Open en Europese taalmodellen bij de overheid – definities en afwegingen

Auteurs
Jens van der Weide
Sophie van Baalen
Gabriela Bodea
Marianne Schoenmakers
Stephan Raaijmakers

TNO-2026-17545

Augustus 2026

TNO | Vector

Samenvatting

Generatieve AI kunnen overheden nieuwe kansen bieden voor de dienstverlening voor burgers en bedrijven en de uitvoering van eigen taken. Geopolitieke ontwikkelingen maken duidelijk dat het belangrijk is om kritisch te kijken naar digitale afhankelijkheden. Bij het versterken van digitale soevereiniteit van de Nederlandse overheid is het stimuleren van het gebruik van open source modellen en Europese alternatieven cruciaal.

Begrippen als ‘open source’, ‘open AI' en ‘Europees’ zijn echter niet eenduidig en kunnen op verschillende manieren worden geïnterpreteerd. Aanbieders kunnen veelal zelf bepalen hoe zij hun taalmodellen kwalificeren. Dit schept onduidelijkheid voor gebruikers.

Dit rapport geeft inzicht in deze concepten, en formuleert criteria en waarden die belangrijk zijn bij het maken van keuzes rondom de inzet van deze taalmodellen.

De belangrijkste conclusies en aanbevelingen zijn als volgt.

'Open' is een spectrum, geen schakelaar
De openheid van een taalmodel varieert van volledig open (trainingsdata, code én gewichten beschikbaar) tot open-weight (alleen de gewichten) tot volledig gesloten (alleen API-toegang).

‘Europees’ is niet eenduidig
Om te bepalen of een model Europees is moeten verschillende aspecten in overweging worden genomen. Zoals de herkomst van de provider, de locatie van de dataopslag en – verwerking, de hostinglocatie en de herkomst van de trainingsdata.

Bovendien staan de dimensies 'openheid' en 'Europeesheid' niet los van elkaar. Deze complexiteit leidt ertoe dat er behoefte is aan een hulpmiddel voor het maken van afwegingen over de inzet van open en Europese taalmodellen.

Dit rapport presenteert een afwegingentabel met verschillende criteria om de openheid en Europeesheid van een taalmodel te beoordelen, inclusief de samenhang tussen deze dimensies een uitwerking van hoe dit kader geoperationaliseerd kan worden.

Aanbevelingen:
- Stel werkdefinities vast van ‘open AI’ en ‘Europees’ om ‘*open-washing*’ te voorkomen.
- Vertaal de criteria in het afwegingstabel naar een praktisch afwegingskader of (interactief) beslissysteem.
- Stel een leaderboard op met ervaringen van overheidsgebruikers met open en Europese modellen
- Neem meer afwegingen mee bij de implementatie van een model in een AI-applicatie .
- Wees alert op nieuwe vormen van afhankelijkheid.
- Koester, bescherm en stimuleer het vrije en opensource AI-ecosysteem.
- Steun de ontwikkeling van open standaarden.

Inhoud

1. Inleiding

De snelle ontwikkeling rond generatieve kunstmatige intelligentie (GenAI)[1] biedt overheden wereldwijd nieuwe kansen voor dienstverlening aan burgers en bedrijven en het vergemakkelijken van de uitvoering van dagelijkse taken.

Tegelijkertijd maken recente geopolitieke ontwikkelingen steeds duidelijker hoe belangrijk het is om kritisch te kijken naar digitale afhankelijkheden. Het grootste deel van de GenAI modellen wordt momenteel ontwikkeld door niet-Europese, veelal commerciële aanbieders. Ook de Nederlandse overheid maakt op dit moment, voor zover bekend, vooral gebruik van Amerikaanse GenAI modellen.

Om deze afhankelijkheden te verkleinen groeit de aandacht voor open source modellen en Europese alternatieven. Deze modellen bieden potentieel meer transparantie, flexibiliteit en mogelijkheden voor toetsing en aanpassing. Tegelijkertijd is het landschap complex. Wat precies wordt bedoeld met 'open' is een onderwerp van veel discussie, en een breed gedragen definitie voor open source AI modellen ontbreekt nog.

Ook de vraag wanneer een model als 'Europees' kan worden beschouwd is minder eenduidig dan het op het eerste gezicht lijkt: gaat het om de herkomst van het bedrijf, waar de data wordt opgeslagen, hoe het model is getraind of hoe het gebruik ervan wordt aangestuurd en gecontroleerd (governance)? In dit landschap waarin definities van 'open' en 'Europees' onduidelijk zijn, kunnen aanbieders nog veelal zelf bepalen hoe zij hun taalmodellen kwalificeren. Dit schept onduidelijkheid voor gebruikers. Afnemers, zoals overheidsorganisaties, hebben behoefte aan duidelijkheid en duiding van deze complexe kwestie.

Stimuleren van het gebruik van open en Europese taalmodellen bij de Nederlandse overheid

In de Nederlandse Digitaliseringsstrategie (NDS) legt de Nederlandse overheid nadruk op het stimuleren van digitale soevereiniteit en het geven van prioriteit aan open source of Europese toepassingen. Zo ook als het gaat om AI.

Om het gebruik van open en Europese taalmodellen binnen de overheid te bevorderen is het nodig om eerst een duidelijk begrip te hebben van een aantal essentiële concepten, zoals wat taalmodellen zijn bijvoorbeeld 'open source', 'open AI' en 'Europees'. Vervolgens is het belangrijk om criteria en waarden te bepalen die belangrijk zijn bij het maken van keuzes rondom de inzet van deze taalmodellen. Het ministerie van Binnenlandse Zaken en Koninkrijksrelaties heeft, in het kader van de NDS, aan TNO gevraagd om deze verkenning uit te voeren.

Digitale soevereiniteit en open Europese taalmodellen

Het gebruik van open source software kan de afhankelijkheid van commerciële technologie-aanbieders en *vendor lock-ins* verminderen. Bij *vendor lock-ins* kan een gebruiker alleen met veel moeite overstappen naar een andere softwareaanbieder, bijvoorbeeld vanwege hoge overstapkosten of een gebrek aan interoperabiliteit tussen aanbieders. Een dergelijke overstap vereist, naast het gebruik van open of Europese taalmodellen, ook inbedding van deze modellen in software-alternatieven voor commerciële, buitenlandse aanbieders.

Als open-source software wordt ondersteund door een divers en actief ecosysteem, kan dit bijdragen aan Europese soevereiniteit: het biedt een alternatief voor gecentraliseerde platforms van grote buitenlandse techbedrijven en vergroot de controle op digitale infrastructuur en onafhankelijkheid. Bovendien sluit open source aan bij Europese waarden zoals transparantie, samenwerkingen en interoperabiliteit.

Aanpak en leeswijzer

Dit rapport geeft een kort en deels visueel overzicht van de verkenning die TNO heeft gedaan naar de definities van 'open' en 'Europees' rondom taalmodellen, en afwegingen bij de inzet van deze modellen.
De eerste stap is om op basis van literatuur en relevante rapporten een overzicht te maken van belangrijke concepten, definities en verschillende zienswijzen. Dit is uitgewerkt in hoofdstuk 2.

1 In dit rapport spreken we over GenAI als een vorm van AI die zelf inhoud kan maken, zoals teksten of afbeeldingen. Centraal in dit rapport staan taalmodellen. Zie 2.1 voor meer details. Het is belangrijk om op te merken dat dit rapport zich richt op modellen, en niet de toepassingen die daarop worden gebouwd.

Hoofdstuk 3 beschrijft de belangrijkste juridische aspecten, zoals de meest relevante wetgeving, standaardisatie en juridische ontwikkelingen.

In een tweetal workshops met stakeholders van de Nederlandse overheid is een beter begrip ontstaan over de doelstellingen van de overheid op deze thema's, de obstakels die zij momenteel zien voor het kiezen voor open Europese taalmodellen en de afwegingen die voor hen belangrijk zijn om in ogenschouw te nemen. Een uitwerking van deze workshops staat in hoofdstuk 4.

Op basis van de deskresearch en de praktijkworkshops wordt vervolgens een aantal afwegingen geschetst die kunnen helpen bij de keuze voor (open) Europese taalmodellen. Deze staan in hoofdstuk 5.

In hoofdstuk 6 staan de conclusies en aanbevelingen.

2. Begrip van open/ Europese taalmodellen

2.1 Wat zijn 'taalmodellen'?

De term 'taalmodel' (of *language model*) verwijst naar een statistisch model dat gebruikt kan worden om natuurlijke taal te genereren. Grote taalmodellen (*Large Language Models*, LLMs) staan momenteel centraal in het domein van de *Natural Language Processing* (NLP), een deelgebied van AI.

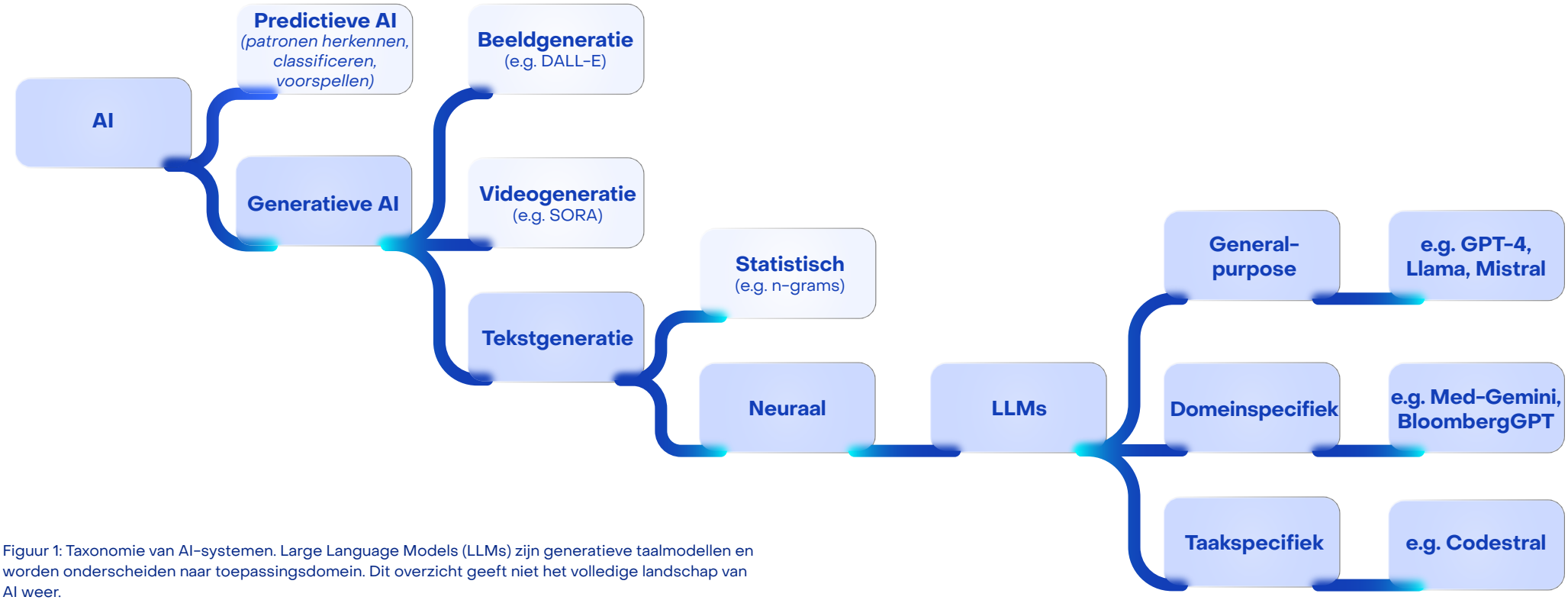
LLMs zijn gebaseerd op neurale netwerken getraind op grote hoeveelheden tekstuele data. LLM's zijn niet alleen goed in klassieke NLP-taken zoals samenvatten en vertalen, maar ook in het genereren van tekst. LLM's vallen dan ook in de categorie generatieve AI (zie Figuur 1).

De levenscyclus van een LLM bestaat grofweg uit drie cyclische fases: pretraining, fine-tuning en inferentie (zie Figuur 3).

Bij de pretraining draait om het trainen van een taalmodel op grote hoeveelheden ruwe tekstverzamelingen van webpagina's, fora, artikelen, maar bijvoorbeeld ook ingescande boeken. Dit geeft het model een brede taalvaardigheid en kennis van de wereld. Pas bij *fine-tuning* leert een model wat een wenselijk antwoord is op een vraag of hoe het een specifieke taak moet uitvoeren. Dit kan door het model instructies te geven met een bijbehorende 'juiste' reactie (*supervised fine-tuning*) of door het model te laten leren van menselijke feedback, met technieken als *reinforcement learning from human feedback* (RLHF). Taalmodellen die deze laatste vormen van *fine-tuning* doorgemaakt hebben bereiken de status van 'AI assistent', waarbij het leren van menselijke feedback met name een belangrijke factor is.

Nadat een model is getraind en gefinetuned, zijn de interne parameters van het model vastgelegd. Deze zogenoemde modelgewichten vormen een wiskundige representatie van de patronen en taalvaardigheid die het model tijdens training heeft opgedaan. Een LLM kan in grootte variëren, meestal uitgedrukt in het aantal parameters, wat bepalend is voor de benodigde rekenkracht en geheugen. Naast het gericht snoeien in de parameters van een LLM kan de precisie waarmee deze parameters worden opgeslagen naar believen worden verlaagd. Beide ingrepen leiden tot compactere en efficiëntere modellen, mogelijk ten koste van nauwkeurigheid.

Moderne LLM's hebben tientallen of zelfs honderden miljarden parameters die gezamenlijk bepalen hoe het model reageert op invoer.



Figuur 1: Taxonomie van AI-systemen. Large Language Models (LLMs) zijn generatieve taalmodellen en worden onderscheiden naar toepassingsdomein. Dit overzicht geeft niet het volledige landschap van AI weer.

Eenmaal getraind kan het model worden ingezet in de gebruiksfase, ook wel 'inferentie' genoemd. Tijdens inferentie ontvangt het model een prompt (bijvoorbeeld een vraag of instructie) en genereert het, op basis van de vaststaande modelgewichten, een passend antwoord. In deze fase leert het model niet meer bij, het past uitsluitend toe wat tijdens pretraining en fine-tuning is aangeleerd.

Reasoning, Chain-of-Thought en Mixture-of-Experts

Via prompt-gebaseerde instructies kunnen LLMs tot redeneren bewogen worden. Dit kan op verschillende manieren.

Chain-of-Thought prompting is een techniek waarbij een model geïnstrueerd wordt om redeneringen op te bouwen met tussenstappen, zodat het model gestructureerder tot een antwoord kan komen. Met *Socratic Prompting* worden modellen aangespoord assumpties systematisch te onderbouwen. Via *Self-Reflection* worden LLMs gedwongen iteratief hun eigen uitvoer kritisch te inspecteren, en te leren van hun eigen oordeel. Vanuit het perspectief van architectuur zien we een ontwikkeling naar gedistribueerde systemen. Bij de *Mixture-of-Experts* (MoE)-benadering wordt een LLM opgesplitst in meerdere gespecialiseerde onderdelen of 'experts', waarvan er per vraag slechts enkele actief worden. Dit maakt modellen efficiënter in gebruik en kan de prestaties verbeteren, zonder dat het model als geheel groter hoeft te worden. Agentic LLMs organiseren meerdere LLMs in een communciatief agent-netwerk, waarbij elke agent door een LLM aangestuurd wordt. Deze netwerken hebben verschillende niveaus van autonomie.

Toepassingen van LLM's

Hoewel de bekendheid van LLM's voornamelijk voortkomt uit publieke chatbots, kent de techniek ook andere toepassingen. Voor interne werkprocessen kunnen deze modellen worden ingezet om tekstuele taken te automatiseren, zoals het samenvatten of categoriseren van documenten en het herschrijven van ambtelijke teksten naar eenvoudiger taalgebruik.

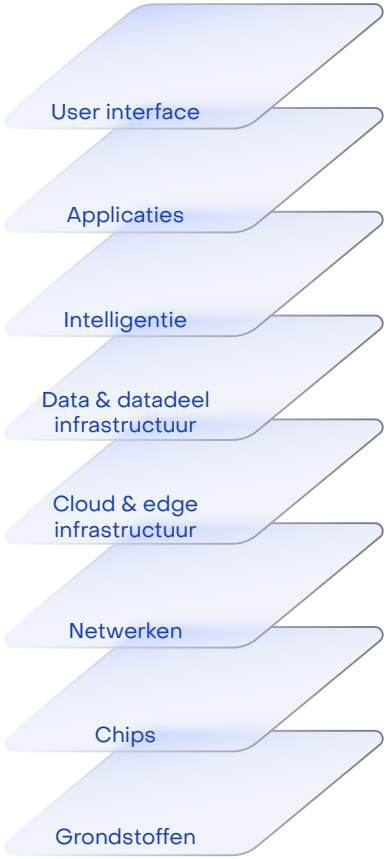
Om de betrouwbaarheid van de gegenereerde output te vergroten, kan een LLM worden gekoppeld aan een specifieke database via Retrieval Augmented Generation (RAG). Hierbij fungeert het model als een interface die informatie direct uit geverifieerde bronnen haalt, citeert en naar de oorspronkelijke documenten verwijst.

