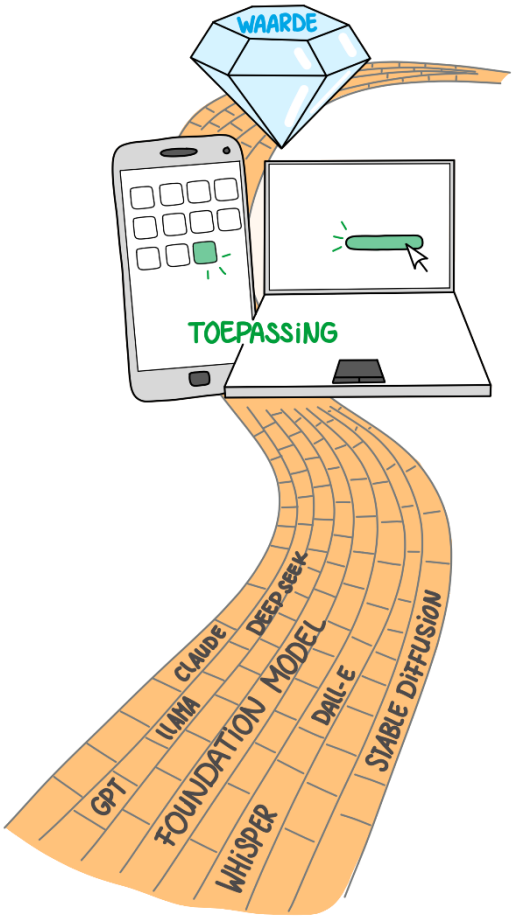


Auteur
Sietske Tacoma
Lectoraat Artificial Intelligence

Inlichtingen
Sietske.tacoma@hu.nl

Datum
Februari 2025

Bouwen op Foundation Modellen

Hoe ontwikkelen we waardevolle AI-toepassingen voor onze organisatie?

Inhoudsopgave

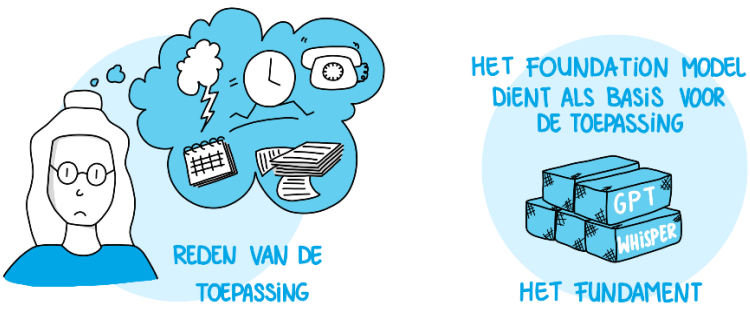
1 Inleiding

De recente snelle opkomst van grote AI-modellen en hierop gebaseerde toepassingen, zoals ChatGPT, heeft AI in veel organisaties hoog op de agenda gezet. Organisaties zien potentie in deze modellen voor het bewerken en creëren van tekst, beeld en audio. Daarom willen ze uitzoeken hoe deze modellen waarde kunnen opleveren binnen hun werkprocessen. Omdat de modellen nog nieuw zijn en de ontwikkelingen razendsnel gaan, levert dit ook veel vragen op, zoals:

- Welke modellen werken het beste voor de taken van onze organisatie?
- Hoe werken we zo met deze modellen dat onze gegevens veilig zijn?
- Is de kwaliteit van de resultaten en de efficiëntie van deze modellen goed genoeg om echt een verbetering van ons werk op te leveren?
- We weten dat dit soort modellen soms foute resultaten produceren, dus hoe beperken we het risico op schade door zulke fouten?
- Als we modellen van grote aanbieders zoals Microsoft, Google of Amazon gebruiken, hoe afhankelijk zijn we dan op langere termijn van hen en in hoeverre maakt dit ons kwetsbaar?
- Zien de beoogde gebruikers meerwaarde in het gebruiken van deze modellen?
- Zijn de beoogde gebruikers bereid en in staat om deze modellen in hun werk te gebruiken?

Deze handreiking is bedoeld om organisaties te helpen grip te krijgen op deze vragen en de zoektocht naar antwoorden. We richten ons hierbij op hoe organisaties toepassingen ontwikkelen met grote AI-modellen als basis, om deze toepassing vervolgens in te bedden in werkprocessen van werknemers en/of klanten. In mindere mate zijn de vragen en afwegingen die we bespreken ook van toepassing op hoe individuele werknemers direct gebruik maken van kant-en-klare AI-toepassingen, zoals ChatGPT[1].

De reden voor deze focus op toepassingen gebaseerd op AI-modellen, is dat de nieuwe generatie AI-modellen ook wel bekend staat als *Foundation Modellen.* Het zijn grote, generieke modellen die getraind zijn op zeer grote hoeveelheden data. Hierdoor presteren ze goed op algemene taken, zoals het beantwoorden van vragen of het creëren van een afbeelding op basis van een beschrijving. Gebruikers hebben met deze algemene modellen echter vaak niet genoeg regiemogelijkheden om op specifieke taken de resultaten te krijgen waarnaar ze op zoek zijn. Door een toepassing te ontwikkelen met een foundation model als basis, maakt de organisatie gebruik van de goede algemene prestaties van het model, en voegt informatie toe om de prestaties op specifieke taken te verbeteren.



[1] Over het opstellen van beleid over het gebruik van AI in organisaties zijn ook andere bronnen beschikbaar. Zie bijvoorbeeld deze blog van ISACA (Information Systems Audit and Control Association) en Nederlandse voorbeelden van de Nederlandse Organisatie voor Wetenschappelijk Onderzoek (NWO), ICT Institute en de Erasmus Universiteit Rotterdam.

1.1 Voor wie is deze handreiking bedoeld?

Voor een goede implementatie van een toepassing op basis van een Foundation Model is niet alleen technische kennis nodig. Er is ook kennis nodig over wie er mogelijk profijt hebben van zo'n toepassing. Daarnaast is er mandaat nodig om de benodigde verkenningen, experimenten en ontwikkelingen in gang te zetten en om succesvolle resultaten vervolgens ook in de praktijk te implementeren. Deze handreiking is bedoeld voor alle betrokkenen in dit proces, om te achterhalen welke vragen beantwoord moeten worden en welke, mogelijk conflicterende, belangen een rol spelen.

