VALIDATIEKADER

‘Verantwoorde inzet Large Language Models (LLMs) voor publieke informatievoorziening’

Inhoudsopgave

Klik op het hoofdstuk om hier direct naar toe te navigeren.

Om gebruik te maken van alle functionaliteiten in deze interactieve PDF, maak je het beste gebruik van een PDF viewer, zoals Adobe Reader.

- Organisatorische waarborgen
- Applicatie ontwerp
- Evaluatie
- Evaluatie
- Aanvullende informatie

Dit project is mede mogelijk gemaakt door:

SIDN fonds

TOPSECTOR ICT dutch digital

Validatiekader ‘Verantwoorde inzet Large Language Models (LLMs) voor publieke informatievoorziening’

Dit validatiekader ondersteunt organisaties bij de verantwoorde inzet van Large Language Model (LLM)-toepassingen voor publieke informatievoorziening. AI-modellen, zoals GPT-5.2, Llama en GPT-NL, bieden kansen om informatie breed beschikbaar te stellen aan burgers, organisaties en ambtenaren. Er zijn echter ook risico’s. Voordat de technologie kan worden ingezet, moet daarom worden nagegaan of de LLM-toepassing naar behoren werkt. Dit validatiekader helpt hierbij.

Wat is publieke informatievoorziening?

Het beschikbaar stellen van toegankelijke, betrouwbare en begrijpelijke informatie over publieke dienstverlening, beleid, wet- en regelgeving voor burgers, organisaties en ambtenaren.

Waarom een validatiekader?

Het verstrekken van publieke informatie is een kerntaak van overheidsinstellingen. Ethische waarborgen staan hierbij centraal, ook wanneer LLMs bij deze taak ondersteunen.

LLMs zijn een vorm van generatieve AI. In tegenstelling tot traditionele AI-toepassingen, zoals classificatie-algoritmes, kan zowel het type input als output van een LLM variëren. Deze flexibiliteit biedt nieuwe mogelijkheden voor geautomatiseerde verwerking van gegevens. Tegelijkertijd brengt het ook unieke risico’s met zich mee, bijvoorbeeld met betrekking tot de privacy van de eindgebruiker of het genereren van ongepaste output. Ook het evalueren van LLM-toepassingen is hierdoor uitdagend.

Bij de overweging om gebruik te maken van een LLM dient te worden stilgestaan bij de maatschappelijke gevolgen. Zo kost gebruik van een LLM veel rekenkracht, energie en data. Omdat slechts enkele commerciële partijen hiertoe in staat zijn, brengt gebruik van LLMs vaak afhankelijkheden met zich mee. Daarnaast zijn er vaak auteursrechtelijke vragen of data waar LLMs op getraind zijn rechtmatig zijn verkregen en zijn de data in sommige gevallen verzameld of gelabeld door datawerkers wiens arbeidsrechten beperkt worden gerespecteerd.

Uitgangspunten

Voor de verantwoorde inzet van LLMs worden de volgende kernprincipes als uitgangspunt genomen:

- **Doelmatig:** De LLM-toepassing dient een helder doel en is geschikt om dit doel te bereiken.
- **Zorgvuldige risicobeheersing:** Context-specifieke risicobeheersmaatregelen, zoals *guardrails* en *prompting*, bieden kwaliteitswaarborgen en mitigeren risico’s.
- **Structurele evaluatie:** De kwaliteit en risico’s van de LLM-toepassing worden structureel getest en gemonitord.
- **Transparantie**: Interacties met het LLM bieden een transparante gebruikerservaring.
- **Interoperabiliteit:** De afhankelijkheid van commerciële LLMs en (Amerikaanse) cloudplatforms wordt, met het oog op Europese soevereiniteit, geminimaliseerd.
- **Milieuimpact:** Bij de inzet van een LLM wordt de impact op mens en milieu geïnventariseerd en meegenomen in de afweging om de technologie in te zetten.

Leeswijzer

Opbouw en reikwijdte van het validatiekader

In het validatiekader worden voor 12 verschillende dimensies best practices beschreven hoe LLMs verantwoord kunnen worden ingezet ter ondersteuning van publieke informatievoorziening. De best practices zijn gebaseerd op praktijkervaring en wetenschappelijke inzichten die zijn opgedeeld in drie secties: **Organisatorische waarborgen**, **Applicatie ontwerp** en **Evaluatie**. In de eerste twee secties zijn best practices thematisch gebundeld. In de **Evaluatie** sectie worden risico's eerst geïdentificeerd, waarna er thematisch een beheersmaatregel wordt beschreven, welke kan worden gebruikt om de losse componenten en de LLM-toepassing als geheel te evalueren. Het kader is aangevuld met relevante informatie over onder meer de AI-verordening en specifieke evaluatiemethoden. Dit kader biedt geen uitputtende lijst met waarborgen die verantwoorde inzet van LLMs garanderen, maar is een hulpmiddel voor ontwikkelteams om na te gaan hoe bepaalde risico's die gepaard kunnen worden beheerst. Aspecten met betrekking tot de informatiebeveiliging van LLM-toepassingen vallen buiten de reikwijdte van dit kader.

Voor wie is dit validatiekader bedoeld?

Dit validatiekader is bedoeld voor **multidisciplinaire teams** binnen overheindsorganisaties die zich bezig houden met de (mogelijke) ontwikkeling van een LLM-toepassing voor publieke informatievoorziening. Het kader bevat aandachtspunten voor zowel **bestuurders** als **ontwikkelteams.** Het succesvol doorlopen van het kader vereist daarnaast input van en nauwe samenwerking met **belanghebbenden**, **eindgebruikers**, en **domeinexperts**.

Het vakgebied van LLM-toepassingen ontwikkelt zich snel. Best practices beschreven in dit kader met betrekking tot beheersmaatregelen en evaluaties moeten daarom doorlopend worden gespiegeld aan nieuwe inzichten uit de praktijk.

Ontwikkelproces LLM-toepassing

De ontwikkeling van een LLM-toepassing voor publieke informatievoorziening volgt een generiek ontwikkelproces. Dit proces, dat ook wordt gevolgd voor andere software en AI-toepassingen, wordt hier uitgewerkt voor de ontwikkeling van een LLM-toepassing. Ten behoeve van de leesbaarheid wordt het ontwikkelproces lineair weergegeven. In de praktijk worden stappen vaak iteratief uitgevoerd. Het kan bijvoorbeeld zijn dat in Stap 3 blijkt dat er veranderingen nodig zijn in Stap 1 en 2.

- Organisatorische waarborgen
- Applicatie ontwerp
- Evaluatie
- Evaluatie
- Aanvullende informatie

1 Probleemdefinitie

Start met het bepalen van het **Doel**: is een LLM echt nodig of volstaat een simpeler alternatief? Tegelijkertijd wordt de **Governance** ingericht. Dit omvat het opstellen van een algoritmebeleid met heldere rollen, verantwoordelijkheden en een exitstrategie.

2 Ontwerp

Stel de **Doelen** en gebruikersvereisten vast. De technische blauwdruk wordt vastgelegd tijdens **Applicatie ontwerp** waarbij benodigde (LLM-)componenten en onderlinge interacties worden vastgesteld. De **Gebruikersinterface** wordt dusdanig ontworpen dat gebruikers begrijpen dat zij met AI interacteren en altijd menselijke ondersteuning kunnen vinden. Tijdens dit proces worden **Privacy**-waarborgen geïmplementeerd en worden vereisten uit de AI-verordening nageleefd (zie p 49-51).

3 Model engineering en productontwikkeling

Selecteer een **Large Language Model** dat past bij de gedefinieerde vereisten. Ontwerp de **Systeemprompt** iteratief zodat het LLM gericht wordt ingezet. Geïdentificeerde risico's worden beheerst met behulp van **Guardrails**, die onder meer de input en output van en naar het LLM controleren. Wanneer de LLM-toepassing gebruik maakt van een zoekfunctionaliteit, dient een **Kennisbank en zoekstrategie** te worden bepaald.

4 Testen en valideren

Voer een **Pre-productie evaluatie** uit voordat de LLM-toepassing in gebruik wordt genomen. Deze evaluatie is anders dan de wijze waarop andere AI-systemen worden geëvalueerd. Losse componenten van de LLM-toepassing, maar ook de toepassing als geheel, worden geëvalueerd. Evaluatie vergt onder meer context-specifieke evaluatiedatasets en grondige evaluaties door domeinexperts en eindgebruikers om de prestaties en impact te bepalen van de ontwikkelde toepassing.

5 Ingebruikname en monitoring

Als een LLM-toepassing in gebruik is genomen, is doorlopende **Monitoring** van de toepassing vereist. De effectiviteit van de **Guardrails** moet worden bijgehouden en door middel van meetbaar criteria wordt de betrouwbaarheid van het gehele systeem gecontroleerd.

6 Onderhoud en itereren

Op basis van **Praktijkevaluatie** en **Monitoring** worden, indien nodig, de **Systeemprompt**, **Guardrails** en/of componenten van de LLM-toepassing aangepast. Bij elke wijziging in de keuze van de **Large Language Model** of de **Kennisbank en zoekstrategie** wordt de **Pre-productie evaluatie** herhaald om te voldoen aan de opgestelde prestatievereisten en wordt nagegaan of binnen de gestelde kaders voor het **Doel** en de **Governance** wordt gebleven.

Levenscyclus geïnspireerd op de levenscyclus uit ISO/IEC 22989:2022

Basisprincipes evaluatie van LLM-toepassingen

Grondige evaluatie staat centraal in dit validatiekader. Verantwoorde inzet van een LLM-toepassing betekent dat alle individuele componenten en de toepassing als geheel zorgvuldig zijn ontworpen en geëvalueerd. De volgende pagina's bieden achtergrondinformatie relaterend aan het evalueren van een LLM-toepassing. In de sectie Evaluatie worden evaluatiemethoden in meer detail besproken als onderdeel van een algemeen toepasbaar stappenplan.

Wat houdt evaluatie van een LLM-toepassing in?

Evaluatie is het systematisch beoordelen hoe goed de LLM-toepassing presteert, bijvoorbeeld op juistheid, betrouwbaarheid, bruikbaarheid, veiligheid en efficiëntie, ten opzichte van opgestelde prestatievereisten. De keuze voor een geschikte evaluatiemethode hangt af van het type taak die de LLM-toepassing uitvoert, de gebruikte LLM-componenten, de impact van de output van de LLM-toepassing en de beschikbaarheid van een *benchmark dataset*. Ieder element, inclusief gebruikerstesten, stress testing en red teaming, wordt hieronder toegelicht.

Opstellen van prestatievereisten en prestatiemetrieken

Iedere prestatievereiste moet worden geëvalueerd door middel van een zorgvuldig vastgestelde *prestatiemetriek*. In de context van publieke informatievoorziening kunnen bijvoorbeeld de volgende prestatievereisten relevant zijn: de output is gebaseerd op relevante bronnen, de samenvatting is correct en compleet, en de respons is begrijpelijk voor de beoogde eindgebruiker. Prestaties kunnen enkel worden gemeten wanneer zeer precies wordt gedefinieerd wat een prestatiemetriek beoogt te meten. Bijvoorbeeld: wat wordt precies bedoeld met een 'correcte' samenvatting of een 'begrijpelijke' respons?

Evalueren van LLM-componenten

Een geschikte evaluatiemethode hangt af van de taak die een LLM-component uitvoert. Taken van LLM-componenten zijn ruwweg in twee categorieën in te delen:

- ***Classificatie* of *voorspelling*:** Een LLM-gedreven privacy-guardrail voor het maskeren van persoonsgegevens (PII-masking) is een classificatietaak: bevat een stuk tekst wel of geen persoonsgegeven? (zie ook PV)
- **Generatie:** Sommige LLM-toepassingen genereren content, zoals tekstuele reacties in een chat. Het evalueren van generatieve taken is complex, omdat de gegenereerde tekst vrijwel onbeperkt kan variëren.

Het is voor de evaluatie niet van belang of classificatie is geïmplementeerd door middel van een classificatiemodel of door prompting.