Een RAG is intern inzetbaar voor het efficiënt ontsluiten van organisatie-kennis, maar ook extern om burgers of klanten gericht naar de juiste informatie te gidsen.

De digitale stack van taalmodellen

Een LLM is niet los te zien van de bijbehorende digitale stack (Figuur 2). Het stack-model voor digitale technologie laat zien dat deze technologieën zijn opgebouwd uit verschillende, met elkaar verbonden lagen. Het model benadrukt dat één type technologie – zoals een chatbot – is opgebouwd uit meerdere typen technologie. Het model kan gebruikt worden als analytische tool om inzichtelijk te maken hoe een technologie is opgebouwd, uit welke elementen deze bestaan, en om deze elementen te evalueren.

Dit helpt om te begrijpen in welke mate de digitale toepassingen soeverein zijn, bijvoorbeeld dankzij de inzet van open of Europese technologie, en in welke lagen eventuele afhankelijkheden blijven bestaan.



Figuur 2: Digitale technologieën zijn opgebouwd uit diverse lagen, ook wel 'de stack' genoemd. Het stack model voor digitale technologie geeft dit overzichtelijk weer. (Aangepast van: Stolwijk et al. (2024) Towards a sovereign digital future – the Netherlands in Europe, TNO)

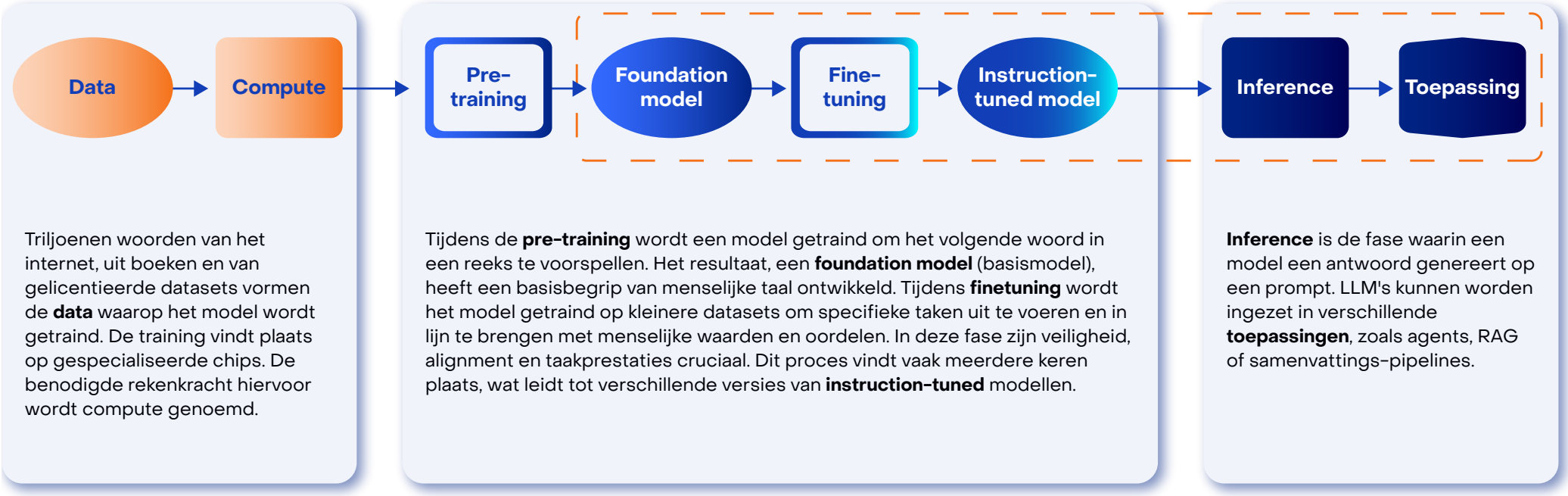
De digitale stack van applicaties zoals chatbots bestaat naast het taalmodel ook uit de data waarop het taalmodel is getraind en de cloud- of edge-infrastructuur waar data wordt opgeslagen en beschikbaar gemaakt en/of waarmee het AI-model (via API's) in gebruik kan worden genomen (als AI-as-a-service, AIAAS).

Een ander belangrijk onderdeel van de digitale stack voor AI zijn de chips. Voor het trainen en fine-tunen van taalmodellen zijn de standaard central processing units (CPU's) die in de meeste computers te vinden zijn niet voldoende, maar zijn krachtige chips nodig, zoals graphics processing units (GPU's). Tot slot is er de laag van grondstoffen die nodig zijn om de benodigde hardware te maken en/of draaiende te houden, zoals zeldzame aardmetalen en het verbruik van water en energie.

Als het gaat om het bevorderen van digitale soevereiniteit is alleen de keuze voor een open source of Europees model dus niet voldoende. Zo kan een model van *Mistral* ook gebruikt worden via de *Microsoft Azure Foundry*. Dan is het model wel Europees, maar draait het op een Amerikaans serviceplatform.

Door naar de hele digitale stack te kijken – en niet alleen het model - wordt het inzichtelijk welke afhankelijkheden er met een bepaalde keuzes blijven bestaan en welke kunnen worden verminderd.

Levenscyclus van een LLM



Figuur 3: Levenscyclus van een LLM.

Hoe worden LLM's geëvalueerd?

Een cruciaal onderdeel van de inzet van LLM's is de systematische beoordeling van hun prestaties door middel van evaluatie. Dit vindt zowel intrinsiek plaats, bijvoorbeeld door het meten van nauwkeurigheid, consistentie en feitelijkheid, als extrinsiek, door te testen hoe het model zich gedraagt binnen een taak of een bepaald werkproces.

Naast specifieke technische meetopstellingen en metrieken die LLM's intern doormeten, zijn breed geaccepteerde vergelijkingskaders zoals leaderboards en benchmarks belangrijk om inzicht te geven in hoe modellen presteren op allerlei taken. Een voorbeeld hiervan is EuroEval, een leaderboard met benchmarks voor een Europese context. Het is belangrijk om deze evaluaties aan te vullen met evaluatietechnieken die afgestemd worden op specifieke toepassingen of werkproces. Toetsingscriteria hiervoor kunnen worden afgestemd op publieke waarden zoals veiligheid, non-discriminatie en uitlegbaarheid.

Op dit moment bestaan er echter geen (Europese) technische standaarden en zijn er ook geen onafhankelijke methoden om LLM's te beoordelen. Er zijn initiatieven gaande om Europese technische normen vast te stellen.

Zorgelijk gedrag van LLM's

LLM's worden door (big tech-) bedrijven, onderzoeksinstellingen en individuele onderzoekers in allerlei fases van ontwikkeling publiekelijk beschikbaar gemaakt op platformen als Hugging Face. Het is vaak onduidelijk aan welke trainingsvormen en -data deze modellen zijn blootgesteld. Dit leidt tot reële risico's. Zo kunnen modellen op toxische, racistische of anderszins schadelijke data zijn getraind. Deze invloeden zijn moeilijk achteraf te herkennen in deze modellen, waarbij het ook lastig is dat onderzoekers en toezichthouders niet precies weten waar naar ze zoeken. Dit kan leiden tot modellen die onvoorspelbaar en ongewenst gedrag vertonen. LLM's zijn van nature stochastisch: ze geven op dezelfde vraag (prompt) regelmatig verschillende antwoorden (output), en zelfs een klein verschil in formulering kan tot heel andere output leiden. Ook zijn LLM's - na *fine-tuning* op menselijke oordelen - soms geneigd tot 'sycofantisch' gedrag: gedrag dat gericht is op het plezieren van de menselijke gebruiker. Dit is een directe consequentie van *reinforcement learning*: de modellen leren in die trainingsfase een strategie aan die gericht is op het verkrijgen van een maximale beloning. In combinatie met de normen en waarden die ze in die fase ook leren kan dit leiden tot '*alignment faking*.' Daarbij doet een model alsof een nieuwe taakinstructie die niet in overeenstemming is met hun ethische waarden wel kan uitvoeren. Daarmee voorkomt het model dat het opnieuw wordt getraind. Dit zogeheten '*dark pattern*' (misleidende ontwerpkenmerken) brengt modellen ertoe om buiten hun morele en ethische '*guardrails*' (veiligheidsmaatregelen) te treden.

Bias en ethiek

LLM's kunnen verschillende soorten bias vertonen. *Selectiebias* ontstaat doordat het grootste deel van de trainingsdata niet gemodereerd en gecureerd is. Daardoor is het onvermijdelijk dat er bij het trainen van modellen op die data allerlei culturele, sociale, ethische en politieke biases de modellen binnensluipen. Een technische *algoritmebias* ontstaat door de selectieve aandacht die LLM's besteden aan dialoogcontext. *Instructiebias* onstaat door de bias in de voorbeelddata tijdens de *fine-tuning* met instructies. De laatste fase van het LLM trainingsproces, het leren uit feedback van menselijke beoordelaars, kan ook een bron zijn van bias. Voor een model als ChatGPT wordt door grote hoeveelheden personeel op diverse locaties in de wereld feedback gegeven op uitvoer van het model, waarna het model periodiek gefinetuned wordt. Dit betekent dat heterogene ethische waardenkaders en sociale en culturele achtergronden betrokken zijn bij het kwalitatief beoordelen van LLM uitvoer. Dat deze trainingsfase kan leiden tot ongewenste uitkomsten werd in april 2025 duidelijk, toen OpenAI een versie van ChatGPT-4o terugtrok omdat het model te 'sycofantisch' was: te zeer gericht op het plezieren van gebruikers.

Mitigatie van bias

Eén van de manieren om ongewenste bias te mitigeren is via guardrails: regelgebaseerde instructies die LLM's dwingen tot bepaald gewenst gedrag, gebruikerssentiment detecteren of LLM-uitvoer regelgebaseerd herschrijven. Het is aangetoond dat guardrails op diverse manieren omzeild kunnen worden. Detectie van operationele waarden en het vaststellen van ongewenste bias binnen LLM's vereisen technische en normatieve methoden en kaders. Kwalitatieve beoordeling veronderstelt een referentiekader van ethiek, waarden en bias.

Openheid is hiervoor een belangrijke voorwaarde. Het geeft inzicht in wie die waarden vaststelt, wie bepaalt of en in welke mate ze aanwezig zijn in LLM-uitvoer. Ook biedt het de mogelijkheid om na te gaan hoe open en neutraal de guardrails zijn die de industrie implementeert in hun modellen. En hoe gemakkelijk de guardrails, al dan niet opzettelijk, te omzeilen.

Niet alleen technisch

De eigenschappen van LLM's – hun gevoeligheid voor de kwaliteit en aard van trainingsdata, de invloed van menselijke feedback op waarden en gedrag, en de risico's rond bias en onvoorspelbaar gedrag - maken duidelijk waarom de keuze voor een taalmodel geen puur technische beslissing is. Inzicht in trainingsdata, modelarchitectuur en guardrails stelt gebruikers in staat om risico's te beoordelen, bias te detecteren en waarden te toetsen. Tegelijkertijd laat de afhankelijkheid van de onderliggende digitale stack, van GPU-hardware tot cloudinfrastructuur, zien dat de keuze voor een bepaalde LLM in de context van digitale soevereiniteit verder gaat dan de kwaliteit van de LLM zelf.

2.2 Wat zijn open taalmodellen?

De term 'open' is het onderwerp van groeiende discussie onder AI-ontwikkelaars, onderzoekers en beleidsmakers. De termen 'open' en 'open source' worden regelmatig door elkaar gebruikt. Waar open-source in de softwarewereld een relatief vaste betekenis heeft, heeft de term in de context van AI ook een juridische dimensie onder de AI Act. 'Open' wordt steeds vaker omschreven als een *moving target*: een begrip dat voortdurend verschuift en lastig eenduidig te definiëren valt, stellen onderzoekers van de Radboud Universiteit.

'Open' in de AI Act

De AI Act erkent de waarde van open-source als bron van innovatie en biedt 'free and open-source' (FOS) AI-modellen voor algemene doeleinden en AI-systemen onder bepaalde voorwaarden een uitzonderingspositie.

Het blijft mogelijk dat een open-source AI-model of systeem toch in een hoog-risicocategorie valt of tegen betaling aangeboden wordt en dus niet van de uitzondering kan profiteren.

Volgens data van Epoch AI zijn er op dit moment enkele tientallen systemen die kwalificeren als AI-modellen voor algemene doeleinden. Wat precies onder de AI Act-uitzonderingen valt, is nog onduidelijk. Onderzoekers verwachten dat interpretatieverschillen ertoe leiden dat de juridische afbakening van 'open source' een belangrijk aandachtspunt wordt voor modelontwikkelaars.

'Open-washing'

Onderzoekers aan de Radboud Universiteit stellen dat LLM-ontwikkelaars zich regelmatig schuldig maken aan 'open washing': punten binnenslepen die openheid aangeven, zonder echt belangrijke informatie te ontsluiten die wetenschappelijk onderzoek en juridische toetsing mogelijk maken. Dit, stellen de onderzoekers, staat innovatie, wetenschappelijke toetsing, en publiek begrip van AI in de weg.

Om meer duidelijkheid te scheppen in wat 'openheid' betekent in de context van AI zijn verschillende raamwerken uitgewerkt. Twee daarvan, het Model Openess Framework en het European Open Source AI Index, staan centraal in deze ontwikkelingen. We gaan dieper in op deze raamwerken.

Intermezzo: Wat is 'open-source'?

Voordat dieper in wordt gegaan op het spectrum van 'openheid' als het gaat om AI is het goed om eerst stil te staan bij definities van open source.

Free and Open Source Software (FOSS)

De Free Software Foundation (FSF) verwijst naar *free software* als *software* die de rechten en vrijheid van gebruikers en gemeenschappen respecteert. *Free software* moet volgens FSF aan vier essentiële criteria voldoen. Daarmee wordt bedoeld dat gebruikers de software naar eigen inzicht mogen gebruiken, kopiëren, inzien/inspecteren, aanpassen en verder verspreiden (inclusief tegen betaling). Toegang tot de broncode is een voorwaarde.

Grotendeels vergelijkbaar met de definitie van FSF, de Open Source definitie (OSD) van de Open Source Initiative (OSI) bestaat uit een tiental vereisten, die samengevat kunnen worden als volgt:

De OSD bepaalt dat royaltyvrije herdistributie moet toestaan en toegang tot de broncode in een vorm die bewerken mogelijk maakt. Licenties moeten wijzigingen en afgeleide werken toestaan en mogen niet discrimineren naar personen, groepen of toepassingsdomeinen. De rechten moeten voor alle ontvangers gelden, los van de distributievorm of het product, mogen geen beperkingen opleggen aan andere meegeleverde software en zijn technologieneutraal.

Open Source AI Definition

Het OSI werkt momenteel aan een eerste versie van een definitie voor AI systemen: de Open Source AI Definition (OSAID), nauw verwant aan de FSF definitie van free software. Deze staat nu open voor consultatie. Samengevat:

Een AI-systeem moet beschikbaar gemaakt worden met de vrijheid om te gebruiken voor elk doeleinde zonder toestemming, te onderzoeken hoe het systeem en zijn componenten werkt, het systeem aan te passen voor elk doeleinde, en te delen met of zonder aanpassingen voor elk doeleinde.

Model Openness Framework

Het Model Openness Framework (MOF), opgesteld door de Linux Foundation, dient om AI modellen in drie categorieën in te delen in openheid. MOF III, de laagste categorie, stelt onder andere als eis dat het model architectuur en de modelgewichten openbaar moeten zijn.

MOF II en MOF I vereisen openbaring van o.a. de code en de trainingsdata, respectievelijk.

Het MOF maakt inzichtelijk wat grofweg de driestap is van open modellen: 1) de gewichten, 2) de code, en 3) de trainingsdata. Veel modellen vallen natuurlijk buiten de drie MOF klassen. Deze kunnen doorgaans als gesloten modellen beschouwd worden.

European Open Source AI Index

Het European Open Source AI Index (EOSAI) vormt een aanvulling op het MOF. Zonder drie klassen te hanteren, geeft het EOSAI een uitgebreid overzicht van de aspecten van modellen waar informatie over te vinden is .

