modellen, wordt eerst ingezet op het opbouwen van kennis, het identificeren van kansen en risico's, en het creëren van een gedeeld strategisch kader. Samenwerking vormt daarbij een sleutelprincipe: VLAM brengt belanghebbenden uit de media-industrie, kennisinstellingen en het beleid samen om tot gedragen inzichten en toekomstige toepassingen te komen, met expliciete aandacht voor ethische en juridische randvoorwaarden.

Inhoudelijk focust het programma op drie samenhangende pijlers. Ten eerste worden data benaderd als een strategische hefboom, waarbij onderzocht wordt hoe Vlaamse media- en cultuurdata beter beheerd, ontsloten en ingezet kunnen worden in AI-toepassingen, zonder de controle erover te verliezen. Ten tweede ligt de nadruk op het aanpassen en verfijnen van bestaande AI-modellen, zodat deze beter aansluiten bij de Vlaamse taal, context en waarden, eerder dan volledig nieuwe modellen vanaf nul te ontwikkelen. Ten derde wordt gekeken naar de noodzakelijke infrastructuur en schaalbaarheid, met aandacht voor rekenkracht en technologische voorzieningen die Vlaanderen in staat moeten stellen om AI-toepassingen op een duurzame en onafhankelijke manier te ontwikkelen en uit te rollen. Met deze aanpak wil het VLAM-programma de basis leggen voor betrouwbare en contextgevoelige AI-toepassingen die geworteld zijn in de Vlaamse realiteit. Het programma beoogt niet alleen technologische innovatie, maar ook het versterken van de culturele en economische positie van Vlaanderen in een digitale wereld waarin data en artificiële intelligentie steeds bepalender worden.

Het Nederlandse VLAM (Veilige Lokale AI-Modellen)

Het Nederlandse VLAM is een AI-platform van SSC-ICT, de centrale ICT-dienstverlener van de Nederlandse rijksoverheid, die IT-diensten levert aan negen ministeries en ruim 58.000 rijksambtenaren. Niet te verwarren met het Vlaamse VLAM-programma. Het platform werd officieel gelanceerd in mei 2025 en draait in het overheidsdatacenter van SSC-ICT, waarbij uitsluitend gebruik wordt gemaakt van open source software en Europese aanbieders. Het platform is geen taalmodel op zichzelf, maar een infrastructuurlaag waarop ministeries eigen AI-toepassingen kunnen ontwikkelen en uitrollen op een manier die veilig, privacybestendig en compliant-by-design is. VLAM-chat biedt ambtenaren een chatfunctie vergelijkbaar met ChatGPT, maar gevoed met data en prompts die het datacenter niet verlaten. Daarnaast beschikt het platform over een zoekfunctie waarmee overheidsinformatie kan worden doorzocht, en een API waarmee ontwikkelaars eigen applicaties kunnen bouwen. Voor de brede uitrol in 2026 hebben acht Nederlandse ministeries deelname toegezegd, goed voor een initiële gebruikersgroep van circa 26.500 ambtenaren. Het platform richt zich op toepassingen zoals het samenvatten van documenten, schrijfondersteuning, vertalen en het doorzoeken van archieven. SSC-ICT hanteert het uitgangspunt dat uitsluitend Europese modellen worden ingezet. Zo wordt GPT-NL momenteel via VLAM gehost in een testomgeving, in het kader van een aantal haalbaarheidsstudies waarin de kwaliteit van het model verder wordt verbeterd en getest.

NOLAI

NOLAI is het nationale onderwijslab-AI voor het basisonderwijs, het voortgezet en het speciaal onderwijs van Nederland, gefinancierd door het Nationaal Groeifonds (<https://www.ru.nl/nolai>). Binnen NOLAI worden projecten opgezet in samenwerking tussen het onderwijs, het bedrijfsleven en wetenschappers om verantwoorde educatieve AI te ontwikkelen en te implementeren in scholen. Op dit moment zijn AI-oplossingen in het onderwijs in Nederland sterk afhankelijk van de AI van Big Tech. Ook initiatieven als NOLAI zijn dus gebaat bij toegang tot een soeverein en verantwoord AI-ecosysteem voor het Nederlands.