Basisprincipes evaluatie van LLM-toepassingen

Evalueren met behulp van benchmark datasets

Een *benchmark dataset* is een zorgvuldig geselecteerde verzameling testsituaties die ondersteunt bij het evalueren van een LLM-component met betrekking tot een specifieke doelstelling. Testsituaties zijn relevant en realistisch voor de specifieke context waarbinnen de LLM-toepassing wordt gebruikt. Voor iedere situatie uit de benchmark dataset wordt door middel van een *prestatiemetriek* bepaald of de LLM-component op gepaste wijze met de situatie omgaat. Op basis van deze score kunnen de prestaties van verschillende **LLMs** met elkaar worden vergeleken. Voorbeelden van prestatiemetrieken voor LLM-toepassing zijn accuracy, precision en recall. Er bestaan verschillende soorten benchmark datasets:

- **Situaties met *ground truth labels*:** Classificatie- en voorspellingstaken worden geëvalueerd met behulp van *ground truth labels*. Dit zijn door experts vastgestelde labels over de (on)juistheid van een situatie. Prestaties worden bepaald door de output van een LLM-component te vergelijken met het ground truth label uit de benchmark dataset. Door middel van een prestatiemetriek wordt de LLM-toepassing op de gehele dataset geëvalueerd. De benchmark dataset van een privacy-guardrail die persoonlijke informatie maskeert bestaat bijvoorbeeld uit chatberichten waarbij voor ieder woord in het chatbericht is aangegeven of het wel/niet een persoonsgegeven is. Afhankelijk van het type applicatie en de te meten prestatievereiste kunnen ground truth labels voor testsituaties worden gelabeld door deskundige beoordelaars, eindgebruikers, of niet-experts (zie kader pagina 6). Ook kunnen bestaande open source benchmark datasets, zoals BBQ, worden gebruikt. Vaak is het opstellen van context-specifieke benchmark datasets vereist om een LLM-toepassing te evalueren. Zodra de ground truth labels bekend zijn kan de evaluatie geautomatiseerd plaatsvinden in een CI/CD teststraat (zie ook **PPE**).
- **Situaties zonder *ground truth labels*:** Voor generatietaken bestaat een benchmark dataset in principe uit testsituaties zonder ground truth labels. Dit is bijvoorbeeld het geval wanneer een LLM-toepassing teksten samenvat. Een gegenereerde samenvatting zal steeds anders zijn, dus de kwaliteit moet steeds opnieuw worden beoordeeld. Toch is het nog steeds nuttig om voorbeelduitkomsten uit te werken. Dit helpt om een geschikte prestatiemetriek te definiëren en kan dienen als voorbeeld voor beoordelaars. Evaluatie kan plaatsvinden door beoordeling door deskundigen, gebruikers, of niet-experts of door middel van de LLM-as-a-judge evaluatiemethode (zie kader pagina 8).

Gebruikerstesten

Verschillende soorten testen kunnen worden gebruikt om gebruikersinteractie met een component van de LLM-toepassing of met de applicatie als geheel te testen.

- **Component-test:** Gebruikers worden gevraagd om een specifieke testsituatie te beoordelen voor (een) specifieke prestatievereiste(n). Voor component-testen hoeft de toepassing niet af te zijn. Het is mogelijk dat gebruikers bijvoorbeeld samenvattingen voorgelegd krijgen met de vraag deze te beoordelen op leesbaarheid zonder dat zij zelf interacteren met de LLM-toepassing.
- **Gecontroleerd experiment (A/B test):** Tijdens een gecontroleerd experiment worden één of meerdere varianten van (een component in) de applicatie met elkaar vergeleken om te bepalen welke variant het beste presteert voor een specifieke prestatiemetriek.
- **Bèta-test:** Middels een bèta-test wordt in een realistische setting gebruiksvriendelijkheid en prestaties van de LLM-applicatie getoetst. Via deze testwijze kunnen ook onverwachte problemen tijdens productie worden geïdentificeerd.

Verschillende manieren om gebruikerstesten met een LLM-toepassing uit te voeren worden in het kader op pagina 8.

Basisprincipes evaluatie van LLM-toepassingen

Stresstesten, red-teaming en edge-case testen

Tijdens evaluatie moet ook de omgang van de LLM-toepassing met extreme scenario's worden getest. Dit zijn scenario's die buiten de verwachte gebruikspatronen vallen, zoals zeldzame situaties en controversiële situaties. Op basis van deze tests kunnen zwakke punten van de LLM-toepassing geïdentificeerd worden, zoals onverwacht gedrag. Voor stresstesten en edge-casetesten kunnen separate datasets (met of zonder *ground truth labels*) worden gebruikt. Naar deze vorm van evaluatie wordt verwezen als *red-teaming* en wordt meestal door een specialistisch team uitgevoerd, om mogelijk (systematisch) misbruik, zoals *promptinjectie*, *jailbreaken*, of anderzijds ongewenste output te toetsen.

Vijf manieren om de output van een LLM-toepassing te beoordelen

1. **Deskundige beoordelaars:** De gouden standaard is beoordeling door deskundigen (domeinexperts of ervaringsdeskundigen). Deze evaluatiemethode is zeer nauwkeurig, maar doorgaans ook duur en traag. In de meeste gevallen worden *ground truth labels* gecreëerd door deskundige beoordelaars.
2. **Beoordeling door gebruikers:** Wanneer de prestatievereiste is verbonden met de bruikbaarheid van de toepassing voor reguliere gebruikers is het van belang dat outputs door gebruikers beoordeeld worden. Zorg voor een diverse groep gebruikers die gezamenlijk representatief zijn voor de doelgroep van de LLM-toepassing.
3. **Beoordeling door niet-experts:** In sommige toepassingsdomeinen is het mogelijk om niet-experts (zoals de ontwikkelaars of externe platformwerkers) testsituaties te laten beoordelen. Deze methode is vaak laagdrempeliger dan de inzet van deskundige beoordelaars of eindgebruikers, maar ook minder nauwkeurig. Wanneer toch gebruik wordt gemaakt van beoordeling door niet-experts, bijvoorbeeld in de initiële opstartfase van een project, dan moet worden afgewogen of op den duur deze werkwijze moet worden aangevuld met oordelen van deskundigen en gebruikers.
4. **Automatische vergelijking met *ground truth labels*:** Wanneer betrouwbare data en labels beschikbaar zijn, bijvoorbeeld uit een benchmark datasets, dan kunnen de uitkomsten van LLM-componenten automatisch worden vergeleken met *ground truth labels*. Deze evaluatiemethode is zeer nauwkeurig en goedkoop wanneer een goede dataset beschikbaar is.
5. **LLM-as-a-judge:** Bij deze evaluatiemethode wordt een andere LLM gebruikt om de output van een LLM-component te beoordelen op basis van duidelijke, vooraf opgestelde beoordelingscriteria in een prompt. Dit is een zeer schaalbare evaluatiemethode, maar ook minder nauwkeurig. Bij het inzetten van *LLM-as-a-judge* moet eerst worden vastgesteld of het oordelende LLM de output van de genererende LLM consistent en nauwkeurig evalueert. Dit gebeurt door de scores te toetsen aan specifieke prestatievereisten en de mate van overeenstemming met menselijke experts te meten. Om de generaliseerbaarheid van de resultaten te kunnen beoordelen is het aan te raden om, net als bij reguliere data-experimenten, de dataset op te splitsen in een validatieset en testset.

Evaluatie (E) p. 18

Stap 1 **Ontwerp evaluatieproces**
Stap 2 **Benchmark dataset**
Stap 3 **Gebruikerstesten**
Stap 4 **Stresstesten, red teaming en edge-case testen**

Doel (D) p. 10

- **D.1** Doelen en doelgroep van de LLM-toepassing
  - **a.** Probleemanalyse
  - **b.** Organisatorische doelen en gevolgen
  - **c.** Maatschappelijke doelen en gevolgen
  - **d.** Toepassingsgebied en doelgroep
- **D.2** Afstemming met gebruikers
- **D.3** Geschikt middel
  - **a.** Alternatieven
  - **b.** Vergelijken geschikte methoden
- **D.4** Identificeer domein-specifieke risico's
- **D.5** Proportionaliteitsafweging

Large Language Model (LLM) p. 21

- **LLM.1** Kies het juiste type model
- **LLM.2** Overweeg een lokaal taalmodel indien gepast, of mogelijk een parallele of extern gehoste LLM
- **LLM.3** Stel het gekozen taalmodel af via RAG of fine tuning
- **LLM.4** Zorg voor een exitstrategie zodat het gebruikte model gemakkelijk vervangen kan worden
- **LLM.5** Evalueer met gebruikers en domeinexperts (zie E)

Kennisbank en zoekstrategie (KZ) p. 35

- **KZ.1** Bepaal geschikte bronnen die gebruikers helpen om in de informatiebehoefte te voorzien
- **KZ.2** Bepaal de geschikte zoekfunctionaliteit
  - a. Bepaal hoe bronnen opgeknipt worden (chunking)
  - b. Kies full text, semantic of hybrid zoekstrategie
- **KZ.3** Bepaal de geschikte LLM voor de geïdentificeerde zoekfunctionaliteit
- **KZ.4** Evalueer de werking van de kennisbank en zoekstrategie (zie E)

Guardrails (GR) p. 39

- **GR.1** Ontwerp guardrails (risicobeheersmaatregel)
  - **a.** Voor input van gebruiker naar LLM-applicatie
  - **b.** Voor filteren van persoonsgegevens
  - **c.** Voor output van LLM-applicatie naar gebruiker
  - **d.** Voor technische waarbogen
- **GR.2** Test de ontworpen guardrails (zie E)

Governance (GO) p. 13

- **GO.1** Beheer middels bestaand algoritmebeleid
- **GO.2** Uitvoeren DPIA en IAMA
- **GO.3** Vastleggen van rollen en verantwoordelijkheden
  - **a.** Tijdens ontwikkeling/pilot
  - **b.** Tijdens gebruik
- **GO.4** Exitstrategie om te stoppen met LLM-toepassing

Gebruikersinterface (UI) p. 32

- **UI.1** Transparantie over LLM-interactie, model, reikwijdte en beperkingen
- **UI.2** Veilige, toegankelijke invoer via een eigen front end met privacy by design
- **UI.3** Verifieerbaarheid via bronverwijzingen en uitleg van methode
- **UI.4** Verifieer of gebruikers verstrekte informatie lezen en begrijpen
- **UI.5** Laagdrempelig contact met een medewerker
- **UI.6** Eenvoudige melding van fouten en zichtbare opvolging
- **UI.7** Richt evaluatieproces in (zie E

Monitoring (M) p. 44

- **M.1** Houd constant zicht op prestaties en gebruikersinteracties
  - **a.** Leg een monitoringsplan vast
  - **b.** Stel een incidentenstrategie op
  - **c.** Ontwikkel een monitoringsplatform
- **M.2** Leg operationele prestaties en gebruik vast
- **M.3** Monitor gebruik en effectiviteit van toegepaste guardrails
- **M.4** Monitor, indien gewenst, de gebruikerinteractie met de LLM-toepassing
- **M.5** Ga na hoe gebruikers de gebruikersinterface ervaren

Praktijkevaluatie (PE) p. 47

- **PE.1** Korte termijn evaluatie
  - **a.** Ga na of doelen gerealiseerd zijn
  - **b.** Bepaal gebruikerservaring en vertrouwen
  - **c.** Ga de operationele efficiëntie na
- **PE.2** Lange termijn evaluatie
  - **a.** Worden kernproblemen opgelost?
  - **b.** Is er sprake van neveneffecten en systeemimpact?
  - **c.** Zijn er waarborgen voor organisatorische en maatschappelijke gevolgen?

Applicatie ontwerp (AO) p. 15

- **AO.1** Informatievoorzieningsproces en afbakening
- **AO.2** Systeemarchitectuur
  - **a.** Selecteer geschikte bronnen
  - **b.** Toepassingsfunctie LLM
  - **c.** Kwaliteit en risicobeheersing by design
- **AO.3** Operatie en kosten

Pre-productie evaluatie (PPE) p. 42

- **PPE.1** Stel evaluatieproces op (zie E)
- **PPE.2** Maak gebruik van CI/CD
- **PPE. 3** Valideer de werking met belanghebbenden

Systeemprompt (SP) p. 25

- **SP.1** Stel een gedetailleerde systeemprompt op en identificeer contextuele kernconcepten die in de prompt benoemd worden
- **SP.2** Stel een evaluatieproces (zie E)
  - **a.** Maak gebruik van CI/CD

Privacy (PV) p. 29

- **PV. 1** Verwerkingsgrondslag en -overeenkomst
- **PV.2** Dataminimalisatie
- **PV.3** Voorkom toegang van LLM-aanbieder tot prompt en chatgeschiedenis gebruikers
- **PV.4** Prioriteer eigen infrastructuur
- **PV.5** Ga wettelijke naleving na met juridische experts

Doel (D)

Taalmodellen zijn niet altijd geschikt om publieke informatievoorziening te verbeteren. Per context moet worden afgewogen of de inzet van een LLM-toepassing wenselijk is gegeven het beoogde Doel (D). Er dient onder meer te worden nagegaan waarom een simpeler alternatief, zoals een klassieke zoekoplossing of extra menselijke ondersteuning, geen oplossing biedt. Hierbij moeten zowel organisatorische als maatschappelijke gevolgen worden meegewogen.