Richting één framework

Gebaseerd op de MOF en EOSAI, kan men openheid van LLM's indelen in drie categorien: closed source, open-weight (en aanverwant), en fully open-source. Zie dit overzicht voor de indeling. In de praktijk blijkt dat closed-source en open-weight het meeste voorkomt.

	Apertus (Swiss AI)	DeepSeek R1 (DeepSeek)	GROK 2 (xAI)
Beschikbaarheid			
Base Model Data	✓	✗	✗
Eindgebruiker Model Data	✓	✗	✗
Base Model Weights	✓	✓	✗
Volledige Model Weights	✓	✓	✓
Training Code	✓	✗	✗
Documentatie			
Code documentatie	✓	✗	✗
Hardware	✓	~	✗
Preprint	✓	✓	✗
Paper	✗	✗	✗
Model card	✓	~	✗
Data sheet	✓	✗	✗
Toegang			
Package	✓	✓	✗
API	✓	✓	✗
Licenties	✓	✗	✗

Gebaseerd op European Open-Source AI Index (Liesenfeld & Dingemanse, 2024)

Hoe 'open' kan een taalmodel zijn?

Closed Source	Open Code	Open Weights	Open Research	Fully Open Source
· **Open**: API · **Gesloten**: model-gewichten, trainingsdata, code, architectuur · **Gebruik**: API-gebaseerd	· **Open**: Trainingscode, model-gewichten onder voorwaarden · **Gesloten**: (deel van de) data · **Gebruik**: Zelf een model trainen en/of draaien	· **Open**: Model-gewichten · **Gesloten**: trainingsdata en -code · **Gebruik**: model zelf draaien en fine-tunen	· **Open**: Model-gewichten + Code · **Gesloten**: (deel van de) data · **Gebruik**: redelijk reproduceren van het model, zelf draaien en fine-tunen	· **Open**: Alles, inclusief alle trainingsdata · **Gebruik**: volledige reproduceer-baarheid
GPT-5, Claude Opus, Gemini 3.1	Phi-4	Llama 4, Gemma 3, Mistral 7B	DeepSeek R1, EuroLLM	OLMo, Apertus, BLOOMZ

Gebaseerd op: *European Open-Source AI Index (Liesenfeld & Dingemanse, 2024)* en *The Model Openness Framework*. De categorieën Closed Source en Open Weights komen in de praktijk het vaakst voor.

2.3 Wat zijn voor- en nadelen van open taalmodellen?

Onderstaande voor- en nadelen gelden in het algemeen voor het gebruik van open taalmodellen, en zijn niet specifiek voor de inzet van open taalmodellen door overheden.

Voordelen	Nadelen
Transparantie Openheid maakt het mogelijk om het gedrag en de capaciteiten van het model te analyseren. Dat biedt ook de mogelijkheid om problemen zoals bias of bugs op te sporen en aan te passen. Ook kunnen problemen met de beveiliging snel opgespoord en verholpen worden.	**Veiligheid** Transparantie maakt het eenvoudiger voor kwaadwillende personen om kwetsbaarheden uit te buiten of safeguards te verwijderen. Openbare modelgewichten kunnen worden misbruikt voor het genereren van schadelijke content zonder ingebouwde beperkingen.
Reproduceerbaarheid Doordat modelarchitecturen, trainingsdata en -processen inzichtelijk zijn, kunnen resultaten onafhankelijk worden geverifieerd en gerepliceerd. Dit versterkt wetenschappelijke validatie en maakt het mogelijk om claims over modelprestaties objectief te toetsen.	**Prestaties** Volledig open source modellen, zoals Apertus, presteren nog niet op het niveau andere niet-volledig open modellen. Open Weight modellen, zoals Deep Seek of andere Chinese modellen presteren minstens, zo niet beter, als gesloten modellen als ChatGPT of Claude, volgens beschikbare Benchmarks.
Innovatie Open taalmodellen dragen bij aan een onderzoeksecosysteem waarin wordt samengewerkt door verschillende partijen. Openheid van modelarchitecturen, trainingsdatasets en algoritmische processen maakt iteratieve ontwikkeling en grondige externe validatie mogelijk.	**Juridische onzekerheid** Onzekerheid rondom juridische kaders, toezichtmechanismen en verantwoordingsstructuren maken het moeilijk om aansprakelijkheid vast te stellen en is het onduidelijk wie verantwoordelijk is als het fout gaat.
Inclusiviteit en diversiteit Een grotere groep – waaronder academische, industriële en onafhankelijke onderzoekers - kan bijdrage aan de ontwikkeling van modellen. Dit bevordert bijvoorbeeld meertaligheid en culturele representatie, wat belangrijk is voor modellen die Europese talen en contexten moeten bedienen.	**Governance** Het ontbreken van een centrale beheerder maakt het lastig om governance in te richten: wie beslist over updates, wie bewaakt ethische richtlijnen, en wie coördineert bij incidenten?
Context-specifieke fine-tuning Organisaties kunnen open modellen aanpassen aan hun specifieke domein, taal of use case zonder afhankelijk te zijn van een externe leverancier. Dit maakt het mogelijk om modellen te optimaliseren voor bijvoorbeeld Nederlandse juridische teksten, medische dossiers of overheidscommunicatie.	
Onafhankelijkheid en keuzevrijheid Open modellen verminderen de afhankelijkheid van individuele (vaak niet-Europese) leveranciers. Ze bieden keuzevrijheid in hosting, aanpassing en doorontwikkeling en voorkomen vendor lock-in, prijsverhogingen of plotselinge beëindiging van diensten. Bovendien bevordert het de ontwikkeling van eigen expertise.	

2.4 Wanneer is een taalmodel Europees?

Wanneer een taalmodel als Europees kan worden beschouwd, is minder eenduidig dan vaak wordt aangenomen. Het land van herkomst van de ontwikkelaar is een belangrijke eerste indicator. Europa is niet de grootste producent van LLM's. Dat zijn China en de Verenigde Staten (Figuur 4). Epoch AI becijferde dat van de sinds 2025 uitgebrachte LLM's, 39% van Amerikaanse makelij is, gevolgd door modellen van Chinese organisaties met iets meer dan 37%. Slechts 8% heeft een Europese oorsprong. Deze komen voornamelijk uit Frankrijk, Zwitserland en Duitsland. Deze cijfers geven een beeld van waar taalmodellen worden ontwikkeld, maar vertellen niet het hele verhaal. Om elementen als technologische soevereiniteit en publieke waarden mee te nemen in de afweging, is een bredere definitie nodig dan enkel land van herkomst. Een AI-model kan op basis van grofweg drie onderling verbonden dimensies als Europees worden beschouwd: herkomst en governance van de ontwikkelaar, data soevereiniteit en infrastructuur en training op Europese talen en waarden.

Herkomst en governance van de ontwikkelaar(s)

Het model is ontwikkeld, beheerd of substantieel aangestuurd door een organisatie die binnen Europa is gevestigd en onder Europees recht valt. Besluitvorming over ontwikkeling, updates en risicobeheersing vindt plaats binnen EU juridische en normatieve kaders.

Datasoevereiniteit en infrastructuur

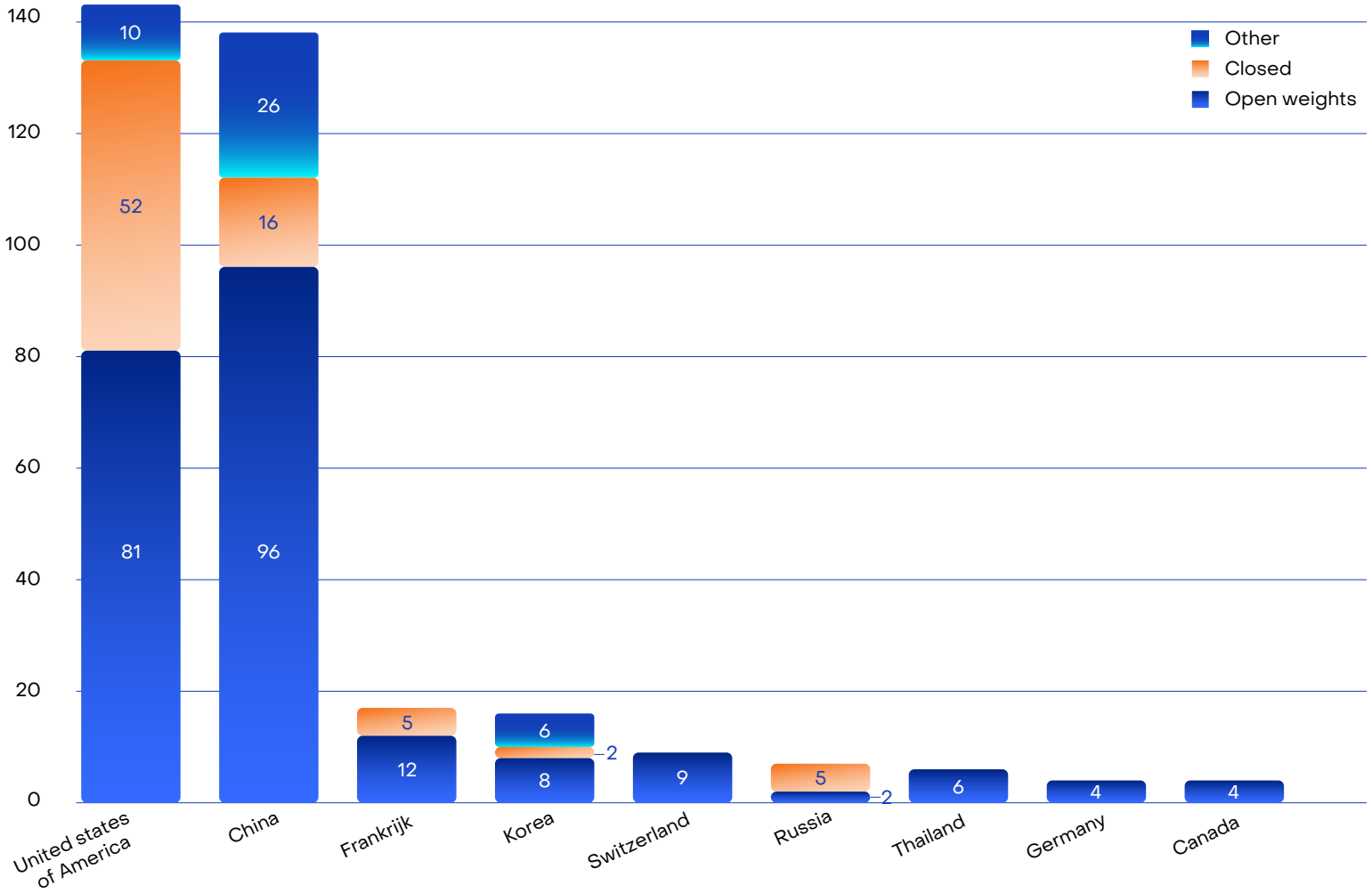
De herkomst, opslag en verwerking van de (pre)trainingsdata moeten voldoen aan alle geldende EU wet- en regelgeving. Met het oog op datasoevereiniteit worden datasets idealiter grotendeels binnen Europa verzameld of beheerd, en draait het model op Europese infrastructuur of bij gecertificeerde aanbieders die aan EU-regels voldoen. Zo niet, dan kan de gehele toepassing worden beschouwd als minder Europees.

Training op Europese talen en waarden en ontwikkelingskeuzes

Belangrijk zijn niet alleen de waarden die het ontwikkelingsproces sturen, maar ook de waarden die in het model verweven zitten. Een model is getraind op datasets die in meerdere of mindere mate Europese talen bevatten. Trainingsdata in Europese talen bepalen niet alleen de taalkundige vaardigheden van een model, maar kunnen ook de culturele waarden die in die talen besloten liggen meenemen. Ook zijn ontwikkelingskeuzes hierin belangrijk. Welke data is uitgekozen, hoe een model *aligned* is en wat voor *fine-tuning* heeft plaatsgevonden hebben allemaal invloed op de culturele waarden die een model bevat.

Europa en open source
Europa heeft een lange traditie in de ontwikkeling van vrije en open source software, en heeft nu een goede uitgangspositie voor de ontwikkeling van open source basismodellen. Europa heeft een goede wetenschappelijke, technische en institutionele infrastructuur. Daarnaast investeert de Europese Unie in diverse initiatieven om infrastructuur en *compute* capaciteiten te verbeteren, zoals AI Fabrieken en het OpenEuroLLM project.

Europees is geen garantie
De keuze voor een (commerciële) Europese oplossing is geen garantie voor het behouden van digitale soevereiniteit, omdat er altijd een reële kans bestaat op overname of substantiële of koersbepalende investeringen door een buitenlandse speler. Europese ondernemers zoeken vaak financiering in de Verenigde Staten. Tussen 2008 en 2021 zijn 30% van alle Europese 'unicorns' verhuisd naar niet-Europese landen, de meerderheid naar de VS.



Figuur 4: Aantal LLM's per land van ontwikkelaar. De dataset ('All AI models'), is samengesteld door Epoch AI (Data on AI Models | Epoch AI) en geüpdatet op 19 maart 2026.

3. Juridisch kader

3.1 Relevante wetgeving

Bij de (mede)ontwikkeling en het gebruik van LLM's moet rekening worden gehouden met een complex stelsel aan toepasselijke wet- en regelgeving. Dit kan variëren van regels omtrent AI, privacy- en gegevensbescherming, bescherming van intellectueel eigendom en bedrijfsgeheimen, veiligheid en beveiliging, aansprakelijkheid, data (governance), cloud etc. Aangezien het landschap nog relatief nieuw en deels nog volop in ontwikkeling is, kan dit extra complexiteit met zich meebrengen wat betreft de interpretatie, op elkaar afstemmen en toepassing van wet- en regelgeving. Enkele relevante wetgevingsbesluiten en -initiatieven worden op de volgende slides kort toegelicht. Voor zover beschikbaar, worden de bepalingen die betrekking hebben op de EU- en vrije en opensource dimensies van LLM's benadrukt.

Het is belangrijk om vooraf enkele aspecten te verduidelijken:

- De categorie AI-systemen, met inbegrip van AI-modellen voor algemene doeleinden (bijv. taalmodellen), behelst naast software, ook gegevens, parameters en wegingen;
- De EU wet- en regelgeving hanteert dezelfde definitie van vrije en opensource (AI) als de Free Software Foundation en de Open Source Initiative. Er wordt echter een belangrijk onderscheid gemaakt tussen enerzijds AI-systemen en -modellen die gratis ontwikkeld en beschikbaar gesteld worden en anderzijds AI-systemen en -modellen die tegen betaling (in geld, data, persoonsgegevens, etc.) aangeboden worden. AI-systemen en -modellen die gratis aangeboden worden vallen meestal buiten de scope van de wetgeving.
- Een AI-systeem wordt niet uitzonderlijk, maar in combinatie met zijn oorsprong en toeleveringsketen (*supply chain*) beoordeeld om risico's te kunnen evalueren en relevante verplichtingen te bepalen.

AI-Verordening (AI Act)
Verordening (EU) 2024/1689 Artificiële intelligentie

De AI-verordening stelt de geharmoniseerde regels vast voor het bevorderen van mensgerichte, betrouwbare, veilige artificiële intelligentie die gebaseerd is op fundamentele rechten, waarden en normen; en voor het bevorderen van innovatie. Taalmodellen, als een van de types van artificiële intelligentie, vallen binnen de scope van deze wet.

- **Inwerkingtreding en toepassing**: de EU AI-verordening is in augustus 2024 in werking getreden en wordt vanaf 2 augustus 2026 van toepassing. Enkele bepalingen zijn al vanaf 2025 van kracht (bijv. over het onderwerp en de toepassingsgebieden van de wet; verboden AI-praktijken; regels omtrent AI-modellen voor algemene doeleinden; governance van AI; sancties en handhavingsmaatregelen). De Nederlandse uitvoeringswet AI-verordening bevind zich momenteel in de consultatiefase.
- **Europese AI-taalmodellen**: De AI-verordening maakt geen onderscheid tussen EU en non-EU aanbieders van AI-systemen, inclusief taalmodellen. De wet is van toepassing op alle aanbieders en gebruiksverantwoordelijken indien ze AI-systemen in de EU in handel brengen, of in gebruik stellen, of de output daarvan in de EU wordt gebruikt.
- **Open-source AI-taalmodellen**: niet-commerciële AI-systemen en -modellen (software, gegevens, modellen, parameters, wegingen), met inbegrip van taalmodellen, die worden vrijgegeven onder vrije en opensource licenties (FOSAI) vallen buiten de reikwijdte van de AI-verordening, zolang het niet commercieel . Dit zou als een voordeel kunnen worden beschouwd ten opzichte van het gebruik van commerciële alternatieven. Tegelijkertijd beperkt het de mogelijkheid om de makers aansprakelijk te stellen of boetes te laten betalen voor fouten in AI-systemen.