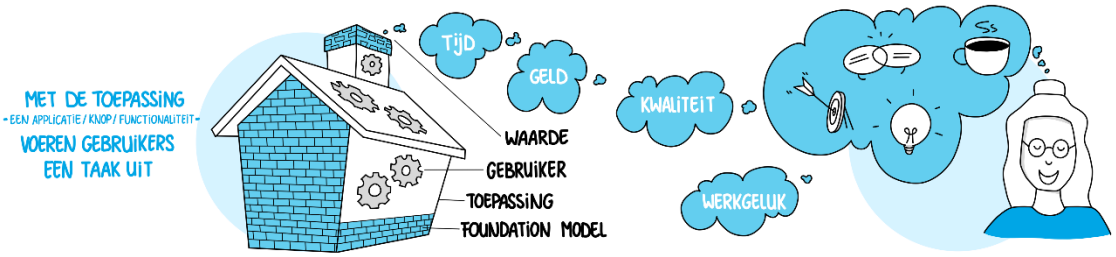
Deze handreiking is tot stand gekomen door onderzoek bij drie mediaorganisaties (ANP, NOS en Triple8) in samenwerking met Hogeschool Utrecht en ICT-coöperatie SURF. De drie mediaorganisaties hebben elk een toepassing ontwikkeld op basis van een Foundation Model. Deze toepassingen lichten we toe in het volgende hoofdstuk. De toepassingen zijn niet media-specifiek: we verwachten dat soortgelijke toepassingen ook in andere organisaties, zoals in de financiële of publieke dienstverlening, zinvol kunnen zijn. Daarom richten we deze handreiking aan organisaties in het algemeen: we verwachten dat de afwegingen die deze drie mediaorganisaties maken voor het overgrote deel ook buiten mediaorganisaties van toepassing zijn.

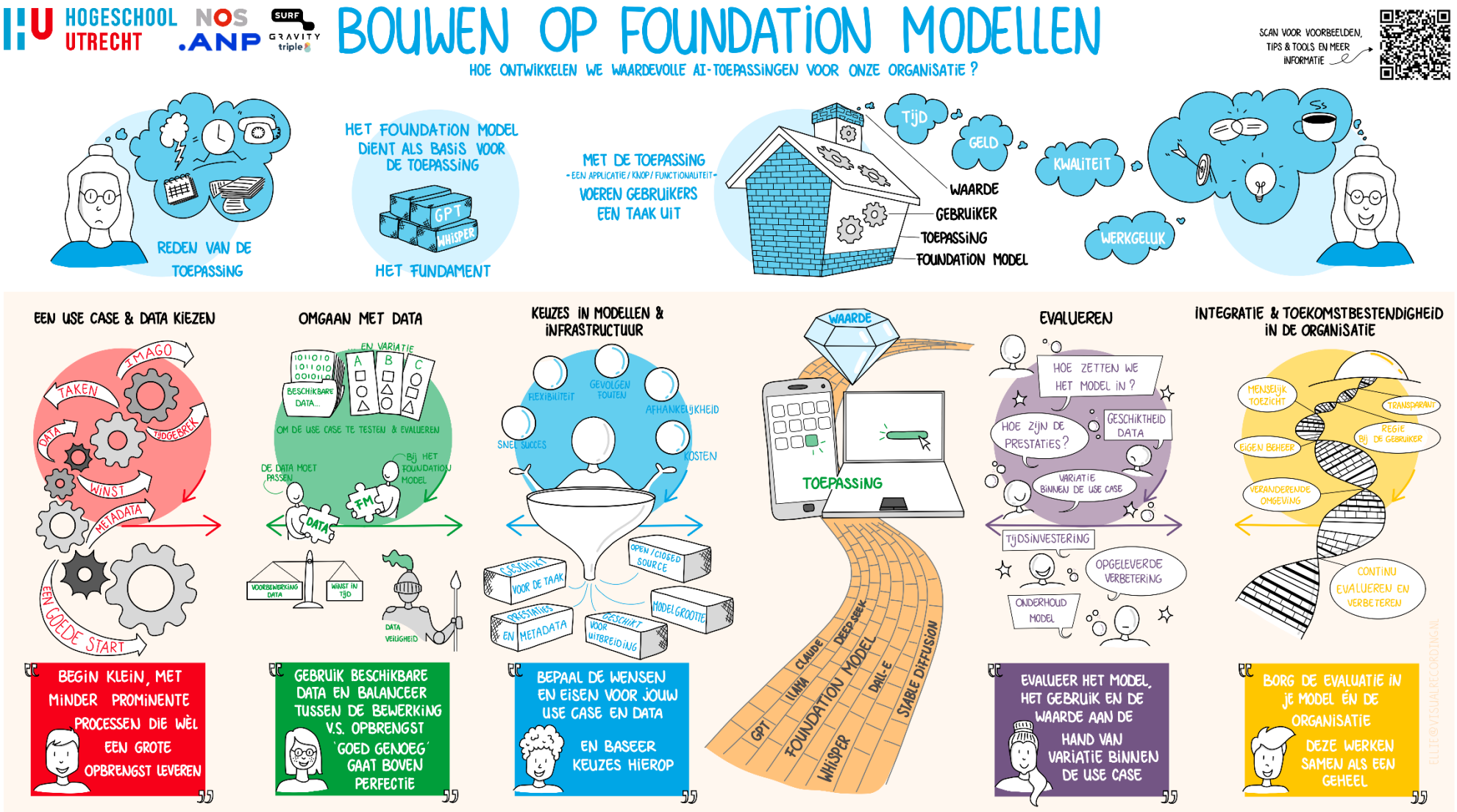
1.2 Leeswijzer

Het vervolg van deze handreiking is als volgt opgebouwd. In het volgende hoofdstuk lichten we de drie voorbeelden vanuit de mediaorganisaties (ANP, NOS en Triple8) toe. Daarna presenteren we een overzicht van te maken keuzes en afwegingen in het ontwikkelen van een toepassing gebaseerd op een Foundation Model binnen een organisatie. In de hoofdstukken daarna gaan we in meer detail in op de vijf hoofdthema's uit dit overzicht:

- Een use case en data kiezen
- Omgaan met data
- Keuzes in modellen en infrastructuur
- Evalueren
- Integratie en toekomstbestendigheid in de organisatie

Tot slot presenteren we de vragenlijst die Hogeschool Utrecht en SURF in dit project hebben gemaakt en gebruikt om de ontwikkelingen in de drie organisaties te monitoren. We verwachten dat niet alleen het overzicht van afwegingen, maar juist ook deze vragenlijst, organisaties kan helpen grip te krijgen op het werken met Foundation Modellen.





2 Voorbeelden

We illustreren afwegingen die organisaties kunnen maken en verschillende invullingen die zij hieraan kunnen geven. Dit doen we aan de hand van drie voorbeelden van toepassingen gebaseerd op Foundation Modellen.