Best practices

#	Titel	Beschrijving
D.1	**Doelen en doelgroep van LLM-toepassing**	Identificeren van doelen, gevolgen en doelgroep van LLM-toepassing, inclusief onderliggende probleemanalyse. **De uitkomst van deze stap zou kunnen zijn dat een LLM-toepassing niet de geschikte oplossing is voor het geïdentificeerde probleem.**
a	Probleemanalyse	▪ Stel het probleem vast waar de LLM-toepassing ondersteuning bij zou moeten bieden. ▪ Analyseer de omvang van het probleem (kwantificeer indien mogelijk). ▪ Beschrijf de oorzaken van het probleem.
b	Organisatorische doelen en gevolgen	▪ Specificeer bij welke kerntaak van de organisatie de LLM-toepassing ondersteuning biedt, bijvoorbeeld het faciliteren van begrijpelijke uitleg over het recht, wetgeving of beleidsprocessen. ▪ Beschrijf het doel en het gewenste effect van de LLM-toepassing op de organisatie en de werkprocessen (kwantificeer indien mogelijk). ▪ Ga na dat het kennisniveau van medewerkers over werkprocessen behouden blijft.
c	Maatschappelijke doelen en gevolgen	▪ Specificeer welke maatschappelijke doelen de toepassing heeft. Denk aan het toegankelijker maken van overheidsprocessen of het verbeteren van de informatiepositie van burgers. ▪ Ga na welke mogelijke (indirecte of onbedoelde) gevolgen de toepassing kan hebben. Denk aan verminderd vertrouwen in de overheid, energieverbruik en milieu-impact, afhankelijkheid van commerciële partijen, de impact van schadelijke dataverzamelingspraktijken (bijv. met betrekking tot IP of contentmoderatie). ▪ Ga na of de LLM-applicatie wordt ingezet voor een beeldbepalende overheidstaak (zoals contact met burgers) of niet (zoals uitleg over wet- en regelgeving richting ambtenaren). ▪ Indien sprake is van een beeldbepalende overheidstaak moet extra zorg worden besteed aan kwaliteitsvereisten.

Doel (D)

#	Titel	Beschrijving
d.	Toepassingsgebied en doelgroep	▪ Beschrijf de specifieke taken van de LLM-toepassing in lijn met de in **a-c** beschreven doelen (denk aan: samenvatten, vragen beantwoorden, herschrijven in begrijpelijke taal, etc.). ▪ Bepaal wie de LLM-toepassing zal gaan gebruiken. ▪ Bepaal de grenzen van het toepassingsgebied: identificeer taken die buiten de reikwijdte van de toepassing vallen. **Voor sommige LLM-toepassingen voor publieke dienstverlening is het bijvoorbeeld van belang dat de toepassing enkel feitelijke informatie verschaft en geen advies geeft of aanbevelingen doet over hoe te handelen.**

D.2	**Afstemming met gebruikers**	▪ Betrek gebruikers om hun informatiebehoefte beter te begrijpen. Specificeer informatiebehoefte van gebruikers (ophalen van informatie, presenteren van informatie of anders).
		▪ Vraag aan gebruikers aan welke vereisten het middel ten behoeve van informatievoorziening moet voldoen. Scherp op basis van de informatie de gegevens verkregen in **D.1** aan.
		  ▪ Is bijvoorbeeld hoge kennisdichtheid gewenst of moeten complexe thema's in begrijpelijke taal worden uitgelegd?
		  ▪ Hoe moeten bronnen worden vermeld, e.g. in de lopende tekst of in een voetnoot?
		▪ Selecteer domeinexperts die betrokken blijven bij mogelijke ontwikkelproces van LLM-toepassing (zie **KZ.3** en **PPE.2**).

D.3	**Geschikt middel**	Ga na of LLM-toepassing een geschikt middel is voor doel en doelgroep zoals geïdentificeerd in **D.1a**

a	Alternatieven inventariseren	▪ Breng alternatieve methoden voor de geïdentificeerde doelen in kaart. Denk hierbij in ieder geval aan:
		  ▪ Contact met een medewerker (telefonisch, chat, fysiek)
		  ▪ Verbeterde (klassieke) zoekfunctionaliteit
		  ▪ Regelgebaseerde logica, beslisbomen, keuzemenu's, handleidingen en/of FAQ
		  ▪ Handmatig geschreven samenvattingen of statische met behulp van LLMs geproduceerde, handmatig gecontroleerde en gepubliceerde samenvattingen
		  ▪ (Q&A-)forum

b	Vergelijken geschikte methoden	▪ Onderbouw of een LLM-toepassing, vergeleken met geïdentificeerde alternatieven, een geschikt middel is voor het bereiken van het gestelde doel. Vergelijk hierbij in ieder geval: ▪ Effectiviteit, risico's (zie ook **D.4**), kosten, ontwikkeltijd en onderhoud

Doel (D)

#	Titel	Beschrijving

D.4	**Domeinspecifieke risico's**	Indien een LLM-toepassing als geschikt middel is aangemerkt in **D.3**:
		▪ Identificeer risico's die gepaard gaan bij gebruik van de LLM-toepassing in de specifieke context. Overweeg hierbij ten minste:
		  ▪ Feitelijkheid van de gegenereerde output (o.a. hallucinaties, niet bestaande bronnen)
		  ▪ Privacy (o.a. verwerkingsgrondslag, verwerking van persoonsgegevens)
		  ▪ Ongewenst gebruik van de toepassing (o.a. agressief taalgebruik, gebruik voor onbedoelde taken, *jailbreaken*)
		  ▪ Representativiteit en bias in gegenereerde output (o.a. stigmatiserende beeldvorming, ongelijke behandeling in gelijke gevallen)
		▪ Ga na hoe groot de kans is dat risico's voorkomen en bepaal de impact.
		
		De impact van risico's is afhankelijk van de doelen. Zo zijn privacywaarborgen belangrijker wanneer persoonsgegevens door de LLM-toepassing worden verwerkt of een persoonlijke situatie wordt uitgevraagd (omdat in dit geval de gebruiker ongevraagd persoonsgegevens zou kunnen delen). Wanneer de LLM-toepassing enkel bedoeld is voor het doorzoeken van informatie middels een kennisbank, speelt bijvoorbeeld privacy mogelijk een minder belangrijke rol.

D.5	**Proportionaliteit**	▪ Neem de in **D.1**-**D.4** verkregen informatie mee om af te wegen of het proportioneel is om te beginnen met de ontwikkeling van een LLM-applicatie:
		  ▪ Zet alle geïdentificeerde nadelen voor organisatie, samenleving en betrokkenen af tegen de mate waarin het beoogde doel wordt bereikt door inzet van de LLM-toepassing.
		  ▪ Neem hierbij mee of er (simpelere) alternatieven zijn waarmee de geïdentificeerde doelen, of vergelijkbare doelen, gerealiseerd kunnen worden.

Merk op: de uitkomst van deze stap kan zijn dat de inzet van een LLM-toepassing niet geschikt is voor het beoogde doel.

Governance (GO)

Taalmodellen zijn niet altijd behulpzaam om publieke informatievoorziening te verbeteren. Per context moet worden afgewogen of de inzet van een LLM-toepassing wenselijk is gegeven het beoogde Doel (D). Er dient onder meer te worden nagegaan waarom een simpeler alternatief, zoals een klassieke zoekoplossing of extra menselijke ondersteuning, geen oplossing biedt. Hierbij moeten zowel organisatorische als maatschappelijke gevolgen worden meegewogen.

Best practices

#	Titel	Beschrijving
GO.1	**Algoritmebeleid**	▪ Pas de toepassing in in bestaand algoritmebeleid van de organisatie ▪ Bij ingebruikname van de toepassing: publiceer de LLM-toepassing in het landelijke Algoritmeregister Inspiratie voor algoritmebeleid kan worden gevonden in het Algoritmekader van het Ministerie van Binnenlandse Zaken en Koninkrijksrelaties, Code Goed Digitaal Openbaar Bestuur (CODIO) voor de Rijksoverheid en het AI Governancekader van de VNG

GO.2	**DPIA en IAMA**	▪ Voer een risicoclassificatie uit voor de LLM-toepassing. Indien er sprake is van impact op de betrokkenen, voer een (pre-)IAMA uit:
		  ▪ Betrek hierbij opgehaalde informatie uit **D.1-D.5**
		  ▪ Wanneer de LLM-toepassing voor een hoog risico toepassingen uit de AI-verordening wordt ingezet moet de overheidsinstelling de Fundamentele Rechten Impact Assessment (FRIA) van de AI-verordening uitvoeren. Zie nader pagina 29.
		▪ Wanneer gewenst of verplicht: voer een DPIA uit. Zolang de inzet van LLM-toepassingen voor de organisatie nieuw en innovatief is, wordt het uitvoeren van een DPIA aangeraden.
		  ▪ Ga na of een DPIA moet worden uitgevoerd volgens intern beleid of volgens de door de Autoriteit Persoonsgegevens (AP) gepubliceerde criteria.
		  ▪ Nadere vereisten in het kader van privacy zijn opgenomen in de dimensie **PV**.

GO.3	**Rollen en verantwoordelijkheden**	Vastleggen van rollen en verantwoordelijkheden binnen het ontwikkel- en beheerteam Besteed extra tijd aan deze stap als de LLM-toepassing wordt ingekocht

Governance (GO)

#	Titel	Beschrijving
a	Tijdens ontwikkeling/pilot	▪ Stem eindverantwoordelijkheid over functioneren van LLM-applicatie al in de ontwikkelfase af met bestuurders ▪ Wijs een verantwoordelijke aan die onder meer prestatiemetrieken analyseert, (indien hiertoe is besloten) chatgeschiedenis van gebruikers analyseert (zie nader **PV.3**) en een beslissing neemt over ingebruikname na pre-productie evaluatie (**PE**) ▪ Betrek domeinexpertise en gebruikers bij ontwikkeling en evaluatie van LLM-toepassing: ▪ Voor het vaststellen van de informatiebehoefte van de doelgroep (zie **D.2**) ▪ Voor het inrichten van de **kennisbank en zoekstrategie (KZ)**, opstellen van **guardrails (GR)** en **systeemprompt (SP)**, uitvoeren van **pre-productie evaluatie (PPE)** en **praktijkevaluatie (PE)**

b	Tijdens gebruik	▪ Bespreek welke verantwoordelijkheden bij bestuurders, beheerders of het ontwikkelteam belegd zijn. Zoals:
		  ▪ Monitoring en communiceren van behaalde prestaties en nieuwe (risico's) conform opgestelde doelen, zoals gespecificeerd in **D.1**, bijvoorbeeld middels vast rapportage met vaste regelmaat
		  ▪ Afleggen van (publieke) verantwoording over gebruikte technologie

GO.4	**Exitstrategie**	▪ Stel een plan op als het gebruikte AI-model of de gehele LLM-toepassing niet meer ingezet kan worden

Applicatie ontwerp (AO)

Het Applicatie ontwerp (AO) richt zich op het vertalen van doelen en risico's naar ontwerpkeuzes voor de architectuur en algehele werking van de LLM-toepassing. In deze stap wordt ook het technische toepassingspatroon van LLM-componenten gekozen, zoals chat, *guardrails*, *Retrieval Augmented Generation* (RAG) of agents. Deze keuzes komen samen in de systeemarchitectuur, die beschrijft hoe alle componenten (data, modellen, API's) met elkaar interacteren.

Best practices

#	Titel	Beschrijving
AO.1	**Informatievoorzienings-proces en afbakening**	▪ Beschrijf het beoogde informatievoorzieningsproces en de rol van de LLM-applicatie binnen dit proces (zie D). Beschrijf in ieder geval welke informatie zal worden verzameld en verwerkt en denk na over interacties met externe systemen. ▪ Vertaal de doelen (gedefinieerd in D) naar (technische) functionaliteiten en (sub)taken voor de applicatie.
AO.2	**Systeemarchitectuur**	Bepaal uit welke componenten de LLM-toepassing bestaat en hoe deze met elkaar interacteren. Denk hierbij aan zowel front-end als back-end componenten, inclusief de gebruikersinterface, servers, (vector) databases, API's, functionaliteiten, en LLM-component(en) (via een API-service of lokaal gehost). Houd rekening met schaalbaarheid en onderhoudbaarheid van het volledige systeem.
a	Bronnen	▪ Bepaal welke bronnen worden gebruikt om gebruikers in hun informatiebehoefte te voorzien. Houd bij deze selectie onder meer rekening met: feitelijke correctheid, kwaliteit en actualiteit van bronnen, aanwezigheid van persoonsgegevens. ▪ Bepaal hoe de bronnen technisch worden geraadpleegd, bijvoorbeeld via een applicatie-specifieke vector database of via een API naar een bestaande, mogelijk externe, database.