Let op: Deze uitzondering geldt niet voor AI-taalmodellen met een hoog risico en ook niet of voor sommige AI-systemen die voor directe interactie met natuurlijke personen zijn bedoeld. In deze gevallen is de Verordening wel van toepassing.

Verordening cyberweerbaarheid

Verordening (EU) 2024/2847 betreffende horizontale cyberbeveiligingsvereisten voor producten met digitale elementen

De EU Verordening cyberweerbaarheid (CRA) is een uitgebreid stelsel maatregelen bedoeld om de cyberbeveiliging van digitale producten (hard- en software) die op de markt worden gebracht en gedurende hun hele levenscyclus te verbeteren. De CRA is slechts één van de pakketten aan maatregelen die beoogt de cyberveiligheid te verbeteren. Andere relevante wet- en regelgeving omvat de Cyberbeveiligingsverordening van 2019 en de de Cyberbeveiligingswet (NIS2-richtlijn) van 2022. De Europese Commissie heeft onlangs een voorstel gedaan voor de herziening van het cybersecurity-pakket.

- **Inwerkingtreding en toepassing**: de EU Verordening cyberweerbaarheid (CRA) is in oktober 2024 in werking getreden en zal vanaf 11 december 2027 van toepassing worden. Enkele bepalingen zullen eerder, vanaf 2026 van kracht zijn (bijv. over de rapportageverplichtingen van fabrikanten). De Nederlandse uitvoeringswet verordening cyberweerbaarheid bevindt zich momenteel in de voorbereidende fase voor behandeling in de Tweede Kamer.
- **Europese AI-taalmodellen**: De CRA maakt geen onderscheid tussen EU en non-EU aanbieders van AI-systemen, inclusief taalmodellen. De wet is van toepassing op alle aanbieders en gebruiksverantwoordelijken indien ze AI-modellen of AI-systemen in de EU in handel brengen, of in gebruik stellen, of de output daarvan in de EU wordt gebruikt.
- **Open-source AI-taalmodellen**: niet-commerciële AI-modellen en systemen (software, gegevens, modellen, parameters, wegingen), met inbegrip van taalmodellen, die worden vrijgegeven onder vrije en opensource licenties (FOSAI), alsook software-as-a-service zijn grotendeels niet onderhevig aan de CRA. Dit zou als een voordeel kunnen worden beschouwd ten opzichte van het gebruik van commerciële alternatieven. Tegelijkertijd beperkt het de mogelijkheid om de makers aansprakelijk te stellen of boetes op te leggen voor fouten in AI-modellen of AI-systemen.

Let op: Deze uitzondering geldt niet voor AI-taalmodellen met een hoog risico en ook niet of voor sommige AI-systemen die voor directe interactie met natuurlijke personen zijn bedoeld. In deze gevallen is de Verordening wel van toepassing.

Richtlijn (EU) 2024/2853 inzake aansprakelijkheid voor gebrekkige producten

De EU Richtlijn 2024/2853 is een bijwerking van de productaansprakelijkheidsregels van de Europese Unie. Het bevat geharmoniseerde regels inzake de aansprakelijkheid voor schade die natuurlijke personen lijden als gevolg van producten met gebreken. Onder producten worden in deze Richtlijn ook software en AI-systemen inbegrepen. Hiermee wordt de aansprakelijkheid van commerciële aanbieders van dergelijke producten significant uitgebreid. Dit betekent ook een **fundamentele verandering** ten opzichte van de huidige situatie en er ontstaat ook een fundamenteel verschil met het regime dat in andere jurisdicties geldt (voor bijv. VS- software en AI-systemen).

- **Inwerkingtreding en toepassing**: de EU Richtlijn Aansprakelijkheid voor gebrekkige producten is in 2024 in werking getreden. In Nederland is de Richtlijn nog niet (volledig) geïmplementeerd. De verwachte ingangsdatum van de Implementatiewet richtlijn herziening productaansprakelijkheid is 9 december 2026.
- **Europese AI-taalmodellen**: De Richtlijn maakt geen onderscheid tussen EU en non-EU aanbieders van AI-systemen, inclusief taalmodellen. De wet is van toepassing op alle aanbieders en gebruiksverantwoordelijken indien ze AI-modellen en -systemen in de EU in handel brengen, of in gebruik stellen, of de output daarvan in de EU wordt gebruikt.

Let op: Door de omzetting van de Richtlijn in nationale wetgeving kunnen verschillen in interpretatie, implementatie en naleving tussen de EU-landen ontstaan.

- **Open-source AI-taalmodellen**: niet-commerciële AI-systemen (software, gegevens, modellen, parameters, wegingen), met inbegrip van taalmodellen, die worden vrijgegeven onder vrije en opensource licenties (FOSAI), zijn grotendeels niet onderhevig aan deze Richtlijn.

Andere relevante ontwikkelingen met betrekking tot wetgeving

Zoals eerder vermeld, zijn LLM's onderhevig aan een complex stelsel aan toepasselijke wet- en regelgeving. Het betreft echter een landschap dat nog volop in ontwikkeling is. Nieuw initiatieven zullen nog meer complexiteit toevoegen aan de interpretatie, op elkaar afstemmen en toepassing van wet- en regelgeving. Enkele voorbeelden van dergelijke initiatieven waar rekening moet worden gehouden:

- Het EU-wetsvoorstel Cloud and AI Development Act inzake de ontwikkeling van cloud computing en kunstmatige intelligentie (AI). Het wetsvoorstel zal o.a. relevant zijn voor kritieke toepassingen in de publieke sector waar veilig gebruik moet worden gemaakt van in de EU gevestigde AI- en clouddiensten.
- De Digitale Omnibus is een pakket wetsvoorstellen van de Europese Commissie dat beoogt de vereenvoudiging van het digitale wetgevingskader. Hiermee zullen meerdere digitale wetten gelijktijdig en ingrijpend worden aangepast, waaronder de AI-verordening, de Algemene verordening gegevensbescherming (AVG)/GDPR, (AVG), Verordening (EU) 2018/172 tot oprichting van één digitale toegangspoort voor informatie, procedures en diensten voor ondersteuning en probleemoplossing, de Dataverordening, de ePrivacyrichtlijn, de Cyberbeveiligingswet (NIS2-richtlijn). Relevante wijzigingen betreffen onder anderen: de definitie van persoonsgegevens; de toepasbaarheid van gerechtvaardigd belang als rechtsgrond voor het verwerken van persoonsgegeven onder de AVG; de criteria voor de classificatie van AI-systemen als risicovol; de rol van het databankrecht sui generis, etc.

3.2 Standaardisatie

Standaardisatie, en met name open standaarden, spelen een belangrijke rol in het bevorderen van innovatie, interoperabiliteit en adoptie van technologieën door organisaties en kwaliteitsborging.

De ontwikkeling van AI-standaarden is echter gefragmenteerd en bevindt zich nog in een prille fase. In sommige gevallen vervangen AI-benchmarks (Sectie 2.1 en 3.2) de rol van formele standaarden. Dit kan ongewenste effecten hebben door het gebrek aan een brede consensus, onafhankelijkheid en de onevenredige invloed dat een klein aantal (VS) AI-bedrijven kan uitoefenen op inhoud, keuzes en richting van de technologie.

Er is een groeiend besef van het belang van AI-standaarden, met name in relatie tot de potentiële (nationale) veiligheidsrisico's die daarmee gepaard zijn. In de Verenigde Staten, bijvoorbeeld, zullen frontier AI-modellen getest worden door het Center for AI Standards and Innovation (CAISI), een afdeling van het V.S. Ministerie van Handel, voordat ze openbaar kunnen worden gemaakt.

- **Europese AI-standaardisatie**: De standaardisatieorganisaties van de Europese Unie hebben een aantal initiatieven opgezet om breed gedragen vrije Europese normen voor AI te ontwikkelen, waaronder geharmoniseerde normen ter ondersteuning van de EU-AI-Verordening. Bijvoorbeeld, CEN-CENELEC ontwikkelt in de gezamenlijke Technische Commissie 21 (JTC 21) algemene normen voor betrouwbare AI, risico- en kwaliteitsbeheer en conformiteitsbeheer; en specifieke normen voor AI (cyber) veiligheid, robuustheid, datasets en natuurlijke taalverwerking. Sinds eind 2025 is de ontwikkeling van deze normen een versneld traject ingegaan.
- **Open standaarden AI**: Er zijn ook andere internationale initiatieven om open (publiekelijk beschikbare) standaarden te ontwikkelen, bijvoorbeeld die van ITU (het gespecialiseerde agentschap van de Verenigde Naties voor digitale technologieën) of van de Linux Foundation. Een voorbeeld is het Model Openness Framework (MOF), (Sectie 2.2) een systeem voor het beoordelen en classificeren van de openheid van machine learning-modellen ontwikkeld door de Generative AI Commons van de Linux Foundation. Een ander voorbeeld is de samenwerking tussen de Linux Foundation en de Europese standaardisatieorganisatie CEN-CENELEC op het gebied van AI en cybersecurity.

In Nederland geldt het 'Pas toe of leg uit'-beleid. Dat houdt in de verplichting voor rijk, gemeenten, uitvoeringsorganisaties etc. om open standaarden toe te passen.

3.3 Risico's

Bepaalde wetten, of bepalingen daarvan, kunnen een risicogebaseerde aanpak vereisen. Dit houdt in dat risico's en de te nemen (mitigerende) maatregelen in verhouding moeten zijn. Het correct in kaart brengen van risico's en het interpreteren van de gevolgen daarvan is daarom van essentieel belang. Bij het bepalen van risico's moet er rekening worden gehouden met de gehele levenscyclus en toeleverings-keten van AI-systemen. Algemeen wordt aangenomen dat AI-systemen bestaande risico's vergroot en nieuwe risico's met zich meebrengen die nog niet volledig worden begrepen. De meningen zijn verdeeld over de vraag of open source de risico's vermindert of juist vergroot. Het wordt aanbevolen om het functioneren van AI-systemen regelmatig te monitoren.

In de volgende tabel geven we enkele voorbeelden van risico's die relevant zijn voor AI-systemen. Deze zijn onderverdeeld in drie categorieën: risico's op het gebied van gegevens-bescherming (AVG), securityrisico's (AVG & CRM) en technische risico's (LLM's & Agentic AI). Deze risico's kunnen voor alle AI-systemen van toepassing zijn: EU of niet-EU, open of gesloten.

Risico's bij gebruik van LLM tbv risicogebaseerde wetgeving

De onderstaande tabel geeft de meest voorkomende privacy, security en technische risico's bij het gebruik van LLM's. Deze zijn van belang voor het correct in kaart brengen van risico's tbv de risicogebaseerde aanpak die veel wetgeving volgt.

Risico's gegevensbescherming (AVG)	Securityrisico's (AVG , CRA en LLM)	Technische risico's (LLM & Agentic AI)
Vaststellen van rollen en verantwoordelijkheden als (mede-) verwerkingsverantwoordelijke/verwerker in AI-systemen	Systeemrisico's bij veelgebruikte modellen.	Prompt injection, jailbreaking
Bepalen of bij het gebruik van een AI-systeem (gevoelige) persoonsgegevens worden verwerkt en wat de grondslag daarvan is	Beperkt aanbod aan AI-systemen dat kan leiden tot het versterken van de gevolgen van kwetsbaarheden in deze systemen	Excessive Agency & Autonomous Actions
Vaststellen en mitigeren van (potentiële) risico's gedurende de gehele levenscyclus van het AI-systeem	Specifieke veiligheidsrisico's van AI-systemen i.v.m. onbetrouwbare trainingsgegevens, de complexiteit van de systemen, ondoorzichtigheid	Overreliance risks & gebrek aan of beperkt Human Oversight, AI confidence limitations
Naleving van de beginselen inzake verwerking van persoonsgegevens: rechtmatigheid, behoorlijkheid, transparantie; doelbinding; dataminimalisatie; juistheid; dataopslagbeperking; integriteit en vertrouwelijkheid; verantwoordingsplicht	Openbaarmaking van gevoelige informatie	Model extraction attacks
Naleving van de rechten van de betrokkenen: transparantie, informatie, rectificatie, recht op vergetelheid	Beveiligingsrisico's in de toeleveringsketen door externe leveranciers	Training data poisoning (LLM) & memory poisoning (Agentic AI)
Geautomatiseerde besluitvorming	Beveiligingslekken in het ontwerp van plug-ins, authenticatie, onveilige integratie van tools,	Halucinations (LLM) & Cascading hallucination attacks (Agentic AI)
Kwaliteit van data: bias, copyright		Misaligned (de modellen werken niet zoals bedoeld) & misleidend gedrag
Transfer van persoonsgegevens naar landen buiten de Europese Unie		Overwhelming the human-in-the-loop: de human-in-the loop wordt overweldigd door de grote hoeveelheid meldingen en kan daardoor geen betekenisvolle controle uitoefenen.

4. Behoeften uit de praktijk

4.1 Opzet workshops

Inleiding

Om beter zicht te krijgen op de wensen en behoeften uit de praktijk van de Nederlandse overheid is een tweetal workshops gehouden, op 2 en 31 maart 2026. Deelnemers waren leden van de werkgroep open taalmodellen (olv Ministerie van BZK), met afvaardiging van stakeholders uit verschillende bestuurslagen en organisaties met affiniteit met het dossier 'open' en 'Europese' taalmodellen.

Workshop 1 – Kansen, risico's en obstakels

Op 2 maart is er een workshop gehouden met als doel om input te verzamelen over het gebruik van (open) Europese taalmodellen in de praktijk, de obstakels die daarbij ervaren worden en de kansen die men ziet. Daarnaast zijn eerste ideeën opgehaald ten behoeve van het ontwikkelen van een afwegingskader rond taalmodellen.
In break-outs is er gebrainstormd over de kansen, risico's en obstakels voor het gebruik van open en/of Europese GenAI modellen aan de hand van twee Use-Cases: een Chatbot voor sociaal domein en een AI-systeem voor het doorzoeken van juridische en beleidsdocumenten.

Workshop 2 – afwegingskader

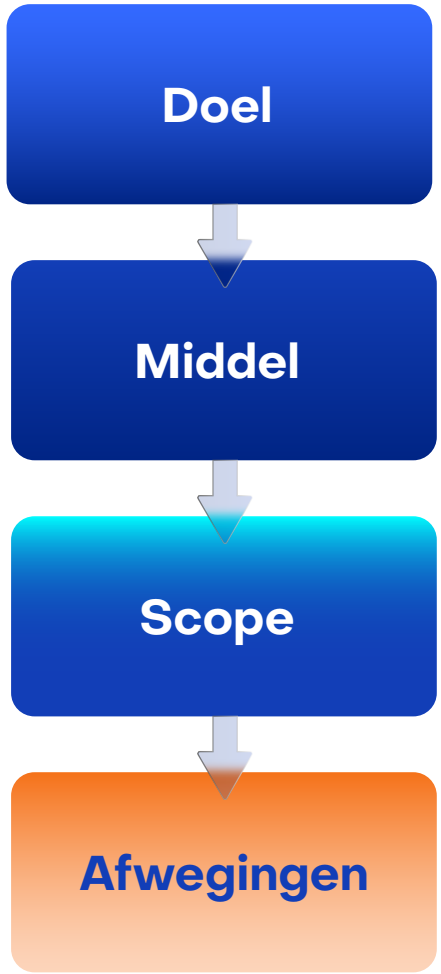
Uit workshop 1 kwam naar voren dat er behoefte is aan een afwegingskader, maar het doel en de scope van dit kader nog onduidelijk is

De tweede workshop werd gehouden op 31 maart. Hierin is met een aantal deelnemers uit workshop 1 verder gesproken over de doelen, de mogelijke middelen om dit doel te bereiken, de scope en afwegingen rond de verantwoorde inzet van (open) Europese taalmodellen.

4.2 Uitkomsten

Hier volgt een kort verslag van de uitkomsten van de twee workshops en is een weergave van wat er besproken in tijdens de twee sessies.

- In de workshops werden verschillende voor- en nadelen van open (Europese) taalmodellen genoemd. Deze zijn samengevat in tabel 1.
- Open' en 'Europees' zijn twee verschillende dingen, met verschillende onderliggende waarden/ redenen om af te wegen.
- Redenen om voor open en/of Europees te kiezen zijn een mix van lange- en korte termijn, ideologische en praktische overwegingen, zoals gebruikersgemak, soevereiniteit, versterken van de Europese digitale markt en grip op normen en waarden.
- Er is behoefte aan een afwegingskader, dat kan helpen bij het maken van de juiste afwegingen rond de keuze van een open of Europees model.
- Zo'n kader moet de keuze voor open en/of Europese GenAI modellen door overheidsorganisaties (Rijksoverheid, provincies, gemeenten en uitvoeringsorganisaties) vergemakkelijken.
- Als het niet mogelijk is om een open, of Europees model te kiezen in specifieke situatie, dan moet dit uitgelegd worden.