2.1 Voorbeeld 1: expertsuggesties bij het ANP

Het ANP verspreidt dagelijks een groot aantal nieuwsberichten naar Nederlandse nieuwsmedia. De informatie die het ANP verspreidt is zo feitelijk mogelijk – het is aan de media om hier duiding aan te geven. Wel heeft het ANP goed contact met veel experts die duiding kunnen geven. Hiervan hebben zij een database aangelegd, die ruim 3000 experts bevat. Met behulp van een Foundation Model wil het ANP de inhoud van nieuwsberichten vergelijken met de inhoud van de cv's van deze experts. Zo wil het ANP automatisch geschikte expertsuggesties genereren, om met de nieuwsberichten mee te sturen naar de nieuwsmedia. Hiermee hopen zij te zorgen dat hun bestand van experts beter benut wordt en bij te dragen aan meer diverse en betere duiding van het nieuws.

2.2 Voorbeeld 2: persconferenties transcriberen bij de NOS

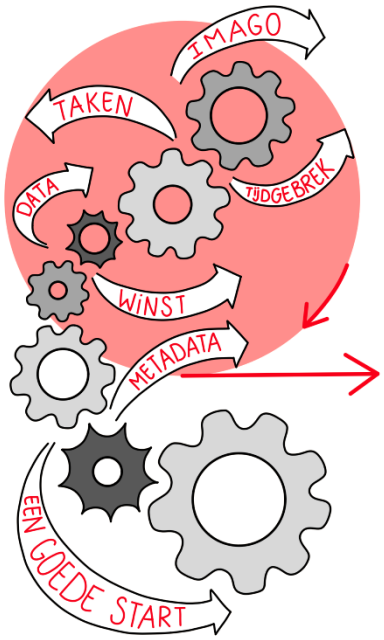
De afdeling DocuMedia van de NOS is verantwoordelijk voor het archiveren van ruw beeldmateriaal. Dit archief bevat veel opnames van persconferenties en Kamerdebatten. Dit materiaal gebruikt de NOS bijvoorbeeld als achtergrond bij nieuwsuitzendingen, maar ook om terug te kunnen halen wat er in eerdere debatten en persconferenties is gezegd over actuele onderwerpen. Om dit laatste sneller te kunnen doen, willen ze met behulp van een Foundation Model transcripties genereren van deze opnames. Dit maakt de opnames veel beter doorzoekbaar en maakt het bijvoorbeeld eenvoudiger om uitspraken die politici in eerdere debatten hebben gedaan terug te vinden.

2.3 Voorbeeld 3: Vlaamse ondertiteling bij Triple8 en Gravity

Een consortium van drie Vlaamse omroeporganisaties heeft Triple8 de opdracht gegeven om Vlaamse ondertiteling te genereren bij hun Vlaams gesproken tv-uitzendingen. Hiermee willen zij hun uitzendingen toegankelijker maken voor doven en slechthorenden en voldoen aan wetgeving die het aanbieden van ondertiteling verplicht De oplossing wordt ontwikkeld door Gravity en Triple8. Voor het ontwikkelen van deze toepassing, trainen ze een Nederlandstalig Foundation Model bij om ook Vlaamse uitspraak en uitdrukkingen goed te transcriberen.

3 Een use case en data kiezen

Binnen een organisatie bestaan veel verschillende taken. Bij het bepalen wat een goede taak is om Foundation Modellen voor in te zetten, houden organisaties rekening met de volgende aspecten:

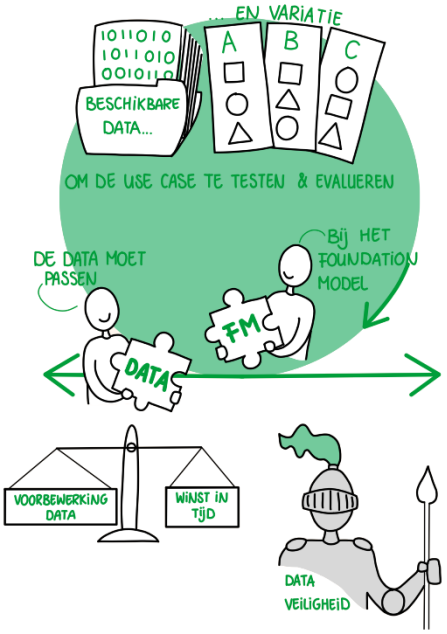


- Welke taken hebben wij als organisatie waar we goed Foundation Modellen voor kunnen inzetten? Foundation Modellen zijn snel befaamd geworden vanwege hun capaciteit om teksten en beelden te genereren. Maar als nieuwsorganisaties Foundation Modellen inzetten voor bijvoorbeeld het schrijven van nieuwsberichten, voldoen de resultaten niet aan de standaarden van de organisaties. Wanneer dit betekent dat medewerkers veel tijd kwijt zijn aan het verbeteren van de resultaten van een Foundation Model, bereikt de organisatie dus niet de beoogde toegevoegde waarde. Er zijn echter binnen een organisatie ook veel taken die minder prominent zijn, maar wel tijdrovend. Of taken die gedaan zouden kunnen worden, maar wegens tijdgebrek niet, of slechts ten dele, worden uitgevoerd. In dit soort taken zagen de organisaties meer waarde voor het inzetten van Foundation Modellen, zoals de voorbeelden ook laten zien.
- Voor welke taken valt met behulp van Foundation Modellen grote winst te behalen? Wat vormt een toevoeging op wat we al hebben? En waar is de gebruikersgroep groot genoeg om de investering te verantwoorden?
    - De Vlaamse omroeporganisaties waar Triple8 mee samenwerkt, merkten dat de werkdruk van hun ondertitelaars flink is gestegen, met name door een groei in online content. Een automatiseringsslag hierin kan de werkdruk dus flink verlagen.
    - De ANP-portal waarin de expertsuggesties zichtbaar worden, wordt dagelijks door honderden journalisten gebruikt. Een groot bereik dus, en daarmee echt potentie om de bronnenbank beter te benutten.
- Welke data hebben wij als organisatie beschikbaar, waar we gebruik van kunnen maken? In alle voorbeelden zagen we dat de organisaties gebruik maakten van data die zij, vanuit hun kerntaken, al hebben. Vaak hebben ze deze database (de bronnenbank van het ANP, het archief van debatten en persconferenties van de NOS, afleveringen van Vlaamse series bij Triple8) in de loop van de jaren opgebouwd en behoort het bijwerken en onderhouden van deze database tot hun kerntaken.
- Hoe voorkomen we imagoschade door het gebruik van Foundation Modellen? Organisaties zien op dit gebied wel risico's, omdat het kan gebeuren dat Foundation Modellen foutieve, verzonnen of irrelevante resultaten geven, die ongewenste gevolgen kunnen hebben. Daarom kiezen ze in eerste instantie voor veilige, voorzichtige implementaties en zetten ze Foundation Modellen niet in voor kerntaken, maar voor minder prominente taken. Zo kiezen organisaties voor het creëren van metadata en geen content. Ook ontwikkelen ze toevoegingen op en geen vervanging van bestaande werkwijzen en bieden ze gebruikers de mogelijkheid zo'n toevoeging uit te zetten.