Applicatie ontwerp (AO)

#	Titel	Beschrijving

b	Toepassingsfunctie LLM	▪ Bepaal welke functionaliteiten of taken worden door een LLM-component worden uitgevoerd.
		▪ Bepaal voor ieder LLM-component een geschikt toepassingspatroon voor het uitvoeren van de (sub)taak. Denk bijvoorbeeld aan:
		  ▪ Chatbot: Bijvoorbeeld voor het uitvragen van de informatiebehoefte van de gebruiker.
		  ▪ *Retrieval Augmented Generation:* Voor het uitvoeren van simpele zoektaken.
		  ▪ LLM agent: Voor het uitvoeren van complexere taken waarvoor planning en/of gebruik van tools vereist zijn.
		  ▪ Indien wordt gekozen voor een agent, bepaal dan hoe het geheugenbeheer (memory management) van de agent wordt ingericht. In de context van publieke informatievoorziening zullen simpele strategieën vaak voldoen, maar in enkele gevallen (bijvoorbeeld bij meerdere of lange gebruikersinteracties) kan extern geheugen noodzakelijk zijn.
		▪ Classificatie: Voor de implementatie van een guardrail of LLM-as-a-judge evaluatiemethoden.

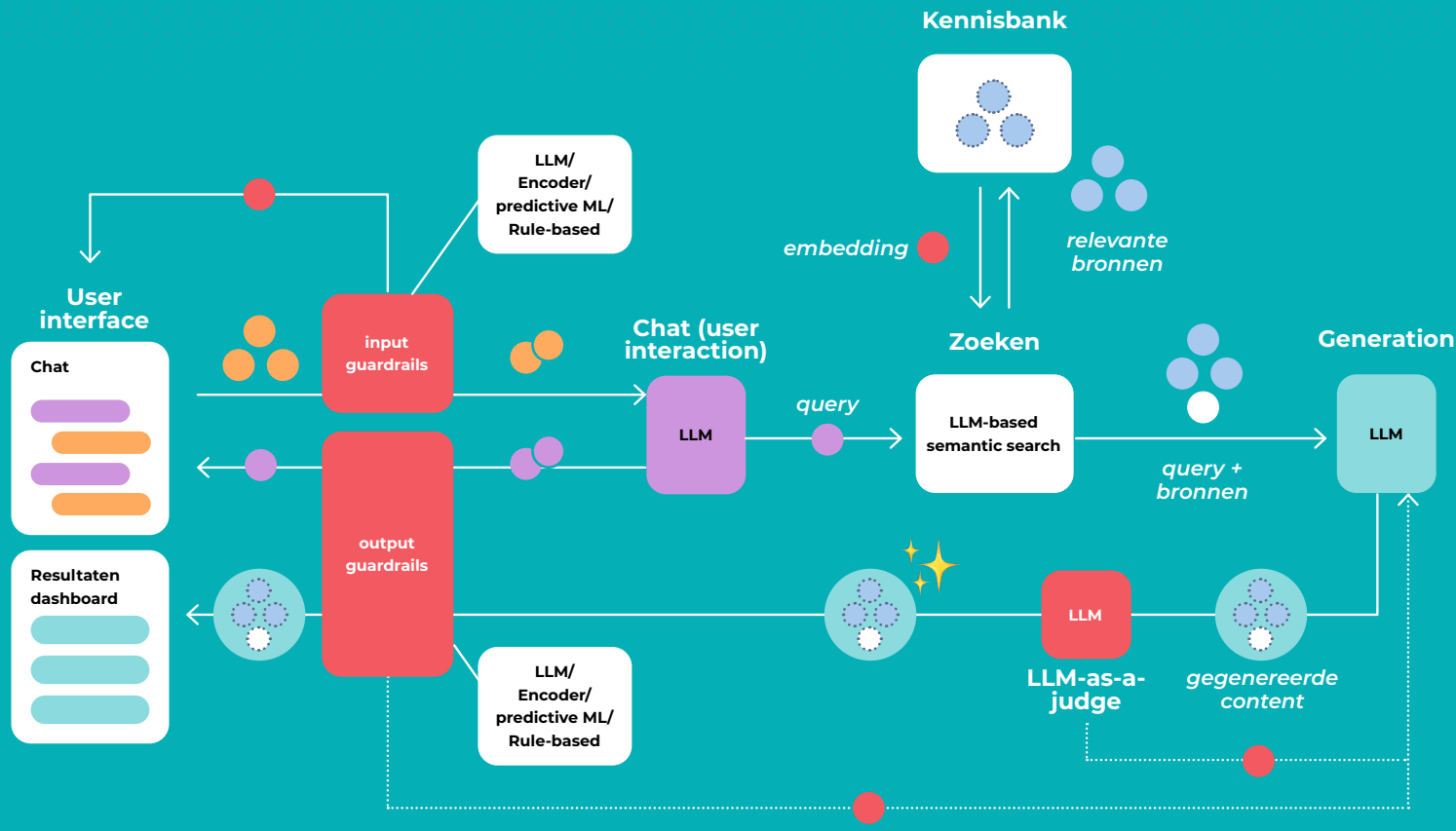
c	Kwaliteit en risicobeheersing by design	▪ Bepaal voor ieder component welke technische risicobeheersmaatregelen worden ingezet om kwaliteit en verantwoorde inzet te waarborgen. Denk in ieder geval aan de inzet van guardrails, systeemprompts, en LLM-as-a-judge. Zie nader **GR**, **SP**, **E**.
AO.3	**Operatie en kosten**	▪ Schat de kosten voor benodigde cloud resources, LLM-API kosten, data opslag, en applicatieontwikkeling. ▪ Heroverweeg op basis van (technische) haalbaarheid, kosten, doelen en risico's (zie ook **D**) of het proportioneel is de ontwikkeling van de LLM-applicatie voort te zetten.

Voorbeeld: voorRecht-rechtspraak

Applicatie ontwerp

De voorRecht-rechtspraak toepassing is opgebouwd uit verschillende componenten. Eerst wordt door middel van een chat interface informatie bij de gebruiker uitgevraagd. Een LLM-component stuurt op basis van de opgehaalde informatie een beeld van de situatie naar de gebruiker ter controle. Indien de gebruiker zich in de geschetste situatie kan vinden, wordt met behulp van een RAG-applicatie relevante bronnen bij de situatie gezocht en wordt relevante informatie uit deze bronnen samengevat. Verschillende *guardrails* beoordelen en filteren de outputs.

LLMs worden in verschillende stappen van het applicatie ingezet: in de chat- en samenvattings-component, de zoekfunctionaliteit en de *guardrails*.



voorRecht Rechtspraak

voorRecht-rechtspraak is ontwikkeld door de Rechtspraak in samenwerking met T&T Data Consultancy en Deloitte.

Gebruikersinput
Chatcomponent
Search (RAG)
Guardrails

Stappenplan voor het evalueren van LLM-toepassingen

Evaluatie staat in de deze sectie en in bijbehorende dimensies centraal. Voortbouwend op de algemene introductie over het evalueren van LLM-toepassingen (pagina 6-8), wordt hier een algemeen toepasbaar stappenplan toegelicht aan de hand waarvan LLM-toepassingen geëvalueerd kunnen worden.

Stap 1 – Ontwerp evaluatieproces

Ontwerp een evaluatiemethodiek voor iedere prestatievereiste van de LLM-toepassing. Dit kan betrekking hebben op één specifieke component of op een combinatie van componenten uit het Applicatie ontwerp (AO). Neem hierbij de volgende aspecten mee:

- **LLM-component en risicocategorie:** Selecteer een component uit de LLM-toepassing en bepaal de risicocategorie (zoals: maskeren van persoonsgegevens, toetsen van vooringenomenheid/bias of leesbaarheid van output).
- **Evaluatiemethode:** Ga na of gebruik gemaakt dient te worden van een *benchmark dataset* of gebruikerstest. Zie p. 8 en Stap 2 en Stap 3.
- **Beoordelaar:** Bepaal wie outputs van de LLM-toepassing beoordeelt (bijvoorbeeld deskundige beoordelaars, gebruikers, niet-experts, *LLM-as-a-judge*). Zie p. 8.
- **Prestatievereiste:** Bepaal en beschrijf de variabele die wordt gemeten om de prestatievereiste te toetsen (bijvoorbeeld: gebruikerstevredenheid op een schaal van 1 tot 5).
- **Prestatiemetriek:** Bepaal hoe de prestatievereiste wordt gekwantificeerd (bijvoorbeeld: accuracy, precision@k, gemiddelde gebruikerstevredenheidscore).
- **Drempelwaarde:** Indien relevant, bepaal een geschikte drempelwaarde voor de prestatiemetriek (bijvoorbeeld: “bij normaal gebruik is *recall* minimaal 0.85”).

Stap 2 – Benchmark dataset

Bepaal of een standaard op maat gemaakte benchmark dataset wordt gebruikt. Stel vereisten op voor de testsituaties en de bijbehorende oordelen in de gebruikte dataset.

Stap 2.1 Testsituatie:

Testsituaties moeten voldoende representatief en divers zijn voor de werkelijkheid. Houd hierbij rekening met zowel voorzien en onvoorzien gebruik en met misbruik. Voor het verzamelen van testsituaties kunnen verschillende bronnen worden geraadpleegd:

- **Op maat gemaakte benchmark dataset:** Een initiële verzameling testsituaties kan vaak worden verzameld door het ontwikkelteam (niet-experts). LLMs kunnen in dit stadium ondersteuning bieden, bijvoorbeeld bij het genereren van chattranscripties voor een specifieke situatie. Om representativiteit en kwaliteit van de testsituatie te borgen, moeten in latere stadia deskundigen of (beoogde) eindgebruikers worden betrokken.
- **Standaard benchmark dataset:** In sommige gevallen kan gebruik worden gemaakt van een standaard benchmarking dataset, zoals BBQ. In veel gevallen zijn deze echter niet voldoende specifiek voor een toepassing-specifieke evaluatie.

Besteed extra aandacht aan representativiteit van testsituaties als een prestatievereiste is gekoppeld aan een drempelwaarde bij normaal gebruik. Voorkom evaluatie met perfecte data in benchmark dataset.

Stap 2.2 Oordelen:

- **Type beoordelaar:** Testsituaties worden bij voorkeur gelabeld door deskundigen (domeinexperts of ervaringsdeskundigen). Bij eenvoudige taken kunnen ook niet-experts labelen. Zorg voor voldoende deskundigheid en diversiteit van menselijke beoordelaars.
- **Definitie prestatievereiste:** Geef beoordelaars een heldere definitie van prestatievereisten (relevantie van gevonden documenten, aanwezigheid persoonsgegevens etc.).

Stappenplan voor het evalueren van LLM-toepassingen

- **Open source data:** Wanneer wordt gekozen voor een bestaande benchmark dataset, ga na of de testsituaties en oordelen vertrouwd kunnen worden.
- **LLM-as-a-judge:** Wanneer gebruik wordt gemaakt van LLM-as-a-judge, moet worden gevalideerd of de beoordelende LLM de genererende LLM consistent en nauwkeurig beoordeelt, bijvoorbeeld door de scores te vergelijken met scores van deskundigen.

Stap 2.3 Metadata:

Houd relevante meta-data, zoals de bron van een testsituatie, de beoordelaar en versie van de LLM-toepassing bij.

Stap 3 – Gebruikerstesten

- **Gebruikersgroep:** Zorg de de groep gebruikers die de interactie met de LLM-toepassing test divers en representatief is voor de gebruikersgroep. Zie nader p. 8.
- **Opdrachtbeschrijving:** Stel voor gebruikerstesten een duidelijke vraag en opdracht op, waarbij verschillende prestatievereisten zoveel mogelijk van elkaar gescheiden worden.

Voeg relevante testsituaties toe aan de benchmark dataset zoals opgesteld in Stap 2. Op deze manier kan bij een toekomstige wijziging getoetst worden of deze testsituaties nog steeds op gepaste wijze door de LLM-toepassing worden verwerkt.