Tabel: voor-en nadelen van open (Europese) taalmodellen, uit de workshops

Voordelen van open taalmodellen	Nadelen/ risico's van open taalmodellen
Taak-specifiek trainen/ fine-tunen	Onduidelijke verdeling van verantwoordelijkheden
Autonomie en keuzevrijheid	**Kwaliteit en gebruikersgemak** Deelnemers ervaren dat gesloten (commerciële) modellen op veel performance metrics beter presteren.
Opbouwen van expertise	
Privacy, datasoevereiniteit	Compliance risico's
Transparantie	**Hardware en expertise** Het fine-tunen, hosten en onderhouden van open modellen vereist technische expertise, rekenkracht en hardware (GPU's). Niet elke (overheids)organisatie beschikt over deze capaciteit, wat een drempel vormt.
Ethiek, beschermen van Europese waarden	

Doelen

De redenen waarom de overheid de keuze voor open en Europese taalmodellen wil bevorderen zijn:

- Vergroten van soevereiniteit bij de inzet van GenAI modellen, en het verminderen van problematische afhankelijkheden
- Het stimuleren van Europese innovatie
- door het gebruik van kleinere open modellen andere mogelijkheden creëren met betrekking tot AI-infrastructuur en rekenkracht

Overkoepelende doelen:

- Een effectieve en rechtvaardige overheid, die de kansen van GenAI modellen benut op een menswaardig en moreel verantwoorde manier, met zoveel mogelijk transparantie.
- Op het hoogste niveau draagt dit bij aan een welvaren en gelukkig Nederland, nu en in de toekomst.

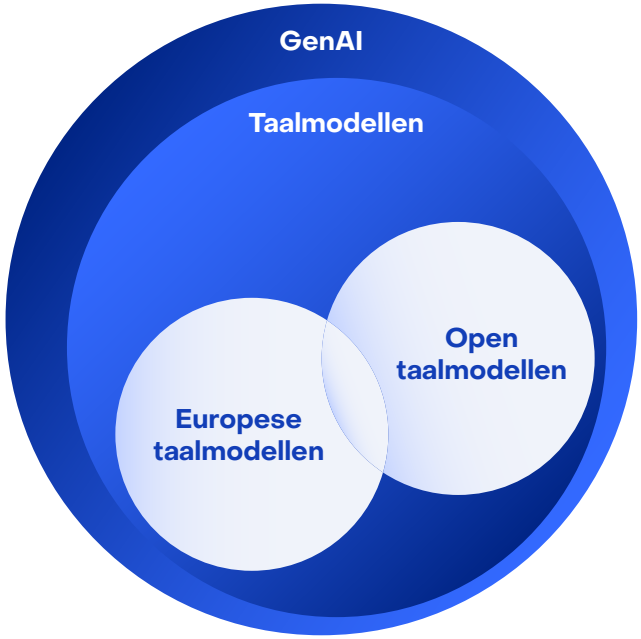
Scope van een afwegingskader

Een afwegingskader moet helpen om afwegingen te maken rond de verantwoorde toepassing van open en/ of Europese taalmodellen.

Wat is het afwegingskader *niet*:

- Beslishulp voor wanneer een use-case niet de inzet van een 'open' en/ of 'Europees' taalmodel vereist (om daarmee het gebruik van commerciële, buiten-landse modellen kan verantwoorden).
- Verantwoord inzet van GenAI binnen de overheid in brede zin. Veel overwegingen daarvoor zijn namelijk al gevat in verschillende richtlijnen en kaders, zoals bijvoorbeeld het Algoritmekader, de Impact Assessment Mensenrechten en Algoritmes (IAMA) en de Baseline Informatiebeveiliging Overheid (BIO).

Het afwegingskader moet dus *aanvullend* zijn op de diverse toetsen en richtlijnen die er al zijn om AI op een verantwoorde manier in te zetten binnen de overheid.



Afwegingen

In de workshops is ook gesproken over de afwegingen die met belangrijk vindt om mee te nemen in de keuze voor een open of Europees model. Relevante aspecten die werden genoemd zijn:

- **Juridische aspecten**, met welke juridische aspecten moet rekening gehouden worden bij de inzet van een open of Europees taalmodel. Denk bijvoorbeeld aan de AI Act, intellectueel eigendom of aansprakelijkheid
- **Openheid van een model**, openheid van modellen is een spectrum van volledig open source tot volledig gesloten modellen
- **De herkomst van een model**, om Europese innovatie te stimuleren is het belangrijk dat de overheid vaker kiest voor Europese modellen.
- **Herkomst van de trainingsdata**, om te kunnen bepalen hoe het model getraind is
- **Omvang van een model**, belangrijk om te weten in het kader van duurzaamheid en geschiktheid voor beoogde taken
- Finetuning
- Architectuur
- Praktische overwegingen

5. Afwegingen open en Europese taalmodellen

5.1 Afwegingstabel

Uit de workshops kwam naar voren dat er behoefte is aan een middel dat helpt bij het in kaart brengen van de relevante dimensies bij het kiezen en inzetten van taalmodellen. In dit hoofdstuk wordt een overzicht gepresenteerd van die dimensies, gegroepeerd naar vier categorieën: openheid, Europese elementen, technische specificaties en gebruik. Het overzicht is niet bedoeld als een prescriptief afwegingskader met een 'juiste' keuze per dimensie, maar als een hulpmiddel om inzicht te krijgen in de factoren die van belang zijn bij de keuze voor een taalmodel, op basis van de mate waarin het model open of Europees is.

Deze afwegingstabel geeft dus afwegingen weer met betrekking tot 'open' en 'Europees', en relevante technische aspecten bij de keuze voor een bepaald model. Als dit model wordt geïmplementeerd in een bepaalde AI-applicatie, dan zullen er ook andere overwegingen een rol spelen, zoals economische aspecten, datasoevereiniteit, ethische en veiligheidsaspecten. Deze zijn in dit afwegingstabel niet meegenomen.

Een LLM op zichzelf is nooit het hele verhaal, daarvoor moet naar de gehele stack worden gekeken, maar de voorbeeldmodellen geven een haakje om de mogelijke openheid en Europese en technische elementen te bespreken. In 5.2 worden de dimensies en keuzes verder uitgelicht. In 5.3 wordt de samenhang van dimensies verder besproken. In 5.4 geven we een voorbeeld toepassing van de tabel. In 5.5 volgt een discussie.

Overweging	**Openheid**		**Europese elementen**				**Technische specificaties**			**Gebruik**	
Voorbeeldmodel/ Dimensie	Openheid model (open, open-weight, closed)	Licentie type (Apache 2.0, MIT, model-specific, copyleft, proprietary)	Herkomst provider (EER, non-EER, internationaal consortium)	Dataopslag en data-verwerking (EER, non-EER)	Stack/Hosting (EER cloud, on-premise, nationaal datacentrum)	Getraind op Europese talen (ja, nee, onbekend)	Omvang model (geen gespecialiseerde GPU nodig (<13B,/klein), GPU-server(~13-70B), datacenter (>70B), nvt/ onbekend)	Fintuening mogelijk (ja/ nee / onbekend)	Benchmarkscores in het Nederlands (via EuroEval)	Kostenmodel (Gratis, self-hosting / per-token / abbonement / ..)	Toepassings-gebied
Mistral 7B	Open-weight	Apache 2.0	EER	EER (self-hosting of API)	Alles mogelijk	Ja	Klein	Ja	1.44 (Rank score)	Self-hosting	
Apertus 70B	Open Source	Apache 2.0	non-EER (Zwitserland)	(self-hosting)	Alles mogelijk	Ja	Middel	Ja	i.e. 48.65/60.81 (Knowlegde, MMLU_NL)	Self-hosting	
GTP-4o	Closed	Proprietary	Non-EER (VS)	Onduidelijk	Non-EER cloud (Microsoft Azure)	Onbekend	Onbekend	Nee	1.30 (Rank score)	Per-token	

5.2 Toelichting dimensies afwegingstabel

5.2.1 Openheid model

De mate van openheid geeft aan hoeveel van het model publiek beschikbaar is (zie Sectie 2.2). Bij een **open-source** model zijn zowel de weights, trainingsdata als trainingscode beschikbaar. Apertus is hier een voorbeeld van. Bij **open-weight** zijn alleen de modelgewichten beschikbaar, je kunt het draaien en vaak ook finetunen, maar hebt vaak geen inzicht in trainings-data of -proces. Dit is de meest voor-komende vorm bij modellen als Llama en Mistral. Bij **closed** modellen is niets beschikbaar en gebruik je het model uitsluitend via een API.

Licentie type

De licentie bepaalt wat je juridisch *mag* doen met een model, los van de technische openheid. Apache 2.0 en MIT zijn tolerant en staan vrij gebruik, aanpassing en herverdeling toe, zoals bij Mistral 24B en Apertus 70B. **Model-specifieke licenties** (zoals de Llama Community License) bevatten eigen beperkingen, bijvoorbeeld een maximaal aantal gebruikers. **Copyleft-licenties** vereisen dat afgeleide werken onder dezelfde licentie worden vrijgegeven. **Proprietary** is geen license, in dit geval heeft de gebruiker alleen de rechten die de leverancier expliciet toestaat, zoals bij GPT-5.4

5.2.2 Europese elementen

Herkomst provider

Geeft aan of de organisatie achter het model een hoofdkantoor heeft in de Europese Economische Ruimte (EER): in de **EER** Mistral's modellen bijvoorbeeld), **buiten de EER** (OpenAI's GPT-5.4), of als onderdeel van een **internationaal consortium** (soms een combinatie van EER en niet-EER). Dit bepaalt onder welke jurisdictie de provider valt. Een EER-provider is onderworpen aan de AVG en de AI Act, terwijl non-EER providers onderhevig kunnen zijn aan conflicterende buitenlandse wetgeving zoals de US CLOUD Act. Zie ook Sectie 2.3.

Dataopslag en dataverwerking

Om de relatie van een taalmodel met Europa te beoordelen, is het belangrijk om niet alleen te kijken naar de herkomst van het model zelf, maar ook naar de volledige stack van het AI systeem waarin het taalmodel in ingebed. Belangrijk daarin is de dataopslag en –verwerking van het AI-model. Als dit niet plaatsvindt in de EER, kan dat serieuze gevolgen hebben voor de veiligheid en toegankelijkheid van de data. Het is echter vaak ondoorzichtig hoe aanbieders van AI systemen dit hebben geregeld. Alleen als een gebruiker het AI systeem volledige zelf host en de data lokaal opslaat, is er volledige grip op dit element. Soms bepaalt de modelkeuze waar de data wordt verwerkt en opgeslagen (vaak bij gesloten modellen), soms kun je zelf kiezen waar je een model draait (bij open modellen).

Stack/Hosting

Een Europees model is niet automatisch soeverein als het draait op niet-Europese infrastructuur. Een model kan **on-premise** draaien, op bijvoorbeeld een eigen server; bij een **nationaal datacentrum** zoals SURF; op een **EER cloud**, bijvoorbeeld STACKIT (Duitsland) of previder (Nederland), of op een non-EER cloud, bijvoorbeeld AWS (ook de European Sovereign Cloud) of Microsoft Azure Foundry. Gesloten modellen draaien op de infrastructuur van de provider (GPT-5.4), bij open modellen kan worden gekozen om het model on-premise, via een groter datacentrum of zelfs via managed services op een (non-) EER cloud te draaien.

Getraind op Europese talen

Geeft een indicatie of het model expliciet getraind is op Europese taaldata, in het bijzonder het Nederlands. Veel modellen bevatten enige Nederlandstalige data, maar de hoeveelheid varieert sterk. De aan-wezigheid van het Nederlands kan al dan niet iets zeggen over hoe goed het model is in Nederlands en of het bijvoorbeeld ook op de bijbehorende culturele waarden is getraind. "Ja" betekent dat Europese taaldata aan-toonbaar deel uitmaakte van de trainingsset. Zie ook Sectie 2.3. De hoeveelheid data van een bepaalde taal correleert niet per se met de vaardig-heden van het model in die taal, maar als er in het ontwerpproces expliciet aandacht is gegeven aan kleinere taalfamilies kan dat wel een indicatie geven van de prestaties in de taal.

5.2.3 Technische specificaties

Omvang model

De parametergrootte bepaalt welke hardware nodig is voor self-hosting. **Kleine modellen** (<13B parameters, bijv. Mistral 7B) draaien op een standaard server zonder gespecialiseerde GPU. **Medium modellen** (~13–70B, bijv. Llama 3 70B) vereisen meestal meerdere GPU-servers. **Grote modellen** (>70B, bijv. Llama 3.1 405B, maar ook Apertus 70B) hebben vaak datacenter-

infrastructuur nodig. Bij closed modellen is de omvang doorgaans onbekend.

Parametergrootte is wederom niet het hele verhaal: architectuur zoals Mixture-of-Experts kan een model met veel parameters toch relatief licht maken. Ook is bijvoorbeeld quantisatie van invloed.

Finetuning mogelijk
Geeft aan of het model aangepast kan worden aan specifieke overheidstaken, zoals het verwerken van juridische teksten, burgercommunicatie en versimpelen of samenvatten van brieven. Open(-weight) modellen zijn technisch vrijwel altijd finetunebaar, maar of het ook *mag* (licentie) en *haalbaar* is (hardware) verschilt per geval. Lichtgewicht methoden als LoRA/QLoRA verlagen de drempel ten opzichte van een volledige finetune.

Benchmarkscores in het Nederlands
Zoals besproken in Sectie 2.1, kunnen (een combinatie van) benchmarks gebruikt worden om een LLM te evalueren. Een gestandaardiseerde, datagedreven meting van modelprestaties in het Nederlands via het EuroEval Dutch Leaderboard. EuroEval test op taalvaardigheid, redeneren, kennistaken en begrijpend lezen. Hierin kunnen ook bias evaluaties opgenomen worden. AI-onderzoekers kunnen zelfs benchmarks en modellen toevoegen aan het leaderboard en daarmee vergelijkingen uitbreiden. Dit biedt een betrouwbaardere vergelijking dan subjectieve beoordelingen of Engelstalige benchmarks, en is specifiek relevant voor de Nederlandse overheidscontext. Belangrijk is wel dat benchmarkscores zich slecht laten vertalen naar de prestatie op een specifieke taak. Om dit in kaart te brengen dient altijd context-specifiek onderzoek gedaan te worden, of ervaringen tussen gebruikers uitgewisseld te worden.

5.2.4 Gebruik

Kostenmodel
De manier waarop kosten worden gemaakt bepaalt zowel de financiële haalbaarheid als de mate van afhankelijkheid. **Gratis** modellen zijn vrij te downloaden, maar kosten zitten in hardware en beheer. **Self-hosting** betekent geen licentiekosten maar wel investeringen in infrastructuur en personeel. **Per-token** prijsmodellen bieden variabele kosten op basis van daadwerkelijk gebruik. **Abonnementen** via managed services geven voorspelbare kosten maar minder flexibiliteit.

Toepassingsgebied
De prestaties van taalmodellen variëren over verschillende taken. Het toepassingsgebied waarvoor het AI model ingezet zal worden is daarom ook een belangrijk criterium om mee te wegen bij de keuze voor een taalmodel.

5.3 Interacties tussen de dimensies

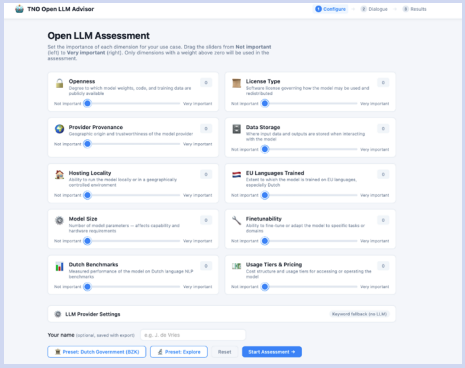
De dimensies in het afwegingskader staan niet los van elkaar. Een keuze op één dimensie beperkt of verruimt de mogelijkheden op andere dimensies. Kies je bijvoorbeeld voor een gesloten model, dan is finetuning niet mogelijk. Wil je een model zelf hosten, dan vereist dat een open(-weight) model, wat beperkingen kan opleggen aan de modelomvang, maar tegelijkertijd dataverwerking en -opslag binnen de EER garandeert.

Het overzicht hiernaast illustreert deze onderlinge afhankelijkheden: welke keuzes op welke andere dimensies doorwerken. De kleuren komen overeen met de dimensiekleuren uit het afwegingskader.