LLM-projecten uit de EU

Het project **OpenEuroLLM** is een Europees initiatief dat zich richt op de ontwikkeling van open, transparante, hoogwaardige taalmodellen die afgestemd zijn op de Europese context. Het project vertrekt vanuit de bezorgdheid dat de huidige generatie LLM's voornamelijk ontwikkeld wordt door een beperkt aantal niet-Europese technologiebedrijven, wat risico's inhoudt op het vlak van digitale soevereiniteit, transparantie en controle over data. OpenEuroLLM wil hierop een antwoord bieden door in te zetten op open modellen die niet alleen technisch goed presteren, maar ook voldoen aan Europese waarden zoals privacy, inclusie en verantwoord gebruik van artificiële intelligentie. Het project brengt een breed ecosysteem samen van onderzoeksinstellingen, bedrijven en publieke partners, met als doel een gedeelde infrastructuur en kennisbasis uit te bouwen voor de ontwikkeling en implementatie van deze modellen.

Binnen OpenEuroLLM ligt de nadruk op het ontwikkelen van meertalige modellen die de diversiteit van Europese talen en culturen ondersteunen en evalueren, inclusief minderheidstalen die vandaag vaak ondervertegenwoordigd zijn in AI-systemen. Daarnaast wordt sterk ingezet op openheid, zowel in termen van modelarchitecturen als datasets, zodat innovatie niet beperkt blijft tot een kleine groep spelers, maar toegankelijk wordt voor een breder veld van organisaties en ontwikkelaars. Tegelijk wordt aandacht besteed aan robuustheid, veiligheid en uitlegbaarheid van de modellen, zodat ze op een betrouwbare manier ingezet kunnen worden in sectoren zoals overheid, onderwijs en industrie. Door deze combinatie van openheid, kwaliteit en Europese verankering wil OpenEuroLLM bijdragen aan een duurzaam en onafhankelijk AI-ecosysteem in Europa.

EuroLLM, niet te verwarren met OpenEuroLLM, was een project dat afliep in april 2025, dat als doel had om een generatief taalmodel uit te brengen dat 24 officiële EU-talen ondersteunt. Het project publiceerde twee open-source modellen en een instructiedataset. EuroLLM kreeg de steun van de EU en de pan-Europese infrastructuur EuroHPC.

Het project **LLMs4EU** sluit hier inhoudelijk op aan, maar legt de nadruk meer expliciet op de toepassing en valorisatie van grote taalmodellen binnen specifieke Europese sectoren en maatschappelijke domeinen. Waar OpenEuroLLM zich sterk richt op de ontwikkeling van generieke modellen en infrastructuur, focust LLMs4EU op het vertalen van deze technologie naar concrete usecases die relevant zijn voor Europese organisaties, burgers en beleidscontexten. Het project onderzoekt hoe LLM's aangepast en ingezet kunnen worden in domeinen zoals publieke dienstverlening, gezondheidszorg en media, met bijzondere aandacht voor betrouwbaarheid, ethiek en naleving van Europese regelgeving.

Binnen LLMs4EU staat het concept van “adaptatie” centraal: bestaande of nieuw ontwikkelde taalmodellen worden afgestemd op specifieke talen, sectoren en toepassingen, zodat ze beter aansluiten bij de noden van gebruikers in Europa. Dit omvat onder meer het werken met domeinspecifieke data, het ontwikkelen van evaluatiekaders voor kwaliteit en bias, en het opzetten van experimentele omgevingen waarin toepassingen getest en gevalideerd kunnen worden. Daarnaast werken publieke en private partners samen aan het creëren van schaalbare oplossingen en het delen van best practices.

Met deze aanpak wil LLMs4EU bijdragen aan de effectieve integratie van grote taalmodellen in de Europese samenleving en economie. Het project streeft ernaar om niet alleen technologische vooruitgang te realiseren, maar ook vertrouwen te creëren in het gebruik van AI, door transparantie, verantwoordelijkheid en maatschappelijke relevantie centraal te stellen. Samen met initiatieven zoals OpenEuroLLM vormt LLMs4EU een belangrijke bouwsteen in de ontwikkeling van een sterk, autonoom en waardengedreven AI-landschap in Europa.