- Hoe kunnen we een goede start maken met het gebruik van Foundation Modellen in onze organisatie? We zien dat organisaties ervoor kiezen klein te beginnen, om relatief snel succes te kunnen boeken. Dat helpt bij het creëren van draagvlak binnen de organisatie voor grotere investeringen. Ook zorgt het voor een context binnen de organisatie om in te experimenteren met Foundation Modellen, zodat de organisatie kan leren wat deze modellen voor ze kunnen betekenen. Klein beginnen houdt in dat organisaties in het begin kiezen voor kleine modellen en modelversies, met beperkte rekenkracht. Ook kiezen ze voor beginnen met schone, duidelijke data. Zo startten Triple8 en de NOS bij hun toepassingen voor het transcriberen van gesproken tekst met data waarin maar één stem tegelijk te horen was. Dit was over het algemeen een duidelijk verstaanbare mannenstem, opgenomen in een studio. Later in het proces breidden de organisaties de toepassing uit naar meer diverse stemmen, meerdere stemmen door elkaar, en meer omgevingsgeluid, door de modellen ook met dit soort meer diverse data te trainen.

4 Omgaan met data

Bij het selecteren van een use case zagen we dat organisaties hier graag data voor gebruiken die zij al beschikbaar hebben en waar ze als organisatie veel mee werken. Door te werken met deze data is veel meerwaarde te behalen. Maar organisaties moeten ook keuzes maken in hoe om te gaan met deze data bij het werken met Foundation Modellen.



Allereerst moet de data qua vorm en kwaliteit goed passen bij het Foundation Model dat een organisatie wil gebruiken. Organisaties besteden dan ook veel tijd en aandacht aan het selecteren, voorbewerken en opschonen van data.

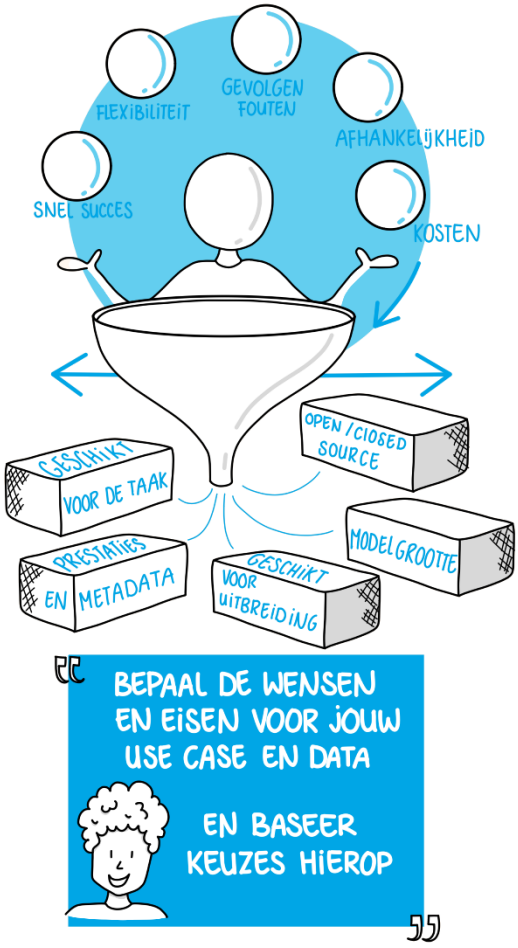
- De NOS gebruikt software voor het detecteren van spraak en presenteert alleen de fragmenten waarin daadwerkelijk verstaanbaar gesproken wordt aan het Foundation Model. Voor fragmenten zonder (goed verstaanbare) spraak levert een Foundation Model namelijk transcripten van slechte kwaliteit. Bovendien kosten dit soort fragmenten meer tijd en rekenkracht, omdat ze voor een Foundation Model moeilijker te transcriberen zijn dan audio met heldere spraak.
- Het team van Triple8 wilde audio presenteren aan een speech-to-text-model en de resultaten vergelijken met de ondertiteling die hun ondertitelaars al hadden gemaakt. Hiermee wilden ze een inschatting maken van hoe goed het model al presteerde en wat het bij moest leren. Ze hebben geëxperimenteerd met de lengte van fragmenten die ze goed aan het model konden presenteren en zowel audio als ondertiteling opgeknipt in korte fragmenten. Een uitdaging hierbij was dat de tijdcodes van de ondertiteling aangaven wanneer de ondertiteling in beeld was en daarom niet precies overeenkwamen met wanneer de tekst uitgesproken werd. Zoeken naar manieren om precies te bepalen welk stuk geschreven ondertiteling hoorde bij welk stuk audio kostte veel experimenteertijd.
- Het ANP gebruikt Foundation Modellen om te bepalen welke cv's van experts goed passen bij nieuwsartikelen. In eerste versies vielen allerlei zaken op. Zo werden bij veel nieuwsberichten experts over Den Haag gesuggereerd, omdat de plaatsnaam Den Haag vaak wordt genoemd bij nieuws dat het ANP vanuit andere kanalen naar buiten brengt. Ook bleek de toepassing associaties te maken op basis van de namen van de experts – een expert met de naam Van Leeuwen werd gesuggereerd als expert bij nieuws over dierentuinen. Daarnaast beschrijven sommige experts in hun cv ook kennisgebieden en onderwerpen waarvoor zij beter niet benaderd kunnen worden. De Foundation Modellen merkten zulke nuances niet op. Voor het ANP was dit inzicht aanleiding om hun database met experts anders in te richten: één veld voor expertise van de expert, en een ander veld voor anti-expertise.

Organisaties ervaren dit proces over het algemeen niet als ingewikkeld, maar wel als tijdrovend. Ze zoeken hierin naar balans. De tijd die ze steken in het voorbewerken van data om goed te passen bij de Foundation Modellen moet opwegen tegen de winst (in tijd, in waarde voor de organisatie) die ze later behalen door de toepassing in de organisatie in te zetten. Daarbij zijn ze, zoals de voorbeelden laten zien, kritisch op de kwaliteit van de data, maar streven ze niet naar perfecte data. Ze streven naar data waar het model goed genoeg mee overweg kan in de context van de geselecteerde use case. Wel kiezen ze, om een goede start te kunnen maken, vaak voor afgebakende doelgroepen en, in eerste instantie, duidelijke schone data. Ook denken ze na over hoe ze dit proces toekomstbestendig inrichten. Zo leggen ze databases aan met opgeknipte fragmenten en bijbehorende metadata, om deze in nieuwe trainings- en ontwikkeltrajecten eenvoudig te kunnen hergebruiken.