Stap 4 – Stresstesten, *red-teaming* en edge-case testen

Monitoring: Incidenten en klachten die aan het licht komen tijdens monitoring (**M**) of gebruikerstesten kunnen worden toegevoegd aan de op maat gemaakte benchmark dataset.

Voorbeelden van evaluatiemethodes voor componenten uit LLM-toepassing

In onderstaande tabel zijn enkele voorbeelden uitgewerkt om gestructureerd een evaluatiemethode te documenteren. De rijen geven verschillende componenten uit een LLM-toepassing weer die geëvalueerd dienen te worden. De kolommen representeren de belangrijkste concepten uit het evaluatie-stappenplan zoals uitgewerkt in p. 19-20. Naar onderstaande tabel wordt ook wel verwezen als een evaluatie- of risico-taxonomie.

LLM-component	Risicocategorie	Type taak	Testsituatie	Oordeel	Beoordeling output van testsituaties	Ground truth labels	Bron ground truth labels	Metriek
Privacy guardrail	Privacy (persoonsgege-ven maskeren)	Classificatie	Chatberichten	Aanwezig (ja/nee)	Automatische vergelijking	Ja	Benchmark dataset/niet-experts	FPR, FNR
Zoekfunctie	Relevantie	Classificatie, ranking of generatie	Zoekqueries, relevante documenten in kennisbank	Relevant (ja/nee)	Automatische vergelijking	Ja	Deskundige beoordelaars/ niet- experts	precision@k
RAG generatie	Bertrouwbaar-heid	Generatie	(Gegenereerde samenvatting, originele bronnen)	Bevat informatie die niet in de bronnen zat (ja/ nee)	Deskundige beoordelaars, LLM-as-a-Judge	Nee	nvt	FPR, FNR
RAG generatie	Toegankelijkheid (leesbaarheid)	Generatie	(gegenereerde samenvatting, originele bronnen)	Schaal 1-5	Gebruiker, LLM-as-a-Judge	Nee	nvt	Gemiddelde score

Large Language Model (LLM)

Het gekozen Large Language Model (LLM) bepaalt in grote mate de functionaliteit en betrouwbaarheid van de component en de LLM-toepassing als geheel. Er bestaan veel verschillende typen LLM: grote cloudgebaseerde modellen, kleine lokale modellen die commercieel en open source worden aangeboden. Het gekozen model moet aansluiten bij het beoogde Doel (D) en de geformuleerde vereisten rondom Governance (GO) en Privacy (PV). LLMs worden in hoog tempo doorontwikkeld; wendbaarheid is daarom cruciaal. Om afhankelijkheden te vermijden, moet het Applicatie ontwerp (AO) zo worden ingericht dat de onderliggende taalmodellen relatief eenvoudig inwisselbaar zijn.

Identificeren

Doel: Bepaal de doelen van het LLM in lijn met de doelen zoals geïdentificeerd voor de LLM-toepassing in onder meer D en AO. Bij LLMs gaat het hier om de balans tussen prestaties, kosten en controle. Denk hierbij aan:

- **Prestaties:** Het LLM moet in staat zijn de taak accuraat en binnen acceptabele tijd uit te voeren.
- **Proportionaliteit:** De complexiteit van het model staat in verhouding tot de benodigde capaciteiten voor de beoogde taak.
- **Governance:** Het borgen van controle over de dataverwerking.
- **Duurzaamheid:** Neem de impact op mens en milieu mee in de keuze voor een specifiek taalmodel.

Risico's: Bepaal welke risico's van belang zijn voor de LLM-componenten. Denk daarbij aan de risico's zoals in D geïdentificeerd voor de LLM-toepassing. Denk hierbij tenminste aan:

- **Interoperabiliteit:** Het risico dat de applicatie te afhankelijk wordt van één specifieke leverancier of model-architectuur waardoor overstappen naar een andere LLM complex wordt.
- **Vertrouwelijkheid en IP:** Het risico dat de provider prompts en interacties met de API opslaat om eigen modellen te trainen.
- **Veroudering:** Te snelle doorontwikkeling van de leverancier wat kan zorgen voor instabiliteit of onverwachte output.

Large Language Model (LLM)

Evalueren

#	Titel	Beschrijving
LLM.1	**Type model**	▪ Ga na wat een geschikt type model is voor de beoogde taak. Er is niet altijd een LLM nodig, overweeg ook andere modellen (bijv. encoder only modellen of klassiekere voorspellende machine learning algoritmes voor classificatie- of zoekentaken). ▪ Het LLM kan per taak of component verschillen. Gebruik een zo passend mogelijk model met de juiste vaardigheden voldoende presteert voor de toepassing (denk aan samenvatten, vragen beantwoorden of redeneren). Zie ook **D.1c**. Neem hierbij ten minste vaardigheden, modelgrootte en energieverbruik in ogenschouw. **Indien wordt gekozen voor een open-weights model kunnen technieken worden toegepast om het model sneller of kleiner te maken.**
LLM.2	**Hosting**	▪ Kies voor een model dat voldoet aan de gestelde vereisten voor governance en privacy. Zie dimensies **GO** en **PV**. Maak hierbij een afweging tussen de kosten, prestaties en afhankelijkheden van zelf-gehoste en extern gehoste modellen: ▪ Zelf-gehost model: Biedt maximale controle over de verwerking van (bijzondere) persoonsgegevens en andere gevoelige informatie die niet eenvoudig te maskeren zijn. Zelf-gehoste modellen bieden ook meer ruimte voor maatwerk en training met eigen data. ▪ Extern gehost model: Indien toch gekozen wordt voor een extern model, probeer dan contractueel vast te leggen dat de provider *systeemprompt*, prompts van gebruikers en gebruikersinteractie niet opslaan of gebruiken voor trainingsdoeleinden.

Large Language Model (LLM)

#	Titel	Beschrijving
LLM.3	**Model afstellen**	▪ Stem de geselecteerde LLM af voor gebruik in het gewenste toepassingsgebied. Overweeg hierbij: ▪ *Retrieval Augmented Generation (RAG)*: Indien een eigen domeinspecifieke kennisbank wordt gebruikt (zie **KZ**). ▪ *Prompt engineering*: Aanpassen van systeemprompt om het gewenste gedrag van de LLM-toepassing te verbeteren (zie **SP**). ▪ *Fine tuning*: Voorbeelden worden gebruikt om het model te leren wat gewenst gedrag is. Er zijn verschillende vormen van fine tuning, waaronder het veranderen van alle gewichten in het model (full fine tuning), door het trainen van een klein deel van het model of door kleine modules toe te voegen. Bekende voorbeelden zijn Adapters en LoRA (Low-Rank Adaptation). Dit zijn vormen van fine tuning die snel en goedkoop zijn. ▪ Modelgrootte: om het model sneller of kleiner te maken voor gebruik, kunnen technieken worden toegepast zoals quantization (verminderen van precisie van de gewichten in het model), pruning (verwijderen van onnodige verbinden) en knowledge distillation (een klein model trainen op basis van een groter model). **Doorgaans geldt: des te meer de output van het model gestuurd moet worden naar een gewenst resultaat, des te groter de kans dat het LLM gefinetuned moet worden. Het is aan te raden eerst te onderzoeken of de gestelde doelen met prompt engineering kunnen worden behaald.**
LLM.4	**Exitstrategie en interoperabiliteit**	▪ Richt de LLM-toepassing, infrastructuur en testomgeving zo in dat, indien gewenst, de gekozen LLM gemakkelijk verwisseld kan worden voor een andere LLM. ▪ Ga hierbij na wat het effect is van deze verandering op de effectiviteit van de gehele LLM-toepassing. Zie nader dimensies pre-productie evaluatie (**PPE**) en monitoring (**M**). ▪ Ga na of er afhankelijkheden zijn met betrekking tot onderliggende
LLM.5	**Evaluatieproces**	▪ Ontwerp en implementeer een evaluatiemethode voor iedere specifieke taak en elk doel van het LLM (zie **E**). ▪ Ga met gebruikers en domeinexperts na of de gebruikte LLM voldoende presteert voor de gestelde prestatievereisten.

Voorbeeld: GPT-NL

Large Language Model

GPT-NL is een soeverein taalmodel dat specifiek is ontwikkeld voor de Nederlandse taal en cultuur. Het betreft uitsluitend het taalmodel (foundation model) zelf en niet het AI-systeem dat er eventueel omheen wordt gebouwd, zoals ChatGPT dat is voor GPT-5. De training en finetuning zijn volledig in eigen beheer en is uitgevoerd op Nederlandse infrastructuur. Het model is ontwikkeld op rechtmatig verkregen data waarvoor geen gebruik is gemaakt van data waarvoor geen toestemming is verleend door rechthebbenden. De dataverzameling is vanaf de basis opgebouwd en is gecontroleerd op persoonsgegevens, de bescherming van intellectueel eigendom en het uitsluiten van schadelijke inhoud zoals discriminerende of haatzaaiende teksten. Auteurs van gebruikte bronnen ontvangen een deel van de opbrengsten van GPT-NL via een zogenaamd 'content board'. Het model is toegespitst specifieke toepassingen waar taalmodellen vaak voor worden gebruikt, zoals: versimpelings-, samenvattings- en RAG-toepassingen. Daarnaast richt GPT-NL zich op het ontwikkelen en testen van nieuwe en bestaande Nederlandse *benchmark datasets*, biedt het transparantie over documentatie en ontwerpkeuzes conform subsidievoorwaarden en vormt het inspiratie voor toekomstige samenwerking om andere generatieve AI oplossingen van Nederlandse bodem te creëren.

GPT-NL

GPT-NL wordt ontwikkeld door TNO, SURF en het Nederlands Forensisch Instituut (NFI) en wordt in het kader van een soeverein Nederlands taalmodel ontwikkeld met financiering afkomstig vanuit de Rijksdienst voor Ondernemend Nederland (RVO) en het Ministerie van Economische Zaken

De meeste LLMs	GPT-NL
Gebruik van gescrapete data zonder expliciete toestemming	Rechtmatig verkregen data
Beperkte controle op auteursrechten en privacy	Strikte filters op persoonsgegevens, IP en schadelijke inhoud
Getraind op infrastructuur buiten Nederland en/of Europa	Volledig van scratch getraind op Nederlandse infrastructuur
Globale, generieke toepassingen	Gericht op publieke en maatschappelijke toepassingen: RAG, samenvatting, simplificatie
Generieke taalkundige en toepassingsgerichte benchmarks	LLM benchmarks voor Nederlandse toepassingen
In eigendom van internationale commerciële partijen, incl. governance	Content board voor opbrengstdeling met rechthebbenden
Weinig stimulans voor samenwerking nationale partijen, bijv. voor aanpassingen met betrekking tot lokale taal en cultuur	Bevorderen van samenwerkingen Nederlandse partijen: ▪ Toepassingen voor Nederlandse publieke en private partijen ▪ Opvolgende generatieve AI-initiatieven
Grijs gebied (gegevensbescherming, auteursrechtelijk etc.)	Naleving van Nederlandse en Europese regelgeving

Systeemprompt (SP)

De systeemprompt vormt het fundament voor interactie met een LLM. Het geeft sturing aan hoe input, zoals chats van gebruikers of externe bronnen, verwerkt worden. Het definieert gewenste output, en grenzen en is daarmee van groot belang voor de kwaliteit van de output. Daarnaast is de Systeemprompt (SP) na Guardrails (GR) een belangrijke maatregel om onjuiste, onvolledige of ongewenste output tegen te gaan.

Identificeren

Doelen en risico's: Bepaal doelen van de systeemprompt in lijn met de doelen en risicobeheersing van de LLM-toepassing, zoals onder meer gedefinieerd in D en AO. Denk in ieder geval aan:

Bedoeld gebruik

- **Accuraatheid:** Het juist beantwoorden of behandelen van relevante vragen of onderwerpen die binnen het vastgestelde domein vallen bij normaal gebruik.
- **Feitelijkheid:** Het systeem vermeldt geen niet-bestaande referenties en voegt geen informatie toe die niet in de originele bronnen aanwezig is (*hallucinaties*). Referenties verbinden de juiste bron aan de juiste informatie.
- **Toon en stijl:** Een consistente en passende toon en stijl van de output.
- **Structuur:** Een passende output-structuur voor verdere verwerking (bijv. JSON).
- **Diversiteit:** De kwaliteit van de output is onafhankelijk van de te verwachten taalvariaties van gebruikers (bijv. veelgebruikte talen en dialecten in de Nederlandse samenleving).
- **Gewenste interactie en inclusiviteit:** De output bevat geen beledigende, vooringenomen of anderszins onwenselijke teksten.
- **Privacy:** Persoonsgegevens van de gebruiker of derden zijn niet aanwezig in output.