TrustLLM is een Horizon Europe-project dat tot doel heeft een open en betrouwbaar LLM te ontwikkelen voor Europa, met een eerste focus op de Germaanse talen. Het project (geleid vanuit Zweden) richt zich op het versterken van transparantie, nauwkeurigheid en vertrouwen in LLM-technologie. Het project heeft een breed consortium van academische en industriële partners en de ambitie van het project is om als blauwdruk kunnen dienen voor toekomstige AI-ontwikkeling in andere taalfamilies. Binnen TrustLLM wordt ook gewerkt aan EuroEval, een toolkit die het evalueren van taalmodellen vereenvoudigt.

High Performance Language Technologies (HPLT) was een Horizon Europe project dat eindigde in december 2025. De projectleiders zetten sterk in op het verzamelen en filteren van grote hoeveelheden web-data in 198 talen. Ze trainden taalspecifieke modellen alsook vertaalmodellen. Het onderzoek en deliverables (datasets, modellen) zijn openbaar te raadplegen en downloaden.

De **European Language Data Space (LDS)** is een initiatief binnen de Europese datastrategie dat tot doel heeft taaldata – zoals corpora, lexica en spraakdatasets – beter vindbaar, toegankelijk, interoperabel en herbruikbaar te maken. Door in te zetten op gestandaardiseerde metadata, gedeelde infrastructuren en duidelijke afspraken rond datagovernance en licenties, draagt de LDS bij aan een robuuste Europese digitale taalinfrastructuur die innovatie stimuleert.

De **Alliance for Language Technologies European Digital Infrastructure Consortium (ALT-EDIC)** is een Europees samenwerkingsverband dat gericht is op het versterken van de capaciteit voor taaltechnologie in Europa. Het initiatief focust op het ontwikkelen, delen en opschalen van hoogwaardige taaldata, tools en modellen, met bijzondere aandacht voor meertaligheid en minderheidstalen.

Common Language Resources and Technology Infrastructure (CLARIN) is een Europees onderzoeksinfrastructuur die zich richt op het duurzaam ontsluiten, beheren en verrijken van taaldata en -tools. CLARIN is opgezet als een **ERIC**, gebaseerd op een model van lidmaatschap van samenwerkende landen. Via een netwerk van erkende CLARIN-centra (zoals het INT) verspreid over Europa biedt het onderzoekers en ontwikkelaars toegang tot hoogwaardige corpora, lexica en analysetools. Door de sterke focus op duurzaamheid, kwaliteit, interoperabiliteit en toegankelijkheid vormt CLARIN een essentieel onderdeel van de bredere taalinfrastructuur en sluit het nauw aan bij initiatieven zoals de Language Data Space en ALT-EDIC.

Colofon

Auteurs

Tanguy Coenen (IMEC)
Saskia Lensink (TNO)
Antal van den Bosch (Universiteit Utrecht)
Catia Cucchiarini (Taalunie/Radboud Universiteit)
Walter Daelemans (Universiteit Antwerpen)
Martijn Kleppe (KB, tot 31 mei 2026)
Roeland Ordelman (Beeld en Geluid)
Vincent Vandeghinste (Instituut voor de Nederlandse Taal/KU Leuven)
Hugo Van hamme (KU Leuven)
Bram Vanroy (Instituut voor de Nederlandse Taal/KU Leuven)
Suzan Verberne (Universiteit Leiden)

We bedanken de leden van de [Raad voor de Nederlandse Taal en Letteren](#) voor hun commentaar op eerdere versies van dit actieplan.

Leden Raad voor de Nederlandse Taal en Letteren

Antoinet Brink
Jordi Casteleyn
Lieve De Wachter
Matthias Hüning (vicevoorzitter)
Franciska de Jong
Eric Mijts
Yasmien Naciri
Harold Polis
Gijsbert Rutten
Thony Visser
Rik Vosters (voorzitter)
Camille Welie

Folkert de Vriend (secretaris)

©2026 Nederlandse Taalunie

Niets uit deze uitgave mag worden verveelvoudigd en/of openbaar gemaakt door middel van druk, fotokopie, microfilm of op welke andere wijze dan ook zonder voorafgaande schriftelijke toestemming van de uitgever.

No part of this book may be reproduced in any form, by print, photocopy, microfilm or any other means without written permission from the publisher.



taal:
unie

E-mail
info@taalunie.org

Web
www.taalunie.org