De data die organisaties gebruiken is vaak waardevol voor de organisaties. Zo is in veel gevallen de data eigendom van de organisatie of van klanten van de organisatie. In sommige gevallen bevat de data ook privacygevoelige informatie. Het is daarom belangrijk voor organisaties om zorgvuldig met de data om te gaan, om datalekken en andere vormen van schade te voorkomen. In de keuze voor modellen (zie volgend hoofdstuk) zien we dan ook dat dataveiligheid een rol speelt.

5 Keuzes in modellen en infrastructuur

Voor het ontwikkelen van AI-toepassingen kiezen organisaties vaak een Foundation Model als basis. Bij het kiezen van specifieke Foundation Modellen en de precieze implementatie in de organisatie spelen veel factoren een rol. Organisaties zoeken naar een balans tussen het boeken van succes op de korte termijn en de mogelijkheid tot uitbreiding op de langere termijn. Ook houden ze in hun keuzes rekening met de gevolgen van mogelijke fouten van de modellen en daarop gebaseerde toepassingen en denken ze na over eventuele afhankelijkheid van Big Tech bedrijven. Tot slot spelen ook kosten op diverse manieren een rol.



5.1 Geschiktheid voor de taak

Allereerst is het zaak om uit te zoeken welke Foundation Modellen passen bij de taak die de organisatie voor ogen heeft. De bekendste voorbeelden van Foundation Modellen zijn grote taalmodellen (Large Language Models). Op basis van een opdracht (een prompt) genereren deze modellen een tekst. Deze modellen kunnen veel verschillende taken uitvoeren, zoals vragen beantwoorden en teksten vertalen, samenvatten en categoriseren. Tegenwoordig zijn deze modellen vaak gecombineerd met beeldmodellen, die plaatjes (en steeds vaker ook video's) als input kunnen verwerken en ook als output kunnen geven. Dit maakt het mogelijk plaatjes te genereren, te beschrijven en te bewerken. Maar er zijn ook Foundation Modellen die andere typen data, zoals audiofragmenten, als input krijgen en daarvan bijvoorbeeld een transcriptie teruggeven. Een overzicht van allerlei taken die Foundation Modellen kunnen uitvoeren is te vinden op Hugging Face, een platform waar veel ontwikkelaars samenwerken aan AI-modellen en -toepassingen.

5.2 Prestaties en metadata

Voor elk van deze taken zijn diverse Foundation Modellen beschikbaar. Overzichten hiervan zijn te vinden op platforms zoals Hugging Face en Ecosystem Graphs for Foundation Models. Met behulp van Leaderboards, bijvoorbeeld beschikbaar op Hugging Face, is te zien hoe deze modellen presteren op specifieke taken (benchmarks). Naast prestaties op dit soort benchmarks spelen ook andere eigenschappen van modellen een rol in de selectie. Het kan bijvoorbeeld uitmaken welke metadata modellen meegeven met hun output.

- Bij het gebruiken van audiotranscriptie voor ondertiteling is het voor Triple8 een pré dat modellen niet alleen de transcriptie, maar ook heel precieze timing van de getranscribeerde tekst in het audiofragment teruggeven.

5.3 Modelgrootte

Ook de grootte van modellen speelt een rol. Het inzetten van Foundation Modellen vraagt om investeringen. Om deze investeringen met vertrouwen te kunnen doen, willen organisaties graag, met beperkte rekenkracht en kleine modellen, al kleine successen boeken. Starten met een klein model, of een kleine versie van een model, heeft als bijkomend voordeel dat niet alleen de benodigde rekencapaciteit, maar ook de benodigde opslagcapaciteit lager is.

- Het ANP gebruikte een Foundation Model om vectorrepresentaties van zowel nieuwsberichten als cv's van experts te maken. Vectorrepresentaties zijn lange rijen getallen. Hoe lang deze rijen zijn, hangt af van de gekozen versie van het Foundation Model. Het ANP experimenteerde met verschillende versies en koos uiteindelijk een relatief kleine versie, dus minder lange rijen, om sneller succes te kunnen boeken. Het rekenen met deze minder lange rijen getallen kost minder rekenkracht en daarmee minder tijd, en voor het opslaan van deze minder lange rijen is ook minder opslagcapaciteit nodig. Tegelijkertijd waren de prestaties van het kleinere model maar een klein beetje minder goed dan de prestaties van de grotere versies.

5.4 Geschiktheid voor uitbreiding use case

Een goede start maken helpt om te leren hoe Foundation Modellen waarde kunnen toevoegen voor de organisatie. Organisaties willen graag bij de ontwikkeling van toekomstige toepassingen kunnen voortbouwen op de nu ontwikkelde kennis en infrastructuur. Daarom houden ze bij de keuze voor modellen niet alleen rekening met wat nodig is voor een goede start, maar ook met wat hen in de toekomst, voor de volledige use case of voor volgende use cases, nodig lijkt.

- Het ANP en Triple8 gaven de voorkeur aan modellen die goed met meerdere talen om kunnen gaan, om internationaal (met buitenlandse persbureaus en met internationale omroeporganisaties) goed te kunnen samenwerken op het gebied van AI-toepassingen.
- Triple8 en de NOS startten met audiofragmenten waarin maar één stem tegelijk te horen was, maar keken in hun modelselectie ook naar hoe goed de modellen presteerden bij meerdere stemmen door elkaar.