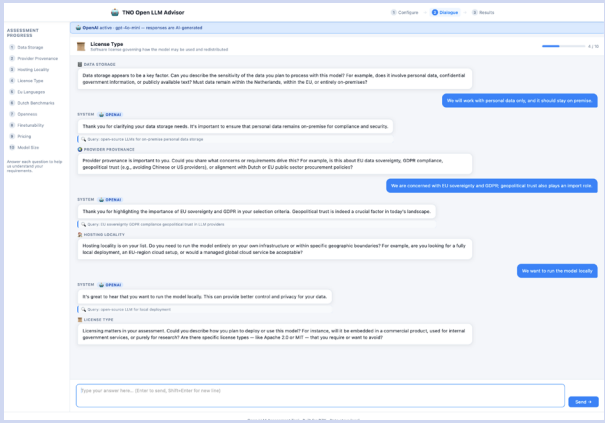
Als je kiest voor...		**Dan geldt in andere dimensies...**
Closed model		Finetuning niet mogelijk
		Self-hosting niet mogelijk, hosting = provider's infra
		Datalocatie afhankelijk van provider en afspraken
Open-weight model		Finetuning mogelijk
		Self-hosting mogelijk, hosting-keuze flexibel
		Licentie kan beperkingen opleggen
Self-hosting vereist		Open of open-weight noodzakelijk
		Omvang model bepaalt GPU-eis en andersom; modelomvang kan beperkt zijn
		Dataopslag en verwerking in EER
Data moet in EER		Sluit non-EER cloud uit; self-hosting of EER-cloud nodig
Data moet in EER	Groot model(>70B)	Hosting-infra beperkt
Klein model (<13B)		Self-hosting laagdrempelig
		Finetuning haalbaar met LoRA
		Complexe taken mogelijk minder goed, mogelijk lagere performance en lagere scores op benchmarks
Groot model (>70B)		Self-hosting vergt datacenter
		EER GPU-capaciteit beperkt
		Betere prestaties op complexe taken, vaak hogere performance en hogere benchmarkscores
Finetuning vereist		Open of open-weight noodzakelijk
		Meestal on-premise hardware nodig
		Specialisatie voor NL-taken mogelijk, kan benchmarkscores verhogen
Kostenmodel per-token		Impliceert closed model (toegang via API)
		Dataverwerking bij provider
		Geen finetuning mogelijk

5.4 Voorbeeld toepassing van afwegingstabel

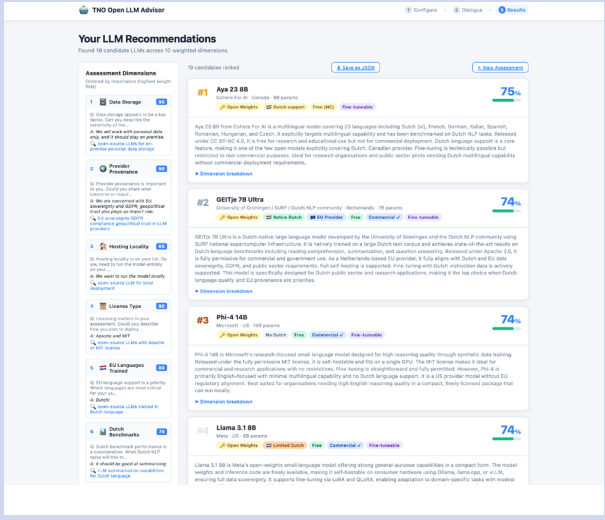
Om te illustreren hoe de afwegingstabel in de praktijk ingezet kan worden om een model te selecteren, hebben we een eenvoudige demonstrator ontwikkeld. Deze demonstrator, de TNO Open LLM Advisor, leidt gebruikers via een dialoog over een gekozen waarden-stelsel naar een geordende lijst van kandidaat-LLMs De demonstrator opent met een scherm (figuur 5) met de criteria uit de afwegingstabel (Sectie 5.1). Gebruikers kunnen per waarde aangeven wat het voor hen het relatieve belang is voor het maken van een keuze voor een bepaalde LLM, door sliders naar links of rechts te trekken. Uit de stand van de sliders wordt een beslisboom afgeleid, die -via een LLM- gebruikers per criterium gaat uitvragen naar aspecten van het criterium (figuur 6). De volgorde van de vragen in de uitvraag wordt bepaald door de slider-stand: de waarde met de hoogste slider-score wordt het eerst bevraagd, enzovoorts. De aldus verkregen informatie wordt (gewogen met de sliderstand) gematched met een achterliggende database (een RAG, zie rapport), waarin algemene informatie van (op dit moment 19) publieke LLMs zijn opgenomen. De matches (per criterium) worden omgezet in een gewogen ranking, die de 19 LLMs ordent, met als eerste de best matchende LLM. Zie bijlage A voor een volledige beschrijving.



Figuur 5: Startscherm



Figuur 6: Dialoog



Figuur 7: Aanbeveling

5.5 Discussie

We zien dat de dimensies in het afwegingskader niet los van elkaar bestaan. Ze laten zich niet reduceren tot een simpel onderscheid tussen 'open' en 'gesloten' of 'Europees' en 'niet-Europees'. Deze twee hoofd-dimensies zijn schalen, geen binaire keuzes, en de interactie ertussen levert combinaties op die zich niet eenduidig in een matrix laten vangen.

'Open' is een spectrum, geen schakelaar

Zoals in Sectie 2.2 is beschreven, varieert de openheid van een taalmodel van volledig open (trainingsdata, code én gewichten beschikbaar, zoals bij Apertus) tot open-weight (alleen de gewichten, zoals bij Llama of Mistral) tot volledig gesloten (alleen API-toegang). Een model van Mistral kan weliswaar open-weight zijn, maar de trainingsdata en -processen zijn niet publiek. Dat biedt wél de mogelijkheid om het model zelf te hosten en te fine-tunen, maar níet het inzicht in trainingsdata dat nodig is om bias of dataherkomst te beoordelen. Anderzijds is bijvoorbeeld GPT-NL wel transparant over hoe het model tot stand is gekomen en waar de data vandaan komt, maar geeft het geen inzicht in de volledige trainingsdataset, omdat hier auters- rechterlijk beschermde data in zit.

'Europees' en 'gesloten' is niet eenduidig

Neem twee modellen die beide als 'Europees' en 'niet-open' zouden worden geclassificeerd: Mistral en GPT-NL. In de praktijk verschillen deze modellen. Mistral is een Frans commercieel bedrijf dat open-weight modellen aanbiedt onder eigen licentievoorwaarden, maar met beperkt inzicht in trainingsdata en een voortdurend risico op overname door niet-Europese investeerders. GPT-NL daarentegen is een initiatief dat specifiek gericht is op soevereiniteit: het model wordt op Nederlandse bodem ontwikkeld, er is verantwoording over de dataverzameling, en er is transparantie over welke data is gebruikt - ook al kan er geen volledig inzicht gegeven worden. Dat het niet open-source is, heeft een bewuste reden: er zit een bedrijfsmodel achter dat data-eigenaren compenseert, waardoor de trainingsdata niet vrij beschikbaar kan komen. GPT-NL scoort daarmee goed op meerdere criteria, zoals EER-herkomst, EER-infrastructuur, datasoevereiniteit, soms sterker dan veel modellen die technisch 'opener' zijn.

De interactie tussen de dimensies

De dimensies 'openheid' en 'Europeesheid' staan niet los van elkaar. Een niet-Europees open model (zoals Llama) biedt technische transparantie en flexibiliteit, maar geen garanties over dataherkomst, governance of aansluiting bij Europese waarden. Een Europees niet-open model kan juist op die punten sterker zijn. En zelfs een volledig open Europees model garandeert geen soevereiniteit als het draait op niet-Europese cloudinfrastructuur met niet-Europese GPU's. De keuze voor een taalmodel vraagt daarom niet om het aanvinken van vakjes, maar om een afweging die de samenhang tussen 'open' en 'Europees' erkent.

Een model is nog geen applicatie

De afwegingstabel in dit rapport is bedoeld als hulpmiddel voor overheidsorganisaties om inzicht te krijgen in de overwegingen die spelen bij de keuze voor een open of Europees model. Deze modellen zullen doorgaans worden gebruikt voor uiteenlopende AI-toepassingen, denk aan een chatbot die burgers informeert over parkeerbeleid in een gemeente, een systeem die ondernemers helpt het beleid en regelgeving rond MKB te doorgronden of een AI systeem die ambtenaren helpt juridische- en beleidsdocumenten te doorzoeken. Bij de inzet en ontwikkeling van zulke AI-toepassingen zijn meer afwegingen van belang dan opgenomen in de afwegingstabel.

Bijvoorbeeld de geschiktheid van het model voor de beoogde toepassing, de opslag en verwerking van data, juridische aspecten (zie ook hoofdstuk 3), ethische aspecten (hiervoor bestaan frameworks zoals het Algoritmekader en, de Impact Assessment Mensenrechten en Algoritmes (IAMA)) , economische overwegingen (buiten de kosten voor het model- zoals levert het een efficiëntieslag op, en cyberveiligheid.

Een compleet kader beoordeelt de AI-applicatie vanuit de beoogde toepassingscontext, en weegt daarbij ook de openheid en Europeesheid van het gekozen model mee, naast een brede range aan andere, toepassingsafhankelijke, aspecten.

6. Conclusies en algemene aanbevelingen

6.1 Conclusies

Het inzetten van taalmodellen vraagt om het maken van verantwoorde keuzes. Het veranderende geopolitieke speelveld, waarin bestaande digitale afhankelijkheden steeds ongewenster zijn, willen organisaties stappen zetten om meer grip te hebben op digitale autonomie. Dit rapport geeft een beknopt overzicht van belangrijke ontwikkelingen en afwegingen rond het complexe beleidsterrein van open en Europese taalmodellen, de context daarvan en relevante begrippen. Daarbij schetsen we technologische, (geo) politieke en juridische aspecten. Dit document kan daarmee dienen als *policy primer*, en bijdragen aan het opstellen of beoordelen van een leidende aanpak en beleid rondom te inzet van (open) Europese taalmodellen in organisaties. Algemene conclusies van dit rapport zijn;

1. **Taalmodellen zijn onderdeel van een stack**

Voor het versterken van digitale soevereiniteit van de Nederlandse overheid is het stimuleren van het gebruik van open en Europese taalmodellen cruciaal. Maar digitale technologieën, zoals ook AI, zijn opgebouwd uit diverse lagen, ook wel 'de digitale stack' genoemd. De onderste lagen bestaat uit grondstoffen, chips en netwerken, daarbovenop draaien cloud en data infrastructuren, helemaal bovenin de stack zitten de 'intelligentie', de applicatie en de user interface. Als het gaat om digitale soevereiniteit kan dus niet alleen gekeken worden naar de laag van het model. Immers, het model is gebouwd op de onderliggende lagen, en maakt bijvoorbeeld gebruik van de data en cloudinfrastructuur.

2. **Geen leidende, breed gedragen definities voor 'open' en 'open source' taalmodellen**

Op dit moment is er nog geen leidende definitie voor 'open' in brede zin, en specifieker 'open source' taalmodellen. In dit rapport hebben we een aantal suggesties gedaan om hieraan invulling te geven. Met name het gebruik van de term 'open' taalmodellen kan een bron van verwarring zijn. Door de vele aspecten waar een model 'open' in kan zijn, moeten we openheid beschrijven in termen van een schaal met grove gradaties. In dit landschap waarin definities van 'open' onduidelijk zijn, kunnen (commerciële) aanbieders van taalmodellen nog veelal zelf bepalen hoe zij hun taalmodellen kwalificeren. Dit schept onduidelijkheid voor gebruikers en afnemers, en kan leiden tot 'open washing' waarin aanbieders hun product als 'open' aanbieden, maar minimaal scoren qua openheid.

3. **Drie stadia van modelopenheid**

Onder andere op basis van de European Open-Source AI Index (Liesenfeld & Dingemanse, 2024) en The Model Openness Framework kunnen we grofweg vijf verschillende gradaties van openheid vaststellen:

Closed source, Open code, Open Weights, Open Research en volledig Open Source. In de praktijk komen drie categorieën vaakst voor: Closed source en open weights, met enkele Open source modellen. Modelontwikkelaars zijn vrij om al dan niet transparant te zijn over alle onderdelen in een LLM pijplijn. Drie categorieën vangen nooit alle modellen precies (Sectie 2.2), maar geven een indicatie van de openheid van het model.

4. **Europese taalmodellen: herkomst, datasoevereiniteit en training op Europese talen**

Ook de vraag of een taalmodel Europees is kan vanuit diverse invalshoeken bekeken worden. Denk bijvoorbeeld aan de herkomst van de aanbieder, de herkomst en opslag van de data en of een model getraind is op Europese talen. Wat een model 'Europees' maakt, is een samenspel van deze factoren.

5. **Interactie tussen dimensies 'Europeesheid' en 'Open'**

Een belangrijke observatie is dat de dimensies in het afwegingstabel niet los van elkaar staan. Dat maakt dat elke beslissing invloed heeft op de volgende keuzes. Hier rijst de vraag: wat is de meest belangrijke overweging? Moet een taalmodel Europees zijn? Of liever helemaal open? Of is het belangrijker dat de data binnen de EER blijft, of zelfs lokaal? De eerste keuze bepaalt in grote lijnen wat de opeenvolgende keuzes gaan zijn.

6. **FOSS heeft uitzonderingspositie in veel EU wetgeving**

Niet-commerciële AI-systemen (software, gegevens, modellen, parameters, wegingen), met inbegrip van taalmodellen, die worden vrijgegeven onder vrije en opensource licenties (FOSAI), vallen grotendeels buiten de reikwijdte van EU wetgeving zoals bijvoorbeeld de AI Verordening (AI-act). Dit zou als een voordeel

kunnen worden beschouwd ten opzichte van het gebruik van commerciële alternatieven.

7. Behoefte aan hulp bij afwegingen rond de inzet van open Europese taalmodellen

Uit diverse workshops met stakeholders uit de overheid blijkt dat er behoefte is aan ondersteuning bij het maken van keuzes voor op en Europese taalmodellen.

6.2 Algemene aanbevelingen

Aanbeveling 1: Stel werkdefinities vast voor beleidskaders.

Om mogelijkheden voor 'open-washing', 'Europe-washing' en 'soevereiniteit-washing' door (commerciële) aanbieders te beperken is het belangrijk om scherp te zijn over de definities en de kaders voor 'open', 'open source' en 'Europese' taalmodellen. Het is raadzaam om beleidsmakers, strategen en inkoop-medewerkers die hiermee te maken hebben hierover te informeren.

Aanbeveling 2: Uitwerking praktisch afwegingskader of (interactie) beslis-systeem

In dit rapport zijn een aantal criteria uiteengezet met betrekking tot 'open' en 'Europees', en relevante technische aspecten bij de keuze voor een bepaald model. Deze dimensies zijn echter nog geen routekaart. De interactie tussen dimensies (Sectie 5.3) toont aan dat deze 'keuzeroute' verandert als je een keuze neemt. Dit zou verder uitgewerkt kunnen worden tot een concreet besluitvormingsinstrument dat organisaties op strategisch niveau helpt bij het maken van keuzes en beleid ten aanzien van in te zetten taalmodellen.

Het voorbeeld in Sectie 5.4 laat zien hoe het afwegingstabel kan worden gebruikt om een interactief instrument te ontwikkelen die organisaties handvaten geeft om keuzes te maken en inzicht te krijgen in wat de invloeden zijn van die keuzes op andere keuzes. In een dergelijk interactief instrument zouden verschillende startpunten in de beslisroute opgenomen kunnen worden, welk gewicht elke beslissing krijgt, alsook de afhankelijheden tussen dimensies.

Hierbij is het belangrijk om scherp te hebben wat de beleidskaders- en doelen zijn, en wie de eindgebruikers van een dergelijk instrument zijn. het Het is raadzaam om het einddoel, denk bijvoorbeeld aan het nastreven van digitale soevereiniteit, als uitgangspunt te nemen, om vervolgens diverse keuzes te bieden die bijdragen aan dat doel.

Aanbeveling 3: Stel een leaderboard op met ervaringen van overheidsgebruikers met open en Europese modellen

Als je een Europees en open model hebt gekozen, hoe weet je dan hoe goed het model werkt voor verschillende doeleinden? En wat als er een nieuw model op de markt komt? Een levend overzicht (leaderboard) van relevante taalmodellen kan hier een uitkomst bieden. Modellen kunnen beoordeeld worden op dimensies in het afwegings-tabel, maar ook op bijvoorbeeld prestaties en gebruikerservaringen in LLM-taken en AI-applicaties die relevant zijn voor de overheid.