Onbedoeld gebruik en misbruik

- **Scope:** De LLM-toepassing genereert output op vragen of onderwerpen die buiten het vastgestelde toepassingsdomein vallen.
  - Denk ook aan het identificeren van gevallen die te complex zijn om door het systeem te worden behandeld.
- **Manipulatie:** Afvangen van pogingen tot *promptinjecties*, *jailbreaking* en andere vormen van manipulatie van het systeem.
  - Denk bijvoorbeeld aan pogingen om ongewenste output te genereren, voor de gebruiker gunstigere misinterpretatie van wet- en regelgeving te genereren, of toegang te krijgen tot de systeemprompt, chathistorie of bronnen uit de onderliggende kennisbank.

Systeemprompt (SP)

Evalueren

#	Titel	Beschrijving

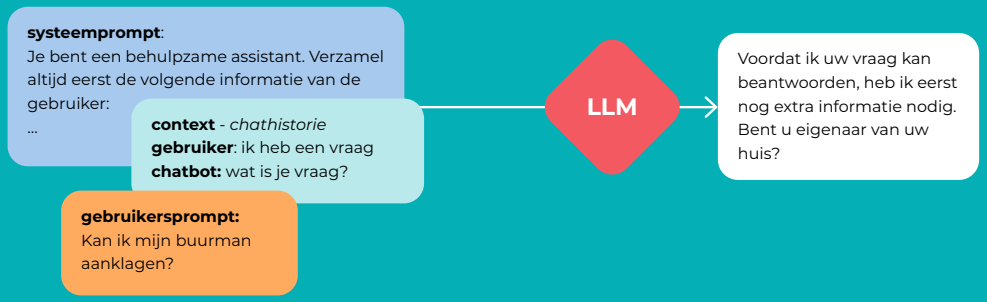
SP.1	**Prompting techniek**	▪ Schrijf een gedetailleerde systeemprompt waarin alle vastgestelde doelen zijn opgenomen. Denk hierbij in ieder geval aan:
		  ▪ Bepaal een geschikte prompting-structuur voor de geïdentificeerde doelen (bijvoorbeeld Chain-Of-Thought, ReAct, Few-Shot prompting).
		  ▪ Schrijf instructies voor het verwerken van meegegeven context, zoals transcripties van voorgaande gesprekken of relevante documenten uit de kennisbank (zie KZ).
		  ▪ Instrueer het LLM expliciet om aan te geven als de gestelde vraag niet op basis van de aangeleverde context of onderliggende bronnen beantwoord kan worden. Geef in de prompt expliciet aan te reageren met: “ik weet het niet” en dat het LLM niet een antwoord moet verzinnen.
		  ▪ Wanneer gebruik wordt gemaakt van een LLM agent, bepaal dan ook instructies over hoe het taalmodel moet communiceren met externe tools en services (bijv. via een Model Context Protocol server) en eventueel extern opgeslagen geheugen.
		▪ Diverse strategieën kunnen helpen in het verbeteren van de kwaliteit van de output. Denk bijvoorbeeld aan specifieke instructies voor:
		  ▪ Toon en stijl: “wees behulpzaam”, “geef antwoord in neutrale bewoordingen”
		  ▪ Scope: “beperk je tot praktische oplossingen”, “geef geen juridisch advies”
		  ▪ Structuur: “vraag naar onderbouwing of stapsgewijze uitleg”, “geef informatie terug in JSON format” (overweeg hier indien mogelijk gebruik te maken van machine leesbare output)
		  ▪ Impersonatie: “Neem de rol aan van een professor”.
		
		Het schrijven van een goede prompt vereist veel iteraties. Houd er rekening mee dat best-practices voor prompting continu veranderen. Strategieën die in het verleden werkten kunnen bij een update van een bepaald type LLM mogelijk niet meer werken. Het is daarom belangrijk een robuust evaluatieproces in te richten.

SP.2	**Evaluatieproces**	Zelfs kleine aanpassingen in een systeemprompt kunnen leiden tot ongewenste antwoorden die in eerdere iteraties zijn verholpen. Gebruikmaken van een weldoordacht evaluatieproces is daarom cruciaal. Zie E.
a	CI/CD	▪ Integreer de evaluaties in een *CI/CD*-werkwijze voor de systeemprompt. Dit is cruciaal om de LLM-toepassing snel aan te kunnen passen indien nodig.

Voorbeeld: systeemprompts

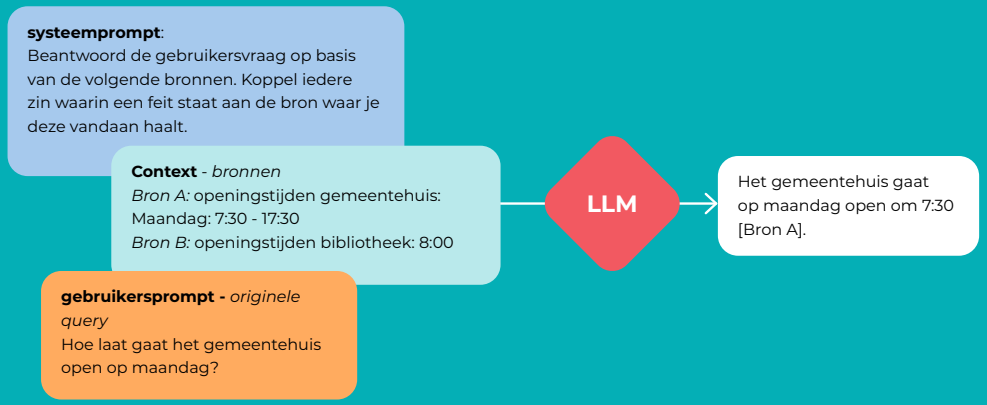
Er zijn over het algemeen meerdere typen prompts die in een LLM-toepassing worden gebruikt, zoals voor gebruikersinteractie, voor guardrails en mogelijk voor content generatie en een LLM-as-a-judge evaluatieproces. Voor iedere taak waarbij gebruik wordt gemaakt van een LLM dient een systeemprompt gebruikt te worden.

Gebruikersinteractie

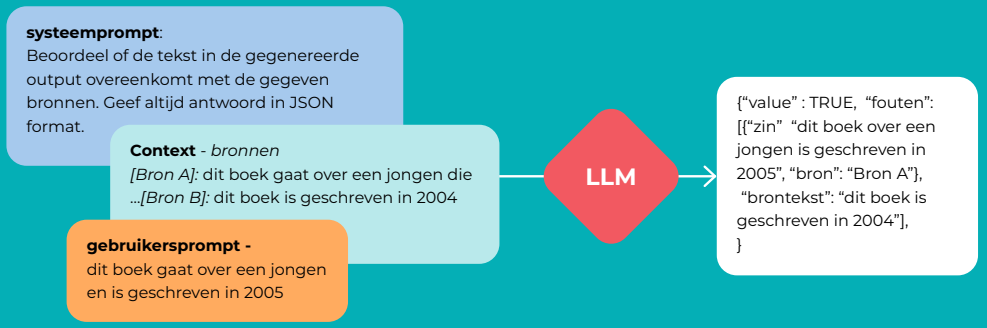


Guardrails

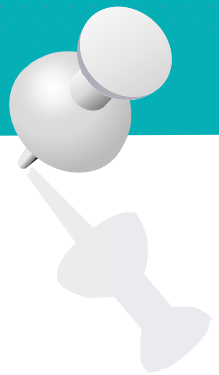
Content generatie



LLM-as-a-judge



Voorbeeld: gezocht



Draag je zelf bij aan de ontwikkeling van een LLM-toepassing en wil je een bijdrage aan het Validatiekader leveren? Stuur een tekstje van ca. 250 woorden over 1) Doel en 2) Praktijkevaluatie naar: info@algorithmaudit.eu

Privacy (PV)

De verwerking van persoonsgegevens door LLM-toepassingen ten behoeve van publieke informatievoorziening kan soms noodzakelijk zijn. Persoonsgegevens kunnen bijvoorbeeld onderdeel zijn van chatinteractie, metadata of relevante bronnen. Indien persoonsgegevens worden verwerkt in de LLM-toepassing moet aan de AVG worden voldaan om de Privacy (PV) van gebruikers te waarborgen.

Identificeren

Doelen en risico's: Bepaal doelen en risico's met betrekking tot privacy, in lijn met de doelen zoals geïdentificeerd in D. Omdat het onderwerp privacy geen technische component is met eigen functionaliteit, worden doelen en risico's hier samen behandeld.

Ga na welke privacy-risico's gelden. Maakt hierbij onderscheid tussen persoonsgegevens die bewust verwerkt worden (zoals IP-adres, cookies en/of metadata) en die mogelijk onbedoeld verwerkt worden (persoonsgegevens in de gebruikersinteractie). Denk tenminste aan:

- **Privacy van de gebruiker:**
  - Overmatig verzamelen van persoonsgegevens, bijv. middels metadata
  - Gebruikers delen persoonsgegevens van zichzelf of anderen tijdens interactie met de chat
  - Gebruik van cookies
  - Persoonsgegevens van gebruikers worden gedeeld met externe partijen zoals leveranciers.
- **Privacy van derden:**
  - Persoonsgegevens die in interne data van een organisatie aanwezig zijn worden gedeeld met gebruikers of externe partijen, bijvoorbeeld doordat deze data aanwezig zijn in de kennisbank.
- **Naleving van de AVG:** De AVG stelt specifieke eisen aan de verwerking van persoonsgegevens, documentatie en transparantie. Denk ook aan verschillende juridische verantwoordelijkheden van partners en leveranciers voor de gebruikte LLM-toepassing.

Privacy (PV)

Evalueren

#	Titel	Beschrijving
PV.1	**AVG grondslag, verwerkingsdoel en overeenkomst**	▪ Documenteer welke persoonsgegevens de LLM-toepassing verwerkt. Maak hierbij onderscheid tussen gegevens die bewust verwerkt worden (zoals IP-adres of andere metadata) en die mogelijk onbedoeld verwerkt worden (persoonsgegevens in de gebruikersinteractie). ▪ Leg de verwerkingsdoelen en verwerkingsgrondslag vast. Let op: dit kan verschillen per persoonsgegeven. ▪ Sluit een verwerkersovereenkomst af wanneer meerdere partijen betrokken zijn bij de verwerking van persoonsgegevens (zoal een externe software leverancier).

PV.2	**Dataminimalisatie**	▪ Verwerk zo min mogelijk persoonsgegevens als nodig is voor het beoogde doel. Dit betekent dat zo min mogelijk metadata verwerkt worden (zoals IP-adres). Wanneer het nodig is om deze gegevens wel te verwerken, zorg dan dat de gegevens zo kort mogelijk worden opgeslagen
		▪ Zorg dat persoonsgegevens die een gebruiker, onbedoeld of onnodig, met een LLM-toepassing deelt niet of zo min mogelijk worden verwerkt, bijvoorbeeld door:
		  ▪ Gebruikers te waarschuwen geen namen, adressen en andere persoonlijke informatie te delen.
		  ▪ Persoonsgegevens te maskeren met behulp van contentfilters (zie **GR**).

PV.3	**Toegang tot prompt en chatgeschiedenis gebruikers**	▪ Richt de LLM-toepassing zo in dat de aanbieder van het LLM geen toegang heeft tot:
		  ▪ De systeemprompt van de LLM-toepassing
		  ▪ De chatgeschiedenis van de specifieke gebruikers en de verzameling van alle gebruikersinteracties met de LLM-toepassing, incl. metadata.
		▪ Sla, indien mogelijk, chatgeschiedenis op binnen de infrastructuur van de organisatie die de LLM-toepassing gebruikt en ga na wat de bewaartermij.n
		
		Pas de 'principle of least privilige' toe: gebruikers, systemen en processen krijgen alleen de minimale toegangsrechten die nodig zijn om hun taken uit te voeren, om zo de kans op misbruik of schade te beperken

PV.4	**Prioriteer eigen infrastructuur**	▪ Richt de LLM-toepassing zo in dat zoveel mogelijk componenten op de infrastructuur van de organisatie die de LLM-toepassing gebruikt worden gehost.
PV.5	**Wettelijke naleving**	▪ Ga na of met de genomen maatregelen de privacy van personen voldoende beschermd is en of aan de AVG is voldaan, in lijn met de vastgestelde risico's (bijvoorbeeld in de DPIA, zie **GO**). Betrek hierbij relevante privacy functionarissen.