5.5 Aanbieder van het model

Het maakt voor organisaties uit wie het Foundation Model dat ze willen gebruiken aanbiedt. Organisaties kiezen in veel gevallen voor modellen van grote aanbieders (Big Tech), zoals Google (Deepmind) en OpenAI. Deze aanbieders bieden goed presterende (state-of-the-art) modellen met goede documentatie, ondersteuning en onderhoud. Daarnaast verwachten organisaties ook dat deze modellen, of verbeteringen hiervan, nog geruime tijd beschikbaar zullen zijn en ondersteund en onderhouden zullen worden. Tegelijkertijd maken organisaties zich zorgen over de hoeveelheid controle die ze hebben over modellen van Big-Tech-aanbieders. Deze modellen zijn in veel gevallen closed source, waardoor niet duidelijk is met welke data de modellen precies zijn getraind en welke keuzes er precies zijn gemaakt tijdens het trainen van de modellen. Bij updates van modellen kunnen aanbieders andere keuzes hebben gemaakt, waardoor de modellen niet meer werken zoals de organisaties verwachten. Ook is niet altijd duidelijk wat er gebeurt met data van de organisatie, die ze tijdens het ontwikkelen of gebruiken van een toepassing aan het model presenteren. Organisaties weten niet waar deze data, die privacygevoelig kan zijn en intellectueel eigendom kan bevatten, terecht komt en of deze veilig is. Ze verdiepen zich in licenties en contracten die de grote aanbieders bieden, maar het is lastig de consequenties hiervan te overzien. Omdat het grote, bekende partijen betreft, hebben organisaties vertrouwen dat hun data er veilig genoeg is. Organisaties zoeken hiervoor ook

samenwerking met juridische afdelingen en denken na over good practices, zoals het centraal coördineren van AI-systemen die binnen een organisatie worden gebruikt en medewerkers informeren over mogelijke consequenties van AI-gebruik met gevoelige data. De Toolbox Ethisch Verantwoorde Innovatie van de Digitale Overheid bevat veel tools en tips om hier goed mee om te gaan.

Open source modellen bieden meer transparantie en mogelijkheden voor controle. Ook kunnen organisaties deze modellen beter aanpassen op eigen specifieke behoeften. Omdat het bij open source modellen goed mogelijk is eigen instanties te maken en implementeren, hoeven organisaties hiervoor geen data met externe partijen te delen. Dit vraagt wel veel technische kennis, waardoor organisaties in eerste instantie toch vaak voor de modellen van Big-Tech-aanbieders kiezen. Omdat organisaties met deze modellen vaak al grote verbeteringen zien voor hun use cases, zonder verdere finetuning met hun eigen data, heeft het investeren in open source modellen (in zowel kennis als financiële middelen) niet altijd de eerste prioriteit.

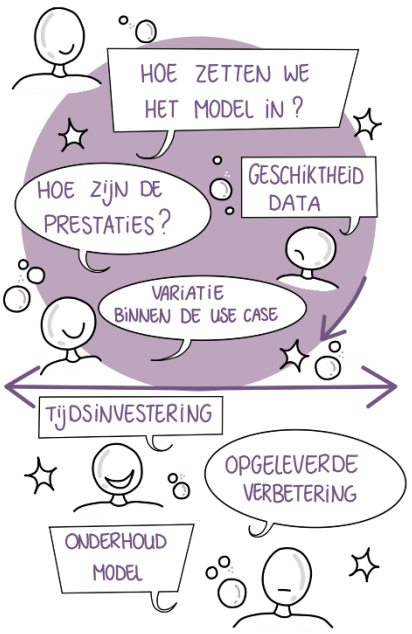
5.6 Keuzes in infrastructuur

Naast de keuze voor *welk* model in te zetten, zijn er ook veel keuzes te maken in *hoe* het model in te zetten. Organisaties zoeken op het gebied van infrastructuur naar flexibiliteit. Ze investeren in het inrichten van een pipeline en architectuur waarin ze eenvoudig verschillende modellen naast elkaar kunnen implementeren en met elkaar kunnen vergelijken. Dit maakt het mogelijk te experimenteren en de snelle ontwikkelingen bij te houden. Ook hier speelt de vraag in hoeverre organisaties hier afhankelijk willen zijn van Big Tech. Voor een goede AI-infrastructuur zijn verschillende onderdelen nodig, zoals opslagcapaciteit, faciliteiten voor databewerking en -uitwisseling, cloud computing en training- en testfaciliteiten. Het afnemen van deze onderdelen als totaalpakket bij dezelfde (grote) partij zorgt voor soepele integratie tussen de verschillende onderdelen, maar maakt de organisatie ook afhankelijk van deze aanbieder en de keuzes die deze aanbieder maakt. Aan de andere kant bieden open source componenten soms wel betere mogelijkheden om de architectuur te laten aansluiten op de eigen systemen, wat een soepele integratie ten goede komt. Kosten en benodigde expertise spelen hierin ook een rol. Bij grote aanbieders kunnen organisaties relatief goedkoop en eenvoudig onderdelen afnemen, terwijl ze voor open source modellen veel zelf moeten hosten en onderhouden. Het ontwikkelen, opzetten en onderhouden van al deze onderdelen wordt ook wel aangeduid met de noemer MLOps. Online zijn veel vergelijkingen voor MLOps tools en platforms te vinden, zoals bij Neptune.ai.

In de infrastructuur ontwikkelen organisaties ook manieren om te zorgen dat de toepassing niet te grote fouten maakt. Een nadeel van Foundation Modellen is dat ze onjuiste resultaten (ook wel aangeduid als hallucinaties) of irrelevante resultaten kunnen geven. Organisaties voegen daarom in de infrastructuur regels en filters toe, om specifieke ongewenste resultaten uit te sluiten of om verschillend te handelen voor verschillende categorieën. Bij het ANP is er bijvoorbeeld een filter ingesteld om aankondigingen, die via dezelfde feed worden verspreid als persberichten, niet ook van bronnensuggesties te voorzien. Een aantal algemenere voorbeelden van het combineren van Foundation Modellen en regels is te vinden in dit artikel op Towards Data Science.

6 Evalueren

Of een Foundation Model waarde toevoegt voor de organisatie, hangt niet alleen af van hoe goed het Foundation Model presteert, maar ook van hoe de organisatie dit model precies inzet. Daarom is het belangrijk om te evalueren hoe de Foundation Modellen en daarop gebaseerde toepassingen presteren in de context van de gekozen use case. Omdat dit context-specifiek is, zijn hier vaak minder kant-en-klare oplossingen voor dan de benchmarks waarmee algemene prestaties van Foundation Modellen worden beoordeeld. Organisaties ontwikkelen daarom zelf methodes om hun toepassing te evalueren. Dit automatiseren ze waar mogelijk.



- Triple8 bouwde een component die aan elke nieuwe modelversie hetzelfde audiobestand voerde. De output van het model werd steeds vergeleken met een handmatig gemaakt transcript, om te bepalen hoe groot de verschillen waren.

In de evaluatie houden organisaties er rekening mee of alleen het beste resultaat geëvalueerd moet worden (zoals bij de transcripties in de use cases van Triple8 en de NOS), of dat de toepassing meerdere mogelijk relevante resultaten geeft (zoals de bronnensuggesties bij het ANP). Hierbij is weer van de context afhankelijk of het erger is een geschikt resultaat te missen, of juist een niet geschikt resultaat toch te presenteren. Ook evalueren de organisaties de toepassing voor de verschillende categorieën en situaties waarin de toepassing gebruikt moet worden.