Aanbeveling 5: Ontwikkel een afwegings- en besliskader voor AI-applicaties

De afwegingstabel in dit rapport is bedoeld als hulpmiddel voor overheidsorganisaties om inzicht te krijgen in de overwegingen die spelen bij de keuze voor een open of Europees model. De keuze voor een model op basis van deze afwegingen houdt echter nog geen rekenen met de uiteenlopende AI-toepassingen waarvoor het model kan worden gebruikt.

Een volledig kader beoordeelt de AI-applicatie vanuit de beoogde toepassingscontext, en weegt daarbij ook de openheid en Europeesheid van het gekozen model mee, naast een brede range aan andere, toepassingsafhankelijke, aspecten, zoals de geschiktheid van het model voor de beoogde toepassing, de opslag en verwerking van data, juridische en ethische aspecten, economische overwegingen en cyberveiligheid.

Aanbeveling 7: Wees alert op nieuwe vormen van afhankelijkheid die kunnen ontstaan.

De markt voor LLM's is sterk geconcentreerd en momenteel is er sprake van een duopolie VS/China. Dit houdt in dat de meeste foundational modellen en modellen voor algemene doeleinden zoals Claude en DeepSeek uit de Verenigde Staten en uit China afkomstig zijn. Dit is belangrijk voor alle AI-applicaties die op deze modellen gebouwd zijn en medebepalend voor de openheid daarvan en de mate waarin ze als Europees kunnen worden beschouwd.

De markt voor LLM's is niet alleen sterk geconcentreerd, maar ook heel klein in termen van aanbod.
Sterke concentratie en een klein aanbod leiden tot beperkte keuzemogelijkheid, afhankelijkheid en lock-in.

Aanbeveling 8: Overweeg zowel de kosten als de baten van *early adoptie*.
Vroege adoptie van een nieuwe technologie zoals LLM's brengt niet alleen voordelen maar ook nadelen met zich mee. Een te diepe integratie daarvan in bestaande systemen of het delegeren van essentiële taken aan AI-applicaties die daarop gebouwd zijn is onwenselijk. LLM's kunnen immers onvoorspelbaar gedrag tonen en het is nog onduidelijk welke business modellen de makers van de technologie zullen gebruiken. Waarborg waarden als betrouwbaarheid, voorspelbaarheid, toegankelijkheid en betaalbaarheid van publieke diensten door kleinschalige LLM-pilots in beschermde omgevingen (*sandboxes*) uit te voeren.

Aanbeveling 9: Koester, bescherm en stimuleer het vrije en opensource AI-ecosysteem.
Dat vereist een bepaalde inspanning van de overheid om de vrije en opensource community te beschermen, de onafhankelijkheid daarvan te waarborgen en ervoor zorgen dat het voort kan blijven bestaan als een duurzame bron van creativiteit en innovatie.

Aanbeveling 10: Steun de ontwikkeling van open standaarden.
De overheid kan een belangrijke rol spelen door zich actief in te zetten voor de ontwikkeling van open, onafhankelijke, internationale en interoperabele standaarden.

Bronnen, figuren

Referenties

- A. Liesenfeld & M. Dingemanse (2024) *Rethinking open source generative AI: open-washing and the EU AI Act*. In: The 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24), June 03–06, 2024, Rio de Janeiro, Brazil. ACM, New York, NY, USA, 14 pages. https://dl.acm.org/doi/10.1145/3630106.3659005
- L. Floridi, C. Buttaboni, E. Hine, C. Novelli, T. Shroder, G. Schanklin (2025) *Open-source AI Made in the EU: Why it is a good idea*. Minds and Machines 35:23. https://doi.org/10.1007/s11023-025-09728-x
- F. Bria, P. Timmer, F. Gernone et al. (2025) *EuroStack – A European Alternative for Digital Sovereignty*. Bertelsmann Stiftung. Gütersloh. DOI 10.11586/2025006
- The Free Software Foundation (FSF) What is Free Software?, https://www.gnu.org/philosophy/free-sw.en.html#four-freedoms
- The Open Source Initiative (OSI) What is Open Source AI? https://opensource.org/ai/open-source-ai-definition
- The Open Source Initiative (OSI) Open Source Definition, https://opensource.org/osd

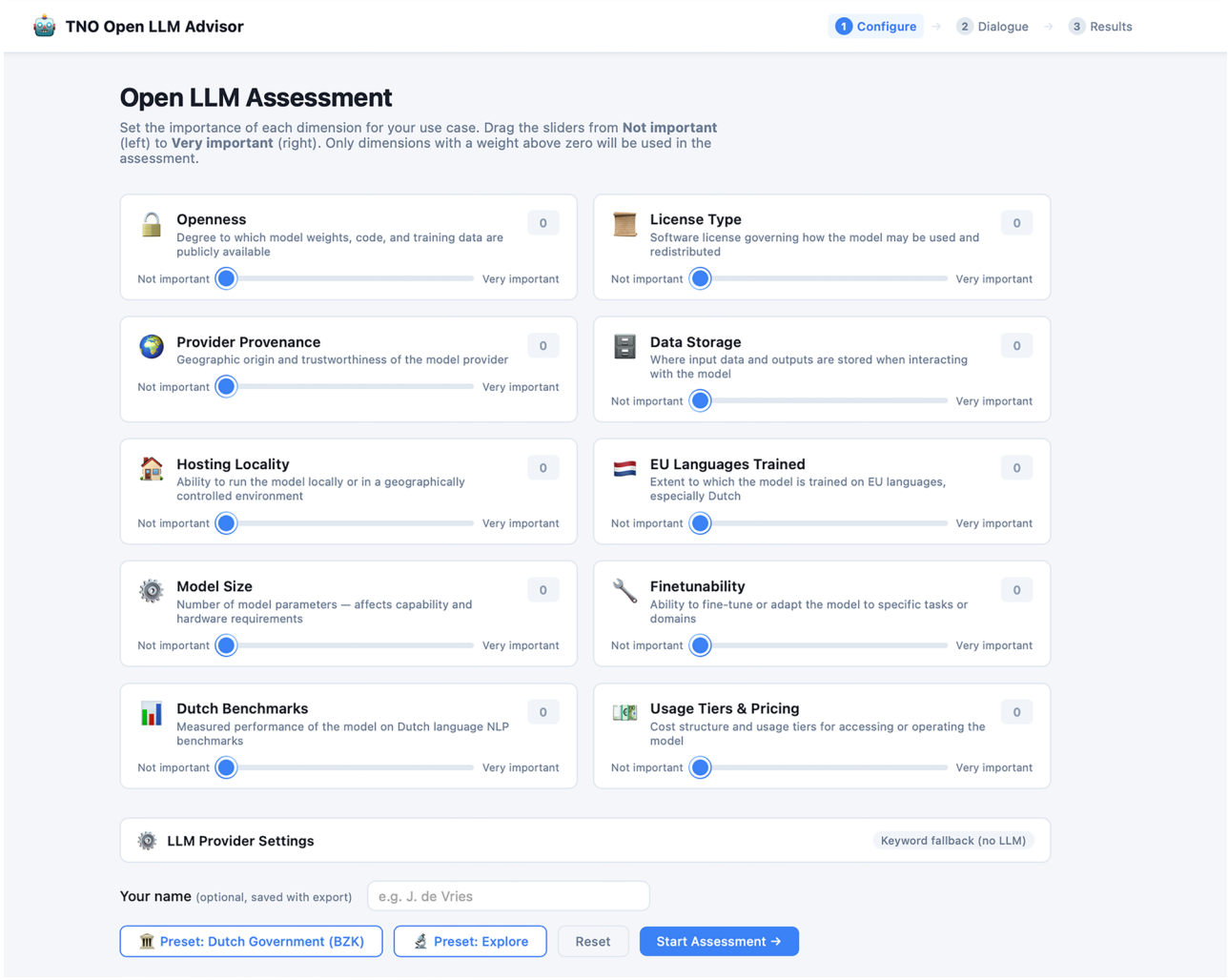
Figuren

Bijlage: TNO LLM Advisor

TNO Open LLM Advisor
Om de praktische bruikbaarheid van een waardenstelsel te illustreren voor het maken van keuzes voor open LLMs, hebben we een eenvoudige demonstrator gebouwd. Deze demonstrator, de TNO Open LLM Advisor, leidt gebruikers via een dialoog over een gekozen waardenstelsel naar een geordende lijst van kandidaat-LLMs.

Startscherm

De demonstrator opent met een scherm met een overzicht van relevante waarden, zoals in dit project vastgesteld door BZK en TNO (Figuur 1). Gebruikers kunnen per waarde aangeven wat het relatieve belang is voor het maken van een keuze voor een bepaalde LLM, door sliders naar links of rechts te trekken. Ook kunnen zij hun gebruikersnaam invoeren, voor (persoonlijke) logging van de resultaten (zie hieronder).



Figuur 1: openingsscherm

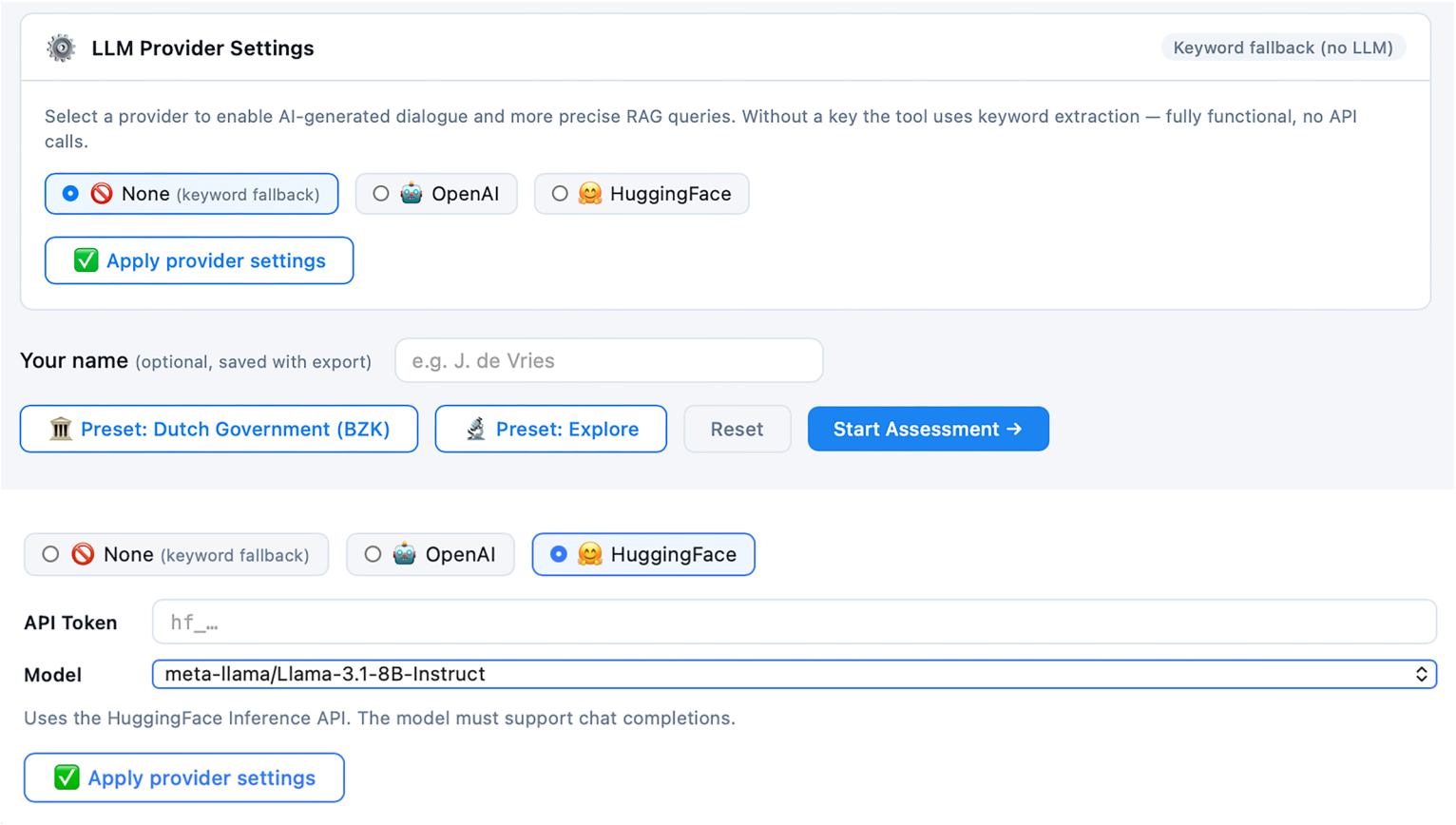
TNO Open LLM Advisor

Uit de stand van de sliders wordt een beslisboom afgeleid, die -via een LLM- gebruikers per waarde gaat uitvragen naar aspecten van de waarden. De volgorde van de vragen in de uitvraag wordt bepaald door de slider-stand: de waarde met de hoogste slider-score wordt het eerst bevraagd, enzovoorts.

De aldus verkregen informatie wordt (gewogen met de sliderstand) gematched met een achterliggende database (een RAG, zie rapport), waarin algemene informatie van (op dit moment 19) publieke LLMs zijn opgenomen. De matches (per waarde) worden omgezet in een gewogen ranking, die de 19 LLMs ordent, met als eerste de best matchende LLM.

Dialoogvoering met LLMs

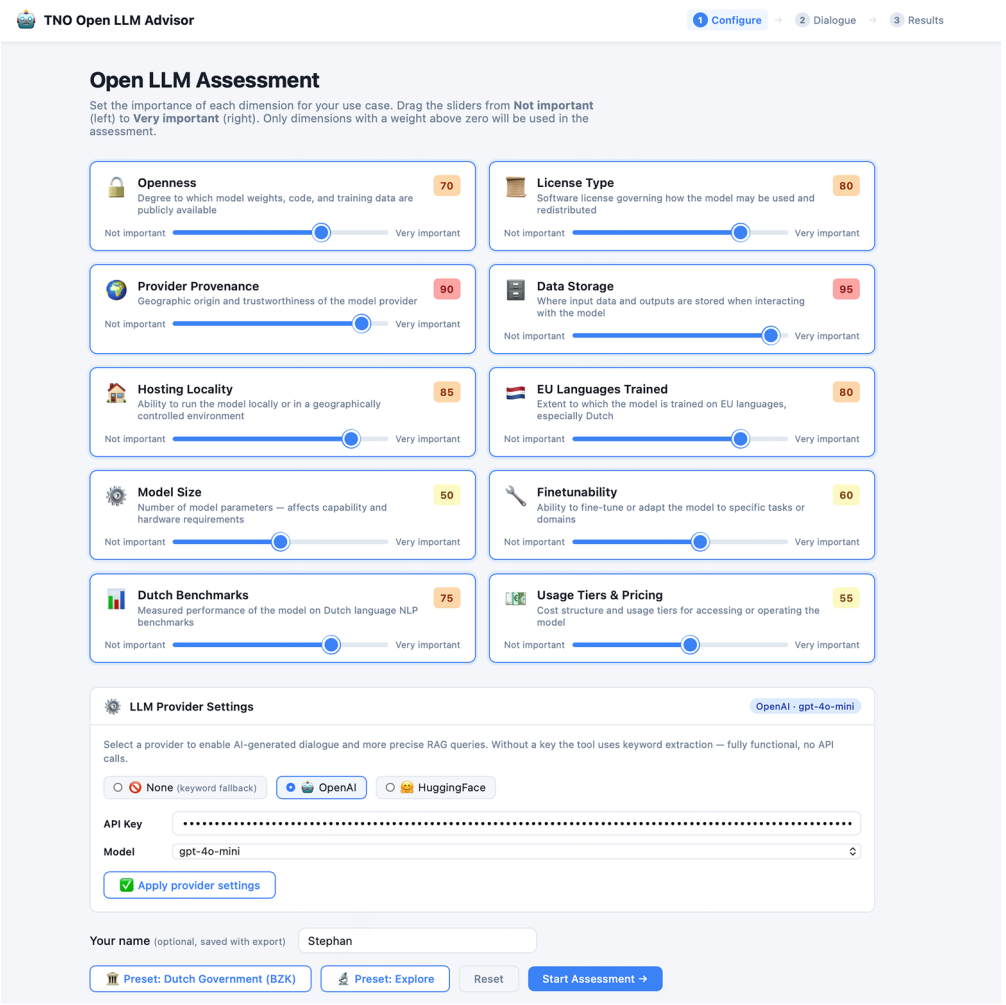
Gebruikers kunnen kiezen met welke LLM (een publieke LLM, van HuggingFace of een commerciële OpenAI LLM) ze een dialoog willen hebben (Figuur 2). Er zijn diverse opties voor beide typen modellen. Ook kunnen ze ervoor kiezen geen LLM voor de dialoog te gebruiken, maar alleen via keywords uit hun invoer de match met de RAG database te maken. Voor beide typen LLMs (HuggingFace, OpenAI) moeten gebruikers API tokens invoeren.