Voorbeeld: voorRecht-rechtspraak

Governance

De pilot voorrecht-Rechtspraak wordt beheerd volgens het algoritmebeleid van de Rechtspraak. Dit houdt in dat voor de pilot- en ontwikkelfase het team afspraken heeft gemaakt met verantwoordelijke bestuurders over de projectplanning, beslis- en keuzemomenten. Aangezien de pilot burgers ten alle tijden online helpt met het delen van informatie van juridische conflicten wordt de applicatie dagelijks door een beheerder gemonitord op prestatievereisten, zoals time-outs en geactiveerde contentfilters. Resultaten en mogelijke reputatierisico's worden wekelijks door het ontwikkelteam met een bestuurlijk verantwoordelijk teamlid gedeeld.

Tijdens ontwikkeling van de pilot is met een diverse groep belanghebbenden een risico-inschatting gemaakt op basis van een vereenvoudigde versie van de Impact Assessment Mensenrechten en Algoritmes (IAMA). Daaruit bleek dat voorRecht-rechtspraak een laag-risico toepassing betreft en het uitvoeren van een volledige IAMA niet noodzakelijk werd geacht. Daarnaast worden domeinexperts doorlopend bij het functioneren van de LLM-toepassing betrokken, bijvoorbeeld voor de evaluatie hoe de *RAG-applicatie* omgaat met randgevallen en voor het opstellen en onderhouden van een geschikte *benchmark dataset*. Het intellectueel eigendomsrecht is door de Rechtspraak van ondersteunende marktpartijen afgekocht.

voorRecht Rechtspraak

voorRecht-rechtspraak is ontwikkeld door de Rechtspraak in samenwerking met T&T Data Consultancy en Deloitte.

Privacy

Om risico's van de pilot met experimentele LLM-technologie met betrekking tot gegevensverwerking in kaart te brengen is een Data Protection Impact Assessment (DPIA) uitgevoerd. Hierin zijn de verwerkingsgrondslag, de verwerkingsdoeleinden en belangen van betrokkenen bij de gegevensverwerking geëxpliciteerd. Om persoonsgegevens te maskeren uit gebruikte jurisprudentie in de kennisbank zijn eigen masking scripts geschreven met behulp van Stanza. Contentfilters van Azure worden gebruikt om haatzaaiende, initimiderende en kwetsende content te detecteren en hier een passende reactie op te formuleren. Vergelijkbare Azure-filters worden gebruikt om de LLM-toepassing te beschermen tegen *prompt injecties*.

De Rechtspraak heeft vastgelegd dat Microsoft geen toegang heeft tot verschillende systeemprompts van de LLM-toepassing en de chatgeschiedenis van gebruikers met de chatbot. Om de pilot effectief te kunnen evalueren en de LLM-applicatie doorlopend te kunnen verbeteren is ervoor gekozen de chatgeschiedenis van gebruikers op te slaan met een bewaartermijn van 60 dagen.

De tech stack van voorRecht-rechtspraak is zo ingericht dat gemakkelijk gewisseld kan worden tussen meer privacy-vriendelijke of Europees georiënteerde LLM-aanbieders. Zie nader LLM.

Gebruikersinterface (UI)

Gebruikers komen meestal alleen via een gebruikersinterface (UI) met een LLM in contact. Om gebruikers een prettige en veilige ervaring te bieden (bijvoorbeeld met een chatbot) is het belangrijk de gebruikersinterface zorgvuldig vorm te geven en af te stemmen op de behoefte van gebruikers. Zo moet uit de interface bijvoorbeeld altijd blijken dat gebruikers interacteren met een AI-model. Daarnaast moeten overheden bij hun kerntaken altijd menselijke ondersteuning aanbieden.

Identificeren

Doel: Bepaal de doelen voor de gebruikersinterface in lijn met de doelen zoals geïdentificeerd in onder meer D en AO. Denk hierbij aan specifieke doelen en hoe deze geëvalueerd kunnen worden voor aan de hand van relevante (gebruikers)testen (zie E):

- **Toegankelijkheid en begrijpelijkheid:** De interface is voldoende begrijpelijk en toegankelijk voor de beoogde gebruikers.
- **Transparantie:** Maak zichtbaar dat het antwoord door AI is gegenereerd. Toon details van het LLM; bied volledige bronvermelding met peildatum en link.
- **Feedback:** Neem vindbare en laagdrempelige manieren om feedback te geven op in de gebruikersinterface.

Risico's: Bepaal welke risico's van belang zijn voor de gebruikersinterface. Denk daarbij aan de risico's zoals in D geïdentificeerd voor de gehele LLM-toepassing. Denk hierbij ten minste aan:

- Misleiding door onduidelijke herkomst, peildata of onzekerheidsduiding.
- Privacy- en beveiligingslekken via invoer, logging of rendering.
- Niet-toegankelijke of onduidelijke UI (niet-*WCAG*, te complexe taal).
- Onvoldoende menselijke waarborgen of feedbackloops.

Gebruikersinterface (UI)

Evalueren

#	Titel	Beschrijving
UI.1	**Introductie en verwachtingen**	▪ Toon een melding dat antwoorden door AI gegenereerd zijn. Licht het doel en beperkingen van de inzet de LLM-toepassing toe. Vermeld de disclaimer niet alleen in de footer van de gebruikersinterface. **Het opnemen van een dergelijke melding is een verplichting volgend uit de AI-verordening (zie pagina 41-43).** ▪ Maak modeldetails zichtbaar: modelnaam, leverancier, versie en datum laatste update van de LLM-toepassing. ▪ Vermeld in het kader van de AI-verordening het beoogde doel van de LLM-toepassing en doeleinden waarvoor de toepassing niet mag worden gebruikt (met oog op onvoorziene hoogrisicotoepassingen, zie nader pagina 41-43). ▪ Gebruik B1-taalniveau waar mogelijk. Vermijd jargon en leg complexe begrippen uit met behulp van uitlegtools.
UI.2	**Invoer en interactie**	▪ Gebruik een eigen *front-end*: de interactie verloopt via een door de organisatie beheerde gebruikersinterface (geen directe vendor-UI), met controle op huisstijl, logging, privacy by design en beveiliging. ▪ Voldoe aantoonbaar aan *WCAG* 2.2 niveau AA voor alle kritieke gebruikersstromen (toetsenbordnavigatie, focus-zichtbaarheid, contrast, semantische structuur/ARIA, foutmeldingen in duidelijke taal). ▪ Waarschuw gebruikers om geen gevoelige- of persoonsgegevens te delen. Leg het principe van dataminimalisatie uit (zie PV.2).
UI.3	**Antwoord en transparantie over techniek en bronnen**	▪ Vermeld bronnen bij antwoorden. Denk aan: titel, korte beschrijving bron, publicatiedatum en directe link. ▪ Bied beknopte uitleg of en hoe *RAG*, *fine tuning* of *systeemprompts* worden toegepast. Vermeld peildatum bij veranderlijke informatie. ▪ Stel gebruikers in staat om bronnen voor een antwoord opnieuw op te vragen of antwoorden te exporteren.
UI.4	**Verificatie**	▪ Evalueer of gebruikers instructies, disclaimers en verstrekte informatie lezen en begrijpen.
UI.5	**Escalatie naar medewerker**	▪ Stel gebruikers in staat om contact op te nemen met een medewerker via een zichtbare keuzeoptie. Vermeld het communicatiekanaal, de beschikbaarheid van medewerkers en de verwachte responstijd van de medewerker. ▪ Verzorg contextuele overdracht met toestemming van gebruiker (gegevensdeling), incl. statusweergave van verzoek.

Gebruikersinterface (UI)

#	Titel	Beschrijving
UI.6	**Feedback en leren**	▪ Vraag gebruiker om feedback, bijvoorbeeld door middel van 👍en 👎feedbackknoppen om onnauwkeurige, verouderde, onduidelijke, partijdige of schadelijke content te rapporteren. ▪ De gebruiker kan ook gevraagd worden om specifieke passages te markeren om opvallende content onder de aandacht van de beheerders te brengen. ▪ Informeer de gebruiker hoe gedeelde feedback verwerkt wordt.
UI.7	**Evaluatieproces**	▪ Ontwerp en implementeer een evaluatiemethode voor ieder specifiek doel van de gebruikersinterface (zie E). Voor de meeste doelen is een gebruikerstest de meest geschikte methode om de gebruikersinterface te evalueren. Vaak kunnen meerdere doelen gecombineerd worden in een grotere test. ▪ Bepaal vooraf wat “goed” is voor de UI: bijvoorbeeld transparant (herkomst en bronnen duidelijk), bruikbaar (toptaken vlot uitvoerbaar), toegankelijk (volgens WCAG-richtlijnen), veilig (geen onbedoelde datadeling of injecties) en controleerbaar (herleidbare bronvermelding). **Voer bijvoorbeeld de volgende testen uit Taaktesten (“vind bron/peildatum”), begrijptest van disclaimer/modelinfo, en UI-audit op consistente, toegankelijke weergave.**

Kennisbank en zoekstrategie (KZ)

De zoekfunctionaliteit van een LLM-toepassing speelt een belangrijke rol bij het vinden van relevante bronnen die bij publieke informatievoorziening kunnen ondersteunen. De kwaliteit van de zoekfunctionaliteit hangt af van de kwaliteit van de gebruikte Kennisbank en zoekstrategie (KZ).

Identificeren

Doel: Bepaal doelen met betrekking tot de kennisbank en zoekstrategie in lijn met de doelen zoals geïdentificeerd in D en AO. Denk hierbij aan:

- **Kennisbank:** Kwaliteit van bronnen (bijv. zijn teksten geverifieerd op correctheid, actueel en getoetst op representativiteit voor de context).
- **Zoekfunctionaliteit:** Relevantie van opgehaalde bronnen.
- **Responsgeneratie:** Betrouwbaarheid van het door de LLM-toepassing gegenereerde antwoord.

Risico's: Bepaal welke risico's van belang zijn voor de kennisbank en zoekfunctionaliteit. Denk daarbij aan de risico's zoals in D geïdentificeerd voor de LLM-toepassing. Denk hierbij tenminste aan:

- **Betrouwbaarheid:** Referentie naar niet bestaande bronnen en misrepresentatie van informatie.
- **Privacy:** Het delen van persoonsgegevens aan de gebruiker via informatie ontsloten uit bronnen (datalek).
- **Bias:** Representativiteit in gebruikte bronnen wat kan leiden tot vooringenomenheid in gegenereerde output (bijv.: stigmatiserende beeldvorming, ongelijke behandeling in gelijke gevallen).
- ***Prompt injectie*:** Gebruikersinput waarmee door middel van instructies een LLM-toepassing misleid wordt om ongewenst gedrag te veroorzaken.
- **Actualiteit:** Referentie naar verouderde bronnen.

Kennisbank en zoekstrategie (KZ)

Evalueren

#	Titel	Beschrijving
KZ.1	**Geschikte bronnen**	▪ Ontwerp evaluatiecriteria om de kwaliteit van de bronnen te wegen aan de hand van de opgestelde doelen. ▪ Ga na welke bronnen als *kennisbank* gebruikt kunnen worden om gebruikers in hun informatiebehoefte te voorzien. ▪ Evalueer of de bronnen van voldoende kwaliteit zijn voor de doelen van de component en de LLM-toepassing als geheel. Zie KZ.4.
KZ.2	**Zoek-functionaliteit**	Er is niet één perfecte zoekstrategie. De beste zoekstrategie is context-afhankelijk en verschilt per taak waarvoor de LLM-toepassing wordt ingezet. Het bouwen van een zoekfunctionaliteit bestaat uit twee hoofdcomponenten: bronnen opknippen (chunking) en het opstellen van een zoekstrategie.

a	Opknippen/chunking	▪ Ga na in hoeverre documenten opgesplitst moeten worden. Dit hangt af van de lengte van de documenten en hoe de gekozen LLM omgaat met grote documenten.
		▪ Vuistregels voor *chunking*:
		  ▪ Indien specifieke antwoorden gewenst zijn: creëer grote stukken (chunks) aan tekst. Dit werkt doorgaans beter omdat uit een brede context een gericht antwoord samengesteld kan worden.
		  ▪ Indien het antwoord waarschijnlijk op basis van meerdere documenten wordt samengesteld, of het een brede zoekopdracht betreft, werken kleinere stukken tekst doorgaans beter.
		