- Het ANP merkte dat de toepassing voor nieuwsberichten in de categorieën economie en buitenland al snel heel geschikte aanbevelingen voor experts kon doen. Voor de categorieën sport en entertainment vond de toepassing vaak nog niet de juiste experts. Daarom besloten zij een regel toe te voegen die zorgde dat de toepassing voor deze categorieën (nog) geen expertsuggesties geeft. Dit gaf het ANP overigens niet alleen inzicht in hoe hun toepassing werkte, maar ook aanleiding om kritisch te kijken naar de samenstelling van hun expert-database.

Organisaties vinden het belangrijk experts te betrekken in de evaluatie en hen steekproefsgewijs de kwaliteit van de resultaten van de toepassing te laten beoordelen.

- Het ANP selecteerde een steekproef van 250 nieuwsberichten en de experts die verschillende modelversies hierbij suggereerden. Journalisten van het ANP beoordeelden deze suggesties en op basis hiervan werden keuzes voor modellen en versies gemaakt.

Wanneer resultaten van de toepassing gebruikt worden in de organisatie, proberen organisaties uit dit gebruik ook informatie te verzamelen die de evaluatie helpt.

- Bij Triple8 is de verwachting dat de resultaten van de toepassing niet perfect zijn en dat ondertitelaars deze dus nog moeten nabewerken. Dit doen de ondertitelaars als onderdeel van hun dagelijks werk, om tot goede ondertiteling te komen. Door deze

handmatige nabewerkingen op te slaan, verzamelt Triple8 automatisch nieuwe trainingsdata om nieuwe versies van het model te trainen.
- Het ANP wil afnemers vragen om de expertsuggesties te beoordelen door een duim omhoog of omlaag te geven. Dit levert informatie over welke expertsuggesties journalisten als waardevol ervaren.

Een factor die op diverse manieren een rol kan spelen in de evaluatie is tijd. Ten eerste is er tijd nodig om een nieuwe versie van een model te trainen of in gebruik te nemen. Dit kan gaan om de vraag hoelang het duurt om een hele database te embedden (zoals bij het ANP), of hoelang het duurt om een nieuwe versie van een model te trainen als er nieuwe data beschikbaar is (zoals bij Triple8). Ten tweede heeft een toepassing tijd nodig om resultaten te genereren bij nieuwe input. Kleinere modellen zijn vaak aanzienlijk sneller in het genereren van transcripten dan grotere modellen, maar leveren ook vaak minder goede kwaliteit. Afhankelijk van de use case kan dit kwaliteitsverlies opwegen tegen de tijdwinst.

Een vraag die in de evaluatie op kan komen is wanneer het resultaat goed genoeg is. Organisaties vinden het vaak lastig dit te duiden. Ze geven aan niet te zoeken naar een zo goed mogelijk resultaat, maar naar verbetering van de huidige situatie door inzet van de toepassingen gebaseerd op Foundation Modellen. Daarbij willen ze ook aandacht geven aan hoe de toepassing de werkzaamheden van de gebruiker uiteindelijk verandert. In de academische wereld gaat veel aandacht uit naar kleine verbeteringen in de nauwkeurigheid van modellen. Maar voor de gebruikers maakt het in de context mogelijk helemaal niet uit of de toepassing net wat minder fouten maakt, als de gebruiker nog steeds alles handmatig moet nalopen.

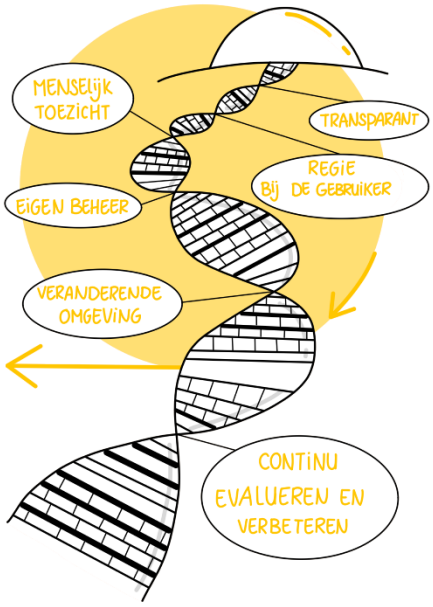
7 Integratie en toekomstbestendigheid in de organisatie

Om van toegevoegde waarde te zijn voor een organisatie moet een toepassing daadwerkelijk gebruikt worden. Ook moet deze op langere termijn blijven functioneren en geen schade veroorzaken voor de organisatie. Daarnaast vinden organisaties het belangrijk om transparant te zijn over hoe ze AI gebruiken.

- Het ANP voegde een disclaimer toe over hoe AI gebruikt is om tot de gegeven aanbevelingen te komen.

Ook besteden organisaties aandacht aan het verantwoord gebruik van de toepassing. Ze hebben bijvoorbeeld discussies over menselijk toezicht op de toepassing, om te voorkomen dat toepassingen autonoom handelen.

- Bij Triple8 wordt de door de toepassing gegenereerde ondertiteling altijd nog door een ondertitelaar nabewerkt.
- Voor het ANP was het belangrijk dat de toepassing aanbevelingen doet en geen besluiten neemt. Het nemen van besluiten over te benaderen experts blijft, bij het gebruik van deze toepassing, de verantwoordelijkheid van journalisten.



Meer manieren om aan de slag te gaan met het verantwoord gebruik van toepassingen zijn te vinden in het Impact Assessment Mensenrechten en Algoritmes.
Om te zorgen dat de toepassing daadwerkelijk gebruikt wordt, besteden organisaties aandacht aan de integratie van de toepassing in hun systemen en met name in de systemen waar de beoogd gebruikers mee werken. Ook om deze reden ontwikkelen organisaties bij voorkeur toepassingen in portals en systemen die ze in eigen beheer hebben. Vaak implementeren ze de toepassing als toevoeging op een bestaand systeem of product en hebben ze in deze implementatie aandacht voor hoe de toepassing in de werkwijze van de gebruiker past.

- Het ANP implementeerde de toepassing in de ANP-app, die journalisten toegang geeft tot de persberichten die het ANP verspreidt. Aan de metadata bij het persbericht worden de aanbevelingen voor experts toegevoegd.
- De NOS implementeert de toepassing als extra zoekmogelijkheid binnen hun archief-database.
- Triple8 implementeerde de toepassing als optie in een versie van het platform Triple8, dat zij aanbieden aan ondertitelaars voor het genereren en redigeren van ondertiteling.