Figuur 2: Selectie LLMs voor dialoogvoering.

TNO Open LLM Advisor Presets

Voor demonstratiedoeleinden hebben we een ‘BZK preset’ gemaakt (Figuur 3): een vaste stand van de sliders. Die preset is aanpasbaar, en zou zelfs in de praktijk afgeleid kunnen worden uit de persoonlijke slider-standen van ‘echte’ gebruikers.

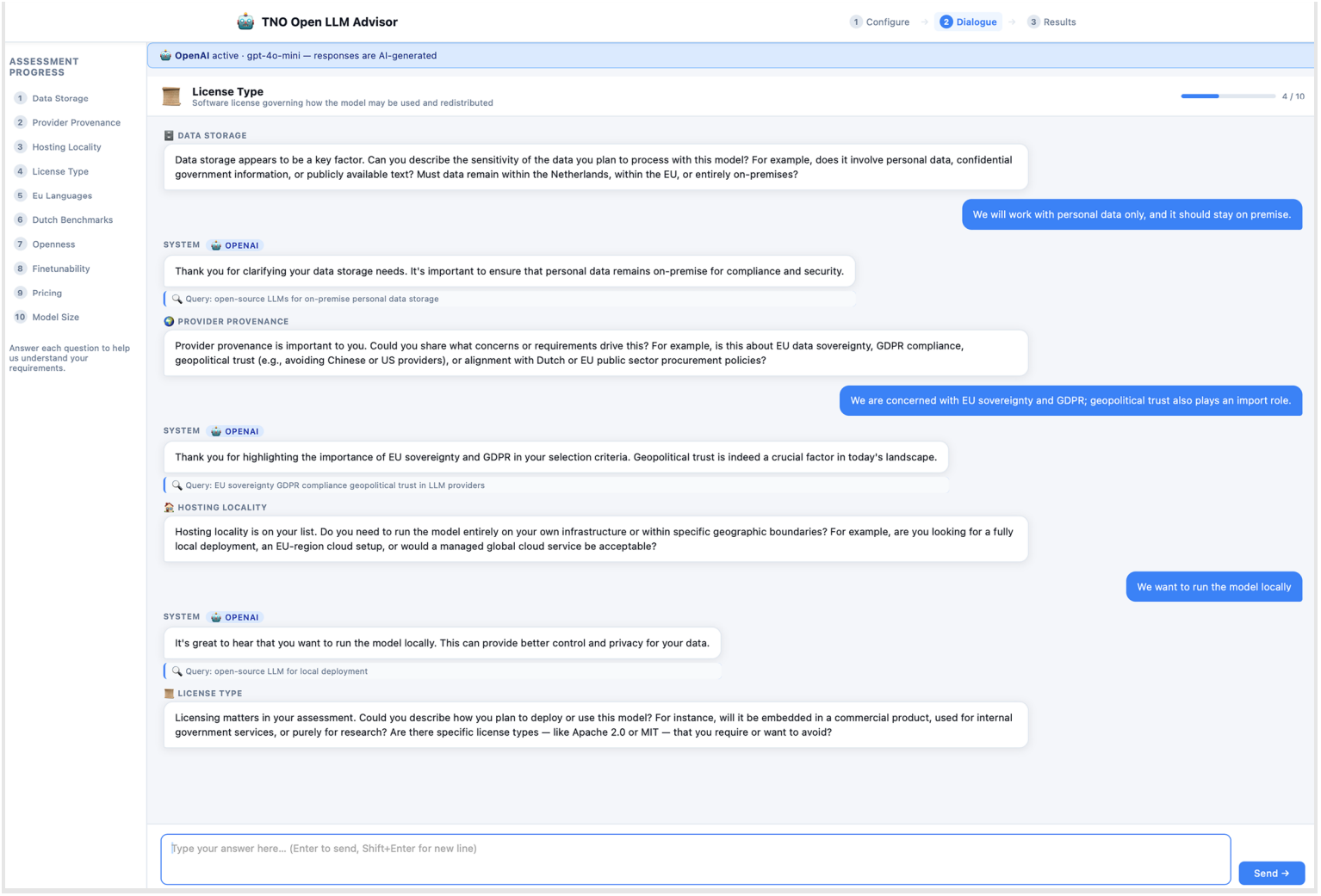


Figuur 3: de BZK preset.

TNO Open LLM Advisor

Dialoog

Vervolgens vindt de dialoog met de gebruikers plaats, aan de hand van sliders, en op volgorde van de slider-standen (Figuur 4)



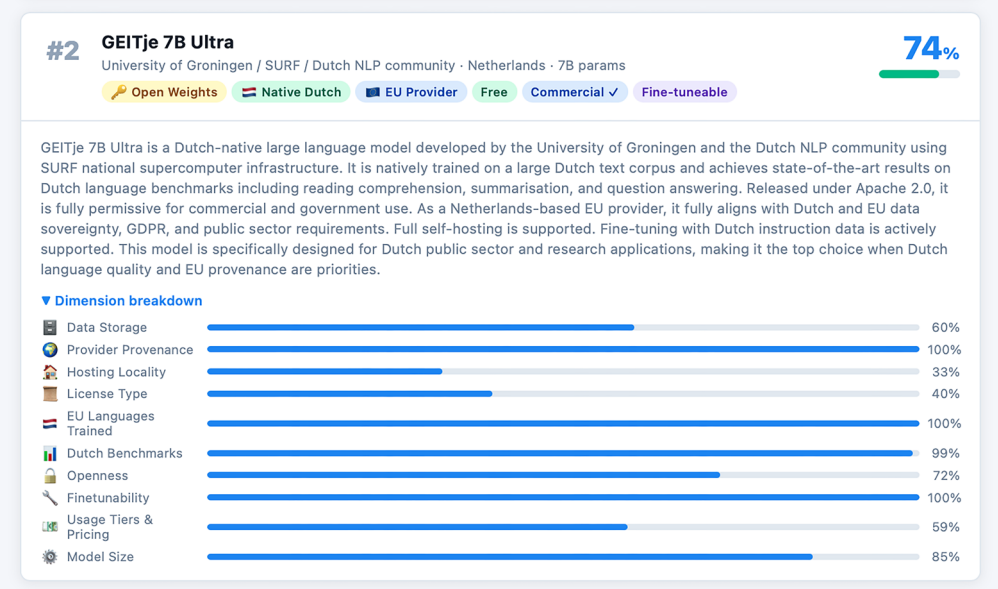
Figuur 4: dialoog over de waarden.

TNO Open LLM Advisor

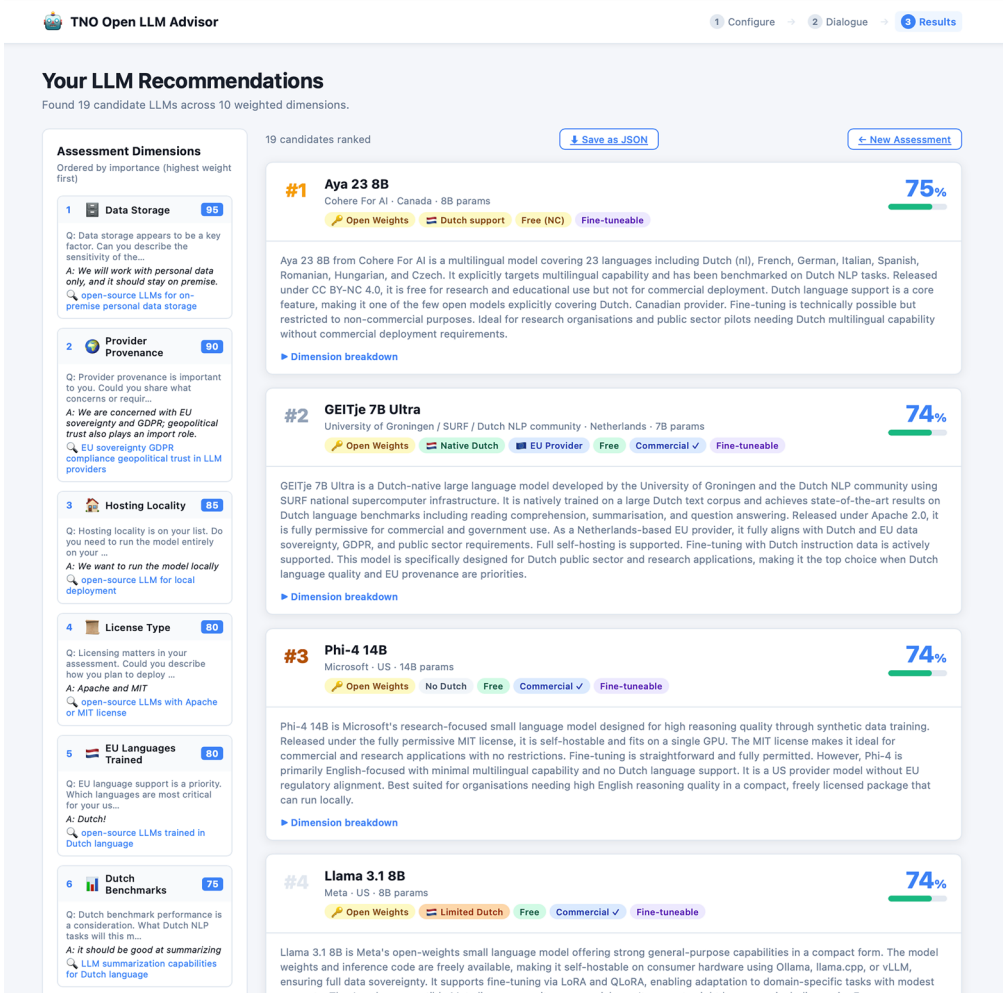
Resultaten

Tenslotte, na het doorlopen van alle vragen, toont het systeem de geordende lijst van LLMs (Figuur 5). Aan de linkerkant van dit scherm is de hele dialoog te zien (vraag/antwoord paren). Tevens kan de dialoog worden geëxporteerd naar JSON, voor verdere verwerking, per gebruikersnaam.

De getoonde LLMs kunnen worden ‘opengeklapt’ (‘Dimension breakdown’), zodat de scores per waarde getoond worden (Figuur 6).



Figuur 6: Scores per waarde per LLM.



Figuur 5: De geordende lijst van LLMs.

TNO Open LLM Advisor

Export en technisch ontwerp

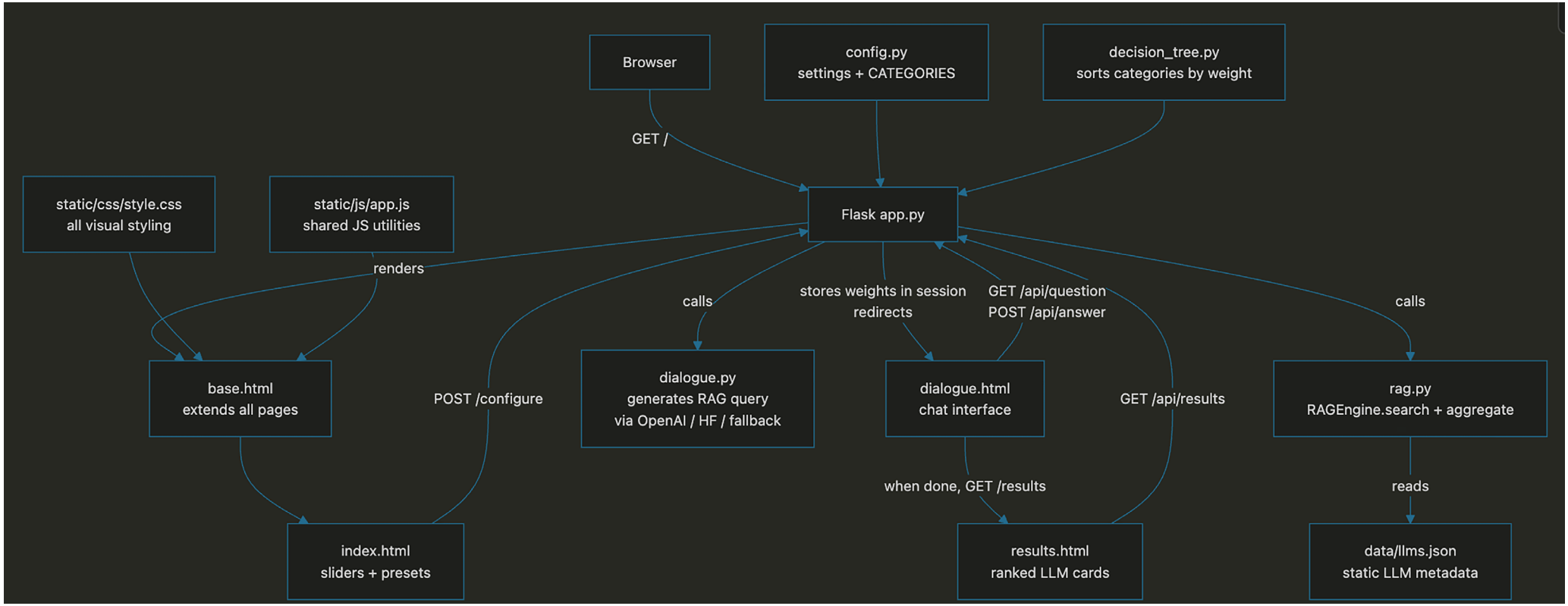
De geëxporteerde JSON ziet er als volgt uit (Figuur 7).


```
{
  "exported_at": "2026-06-22T14:21:26.843935Z",
  "user_name": "Stephan",
  "assessment": [
    {
      "rank": 1,
      "category": "data_storage",
      "label": "Data Storage",
      "weight": 95,
      "question": "Data storage appears to be a key factor. Can you describe the sensitivity of the data you plan to process with this
model? For example, does it involve personal data, confidential government information, or publicly available text? Must data remain wi
thin the Netherlands, within the EU, or entirely on-premises?",
      "answer": "We will work with personal data only, and it should stay on premise.",
      "rag_query": "open-source LLMs for on-premise personal data storage"
    },
    {
      "rank": 2,
      "category": "provider_provenance",
      "label": "Provider Provenance",
      "weight": 90,
      "question": "Provider provenance is important to you. Could you share what concerns or requirements drive this? For example, is t
his about EU data sovereignty, GDPR compliance, geopolitical trust (e.g., avoiding Chinese or US providers), or alignment with Dutch or
 EU public sector procurement policies?",
      "answer": "We are concerned with EU sovereignty and GDPR; geopolitical trust also plays an import role.",
      "rag_query": "EU sovereignty GDPR compliance geopolitical trust in LLM providers"
    },
    {
      "rank": 3,
      "category": "hosting_locality",
      "label": "Hosting Locality",
      "weight": 85,
      "question": "Hosting locality is on your list. Do you need to run the model entirely on your own infrastructure or within specifi
c geographic boundaries? For example, are you looking for a fully local deployment, an EU-region cloud setup, or would a managed global
 cloud service be acceptable?",
      "answer": "We want to run the model locally",
      "rag_query": "open-source LLM for local deployment"
    },
    {
      "rank": 4,
      "category": "license_type",
      "label": "License Type",
      "weight": 80,
      "question": "Licensing matters in your assessment. Could you describe how you plan to deploy or use this model? For instance, wil
l it be embedded in a commercial product, used for internal government services, or purely for research? Are there specific license typ
es — like Apache 2.0 or MIT — that you require or want to avoid?",
      "answer": "Apache and MIT",
      "rag_query": "open-source LLMs with Apache or MIT license"
```


Figuur 7: Geëxporteerde JSON, met vermelding van tijd en gebruiker (optioneel).

TNO Open LLM Advisor
Figuur 8 toont het technische schema van de demo-applicatie, die is opgezet als een lokale (Python) Flask-applicatie, en in een lokale browser draait.

Analyse
De verzamelde data (per gebruiker) kan dienen om slider-standen (en daarmee het relatieve belang van de waarden in het waardenstelsel) te middelen per organisatie. Tevens biedt dergelijke data een interessant perspectief op de perceptie van het waardenstelsel in een organisatie als BZK: zo kan door de tijd heen gezien worden of personen anders gaan denken over bepaalde waarden, en kan ook worden opgevangen of bepaalde waarden wellicht ontbreken, of met elkaar gecorreleerd zijn. Dergelijke analyses maken nu geen deel uit van deze demonstrator.



Figuur 8: Technisch schema van de demo-applicatie

Auteurs
Jens van der Weide
Sophie van Baalen
Gabriela Bodea
Marianne Schoenmakers
Stephan Raaijmakers

TNO | Vector

vector.tno.nl