Organisaties houden er rekening mee dat de omgeving waarin hun toepassing moet werken kan veranderen. Nieuwsorganisaties merken bijvoorbeeld dat er regelmatig nieuwe woorden ontstaan of belangrijk worden (zoals in het recente verleden corona/covid en ChatGPT). Ze denken na over hoe ze hun toepassing op regelmatige basis moeten blijven evalueren, onderhouden en laten bijleren om deze veranderingen bij te houden. Voor het gebruiken van nieuwe data (zoals langere video's of video's met meerdere stemmen door elkaar in de toepassingen van de NOS en Triple8)

stellen ze procedures op om te evalueren of andere voorbewerking nodig is op deze data en of de prestaties van de toepassing ook voor deze nieuwe data aan de verwachtingen voldoen. Ook hier helpen structuren als MLOps om dit onderhouden en monitoren vorm te geven.

8 De vragenlijst

Hieronder volgt de vragenlijst die tijdens het project is gebruikt om de ontwikkeling van toepassingen gebaseerd op Foundation Modellen in kaart te brengen. Deze vragen kunnen ontwikkelaars en andere betrokkenen binnen organisaties helpen om grip te krijgen op de verschillende aspecten van zo'n ontwikkeling en keuzes die ze daarin moeten maken. Ons advies is om, met de betrokkenen, gedurende de ontwikkeling meermaals deze vragen te beantwoorden. Hierbij adviseren we ook om bij te houden welke alternatieven zijn overwogen en welke argumenten de doorslag hebben gegeven bij het maken van keuzes.

De vragenlijst is tot stand gekomen op basis van onderzoek door de Hogeschool Utrecht en SURF, op basis van wetenschappelijke literatuur en raadpleging van experts. In dit paper lichten we de totstandkoming verder toe.

Vragenlijst Toepassingen ontwikkelen gebaseerd op Foundation Modellen

Beoogd gebruik
De vragen in deze categorie gaan over het beoogd gebruik van de toepassing die jullie ontwikkelen.

Doel
Met welk doel voor ogen is de toepassing ontwikkeld?
Welke taak heeft de toepassing die jullie ontwikkelen?
In welke context moet de toepassing worden gebruikt?
Welke taken zijn out-of-scope? Dus waarvoor is de tool niet bedoeld?

Beoogde gebruikers
Wie zijn de gebruikers van de toepassing?
Hoe groot is de gebruikersgroep?
Hoe zijn de beoogd gebruikers betrokken bij het ontwikkelproces?

Rol in werkzaamheden
Op welke manier integreert de gebruiker de toepassing in zijn of haar werkzaamheden?
Hoe vaak gebruikt de gebruiker de toepassing?
Hoe hangt de toepassing samen met andere toepassingen die de gebruiker gebruikt?

Modeleigenschappen
De vragen in deze categorie gaan over het AI-model dat jullie gebruiken in de ontwikkelde toepassing.

Architectuur
Wat voor type architectuur is gebruikt in de toepassing?
Welk(e) foundation model(len) is/zijn gebruikt als basis voor de toepassing?
Gebruiken jullie het foundation model zoals het is aangeboden, finetunen jullie het met trainingsdata, of passen jullie de architectuur van het model aan?
Welke (hyper)parameters kunnen worden aangepast en welke waarden zijn gekozen?

Trainingsdata
(alleen nodig wanneer jullie een bestaand model aanpassen of finetunen, niet wanneer jullie een bestaand model direct gebruiken binnen de toepassing)
Welke dataset(s) is/zijn gebruikt voor de ontwikkeling van de toepassing?
Hoe is de dataset (inclusief annotatie) tot stand gekomen?
Hoe is de data voorbewerkt?
Wat is de kwaliteit van de trainingsdata?
Welke selectiecriteria voor het wel of niet gebruiken van de data zijn gehanteerd?

Ontwikkelaars
Wie ontwikkelt het model?
Welke delen van de toepassing ontwikkelen jullie binnen de organisatie en welke delen worden ontwikkeld door andere partijen?

Training en prestaties van het model en de toepassing

Metrics
Welke metrics gebruiken jullie om het model te evalueren?
In hoeverre bepalen en monitoren jullie metrics voor verschillende groepen of categorieën? (Zie ook omgevingsfactoren)
Indien van toepassing: welke decision tresholds gebruiken jullie?
Welke mate van variatie zien jullie in de evaluatiemetrics?
Hoe evalueren jullie of de toepassing daadwerkelijk geschikt is voor de taak?

Trainingsprocedure
(alleen nodig wanneer jullie een bestaand model aanpassen of finetunen, niet wanneer jullie een bestaand model direct gebruiken binnen de toepassing)
Hoe ziet de trainingsprocedure eruit?
(Op welke manier) is gebruik gemaakt van cross-validation?
Zijn resultaten van meerdere runs gecombineerd?

Evaluation data
Welke dataset(s) is/zijn gebruikt voor de evaluatie van de toepassing?
Hoe is de dataset (inclusief annotatie) tot stand gekomen?
Hoe is de data voorbewerkt?
Hoe is gezorgd dat deze dataset geschikt is voor evaluatie (rekening houdend met omgevingsfactoren en representativiteit)?

Reikwijdte toepassing van het model (omgevingsfactoren)

Groepen
Voor welke verschillende groepen (cultureel, demografisch, fenotypisch, …) moet de toepassing presteren?
Hoe houden jullie in de trainingsdata, trainingsprocedure en evaluatie rekening met deze groepen?

Instrumentatie
Voor welke variatie in instrumentatie (bijvoorbeeld beeldkwaliteit, geluidskwaliteit, …) moet het model presteren?
Hoe houden jullie in de trainingsdata, trainingsprocedure en evaluatie rekening met deze variatie in instrumentatie?

Omgeving
Voor welke variatie in omgevingsfactoren (bijvoorbeeld licht, weersomstandigheden, …) moet het model presteren?
Hoe houden jullie in de trainingsdata, trainingsprocedure en evaluatie rekening met deze variatie in omgevingsfactoren?

Implementatie, onderhoud en ontwikkeling	
	Implementatie Hoe wordt de toepassing in de organisatie geïmplementeerd? Wie besluiten mee over de daadwerkelijke implementatie? **Onderhoud** Hoe wordt de toepassing onderhouden? Wie zijn betrokken bij het onderhouden van de toepassing? Hoe zijn jullie van plan te monitoren of het model blijft presteren zoals beoogd? (Model drift) Welke risico's hebben jullie geïdentificeerd en hoe gaan jullie hiermee om? **Ontwikkeling** Hoe en hoe vaak wordt een nieuwe versie van het model getraind? Hoe wordt omgegaan met nieuwe beschikbare data? Wie zijn betrokken bij het verder ontwikkelen van de toepassing?