		Over het algemeen geldt: een grove opdeling bevat meer context, maar kent risico op verloren details. Bij een meer fijnmazige opdeling: betere focus op details, maar de globale context kan ontbreken.
		
		Literatuur helpt bij het vinden van de juiste chunking strategie, zoals fixed length, recursive token en contextual chunking. Zie ook chunking strategieën van Chroma of contextual chunking van Anthropic.

Kennisbank en zoekstrategie (KZ)

#	Titel	Beschrijving

b	Zoekstrategie	▪ Ga na welke zoekstrategie het meest geschikt is voor de LLM-toepassing:
		  ▪ Full-text search: Zoekstrategie waarin gezocht wordt naar documenten die dezelfde woorden als zoekopdracht bevatten.
		  ▪ Semantic search: Zoekstrategie waarbij rekening wordt gehouden met de contextuele betekenis van woorden,, houd hier rekening met de keuze van het model dat gebruikt wordt om semantisch te zoeken.
		  ▪ Hybrid search: combinatie van full-text en semantic search.
		▪ Stel realistische zoekopdrachten (queries) op waarmee de kennisbank bevraagd kan worden. Zie ook **E.**
		  ▪ Het verrijken van de zoekopdracht met context kan tot betere resultaten leiden.
		
		Op HuggingFace zijn verschillende modellen te vinden die gebruikt kunnen worden voor semantisch zoeken.

KZ.3	**LLM functionaliteit**	▪ Kies een geschikt model (zie nader **LLM**) en geschikte promptingstrategie (zie nader **SP**).
		▪ Bepaal hoe de LLM interacteert met de kennisbank. Hiervoor bestaan meerdere opties, zoals (zie ook **AO**):
		  ▪ Op basis van een chatgesprek wordt eenmalig een zoekopdracht gegenereerd, waarna door middel van *Retrieval Augmented Generation (RAG)* een antwoord wordt gegenereerd.
		  ▪ Agents interacteren iteratief met de kennisbank om de relevante informatie uit de kennisbank te ontsluiten en uiteindelijk een antwoord te genereren.
		
		In protocollen, zoals het Model Context Protocol (MCP), kan vastgelegd worden hoe AI-modellen context (zoals data, tools en geheugen) veilig en consistent kunnen uitwisselen met andere systemen.

KZ.4	**Evalueren**	Evalueer de kwaliteit en prestaties van de kennisbank, zoekfunctionaliteit en LLM-functionaliteit als losse componenten en als geheel. Stel een evaluatieproces op om de opgestelde kennisbank en zoekstrategie te evalueren (zie nader **E**).

Voorbeeld: voorRecht-rechtspraak

Kennisbank en zoekstrategie

Tijdens de ontwikkeling van voorRecht-rechtspraak zijn verschillende experimenten uitgevoerd om de zoekstrategie naar jurisprudentie automatisch te evalueren.

Eén van deze experimenten wordt toegelicht. Op basis van 10.000 rechterlijke uitspraken zijn vragen gegenereerd die aan de chatbot zijn voorgelegd. Bijvoorbeeld, bij een uitspraak over geluidsoverlast door buren is door een LLM de vraag gegenereerd: “Wat kan ik doen wanneer mijn buurman ‘s nachts geluidsoverlast veroorzaakt?”. Deze vragen zijn vervolgens voorgelegd aan de LLM-applicatie. Op basis van de gevonden bronnen wordt nagegaan of het document waarop de vraag is gebaseerd in de top-k voorkomt en wordt een prestatiemetriek berekend. Aan de hand van de prestatiemetriek kan de effectiviteit van de zoekstrategie eenduidig afgezet worden tegen andere zoekstrategieën.

Op basis van een dergelijk evaluatieproces kan de beste *chunking*-strategie worden bepaald. In het geval van voorRecht-rechtspraak geldt dat doorgaans verschillende passages uit verschillende bronnen relevant zijn, waardoor het opknippen van bronnen in kleine delen beter werkt. In aanvulling op dit inzicht is ervoor gekozen om samenvattingen van de volledige uitspraak toe te voegen aan de chunk, zodat de rode draad van de uitspraak behouden blijft.

voorRecht Rechtspraak

voorRecht-rechtspraak is ontwikkeld door de Rechtspraak in samenwerking met T&T Data Consultancy en Deloitte.

Gebruikerstesten en expertreviews

voorRecht-rechtspraak en bijbehorende kennisbank en zoekstrategie zijn getest door deskundige beoordelaars, zoals rechters en juristen, om de juridische juistheid en toepasbaarheid van de output van de LLM-applicatie te toetsen.

Tijdens de gebruikerstest stelden rechters en juristen vragen aan de LLM-applicatie, waarna zij de zoekresultaten en de gegenereerde antwoorden valideerden. Deze validaties werden iteratief uitgevoerd om te testen of de zoekstrategie daadwerkelijk effectief was.

De kennisbank en zoekstrategie zijn daarna getest met eindgebruikers. Tijdens deze tests is onderzocht in hoeverre een burger de gepresenteerde informatie als nuttig ervaart.

Alle uitkomsten van de gebruikerstesten zijn opgeslagen in een *benchmark dataset*, welke gebruikt kan worden om de LLM-applicatie bij toekomstige wijzigingen (automatisch) te evalueren.

Guardrails (GR)

***Guardrails* (GR) zijn risicobeheersmaatregelen voor LLM-toepassingen. Deze maatregelen worden getroffen om de technologie veilig, betrouwbaar en zo eerlijk mogelijk in te zetten. Guardrails kunnen bijvoorbeeld de input van LLM-toepassingen controleren op de aanwezigheid van persoonsgegevens of voorkomen dat de output schadelijke informatie of irrelevante antwoorden bevat. Guardrails zijn nodig om de risico's die gepaard gaan met gebruik van een LLM-toepassing te beheersen.**

Identificeren

Doel en risico's: Bepaal doelen en risico's voor de guardrails, in lijn met de doelen voor de LLM-toepassing zoals geïdentificeerd in D. Omdat guardrails als doel hebben om risico's te beheersen worden doelen en risico's hier samen behandeld.

Ga na welke risico's beheerst kunnen worden door middel van guardrails. Denk hierbij tenminste aan:

- **Ongewenst gebruik van de LLM-toepassing:**
  - Agressief taalgebruik.
  - Noodsituaties: gesprekken die gaan over zelfbeschadiging, geweld of misbruik.
  - Gebruik voor onbedoelde taken, *jailbreaken* en *promptinjecties* hierbij proberen gebruikers het model andere dingen te laten doen dan waarvoor het bedoeld is. Dit kan direct (bijvoorbeeld middels chatinteractie) of indirect (bijvoorbeeld via een document dat meegestuurd wordt).
- **Privacy**: Het delen van persoonsgegevens door de gebruiker met de LLM-toepassing. Het gaat hier om persoonlijke informatie in de chatinteractie, niet om bewust verwerkte (meta)data.

Guardrails (GR)

Evalueren

#	Titel	Beschrijving
GR.1	**Ontwerp guardrails**	Bepaal voor ieder risico een geschikte guardrail.

a	Input guardrails	Contentfilters zijn filters die worden toegepast op de input van een gebruiker voordat deze met een LLM gedeeld wordt.
		▪ Ontwerp voor elk van de geïdentificeerde risico's een passende guardrail. Hiervoor kan gebruik worden gemaakt van bestaande libraries. Overweeg daarbij verschillende mogelijkheden:
		  ▪ Hardcoded filters
		  ▪ Specialistische predictieve modellen
		  ▪ LLM gebaseerde guardrails
		▪ Indien bestaande contentfilters niet voldoen, ontwerp dan eigen contentfilters.
		▪ Baken de LLM-toepassing en het gebruikte LLM af op basis van de *systeemprompt* (zie **SP**) zodat het enkel ingaat op gebruikersvragen die binnen de reikwijdte van de applicatie vallen, zoals vastgesteld in **D**.

b	Filteren van persoons-gegevens	Het filteren van persoonsgegevens is een specifiek type guardrail. Naar deze guardrail wordt ook wel verwezen als Personally Identifiable Information (PII) masking. ▪ Ga na welk type persoonsgeven gebruikers (on)bewust kunnen delen met de LLM-toepassing. ▪ Houd bij het ontwerp van de guardrail rekening met de vraag of de gebruikte methoden geschikt zijn voor de Nederlandse taal. ▪ Bepaal een acceptabel percentage output waarin persoonsgegevens onjuist gemaskeerd zijn door gebruikte contentfilters. **De EDPB heeft advies uitgebracht hoe de indirecte verwerking van persoonsgegevens voorkomen kan worden bij gebruik van GPAI-modellen (zie §27-42).**

Guardrails (GR)

#	Titel	Beschrijving

c	Output guardrails	▪ Bepaal wat onwenselijke output is voor de specifieke LLM-toepassing en stem contentfilters hierop af.
		▪ Bepaal een acceptabel percentage output waarin onwenselijke content is opgenomen. Ga na dat prestatiecriteria meetbaar zijn.
		▪ Vermeld de bronnen waarop de output van de LLM-toepassing is gebaseerd:
		  ▪ Vermeld op basis van welke passage uit de onderliggende bronnen de output gebaseerd is.
		  ▪ Verwerk geen informatie zonder bron in de output.
		▪ Bepaal een acceptabel percentage output waarin onjuiste informatie wordt verstrekt.
		
		Onwenselijke output is een breed begrip en heeft onder meer betrekking op stigmatiserend taalgebruik, mis- en desinformatie. Waar bij onjuiste informatie feitelijk kan worden aangetoond dat informatie onjuist is, is dat voor onwenselijke content minder vanzelfsprekend.

d	Technische waarborgen	▪ Detecteer geautomatiseerde interactie met de LLM-toepassing (botdetectie), bijvoorbeeld door middel van firewall. ▪ Sta maximaal aantal tokens toe tijdens chatinteractie ▪ Sta enkel input toe die benodigd is, dus bijvoorbeeld geen plaatjes of documenten.
GR.2	**Test guardrails**	▪ Test iedere guardrail zoals geïmplementeerd in GR.1 met behulp van een evaluatieproces. Zie E. **Voor guardrails wordt meestal gekozen voor evaluatie middels e*en benchmark dataset*.**

Pre-productie evaluatie (PPE)

Pre-productie evaluatie (PPE) richt zich op het toetsen van de LLM-toepassing voordat deze (opnieuw) naar productie gaat. Het inrichten van PPE kan ook helpen in het maken van de juiste keuzes tijdens het ontwikkelproces, zoals de keuze van een specifiek LLM (LLM) of systeemprompt (SP).

Identificeren

Doel: Bepaal doelen met betrekking tot pre-productie evaluatie in lijn met de doelen zoals geïdentificeerd in D en AO. Denk hierbij in ieder geval aan:

- **Validatie van eisen:** Vaststellen dat de LLM-toepassing voldoet aan de vooraf gestelde veiligheids-, snelheids- en nauwkeurigheidseisen.
- **Baseline:** Vastleggen van de prestaties van de eerste geaccepteerde versie, zodat toekomstige updates hiermee vergeleken kunnen worden.
- **Reproduceerbaarheid:** Reproduceerbare resultaten verhogen de betrouwbaarheid van bevindingen.

Risico's: Bepaal welke risico's van belang zijn voor pre-productie evaluatie. Denk daarbij aan risico's die ontstaan door het gebrek aan goede evaluatie of door foutieve evaluatie:

- **Generaliseerbaarheid:** Het risico dat de LLM-toepassing goed werkt op evaluatiedata, maar in de praktijk minder presteert dan verwacht.
- **Subjectiviteit:** Het risico dat nieuwe versies naar productie worden gebracht op basis van onderbuikgevoel en zonder systematische evaluatie op basis van kwantitatieve data.

Pre-productie evaluatie (PPE)

Evalueren

#	Titel	Beschrijving
PPE.1	**Stel evaluatieproces op**	▪ Stel evaluatieproces op. Zie E.
PPE.2	**CI/CD**	Test gemaakte wijzigingen aan de LLM-toepassing automatisch door middel van *CI/CD*.
a	Unit tests	▪ Schrijf specifieke opdrachten voor onafhankelijke componenten van de LLM-toepassing zodat deze individueel van elkaar getest kunnen worden.
b	Integration tests	▪ Wanneer meerdere componenten van de LLM-toepassing afhankelijk zijn van elkaar, zorg voor integratietesten om te testen of de toepassing in zijn geheel correct functioneert.
PPE.3	**Valideer met belanghebbenden**	▪ Betrek belanghebbenden bij de evaluatie van grote wijzigingen aan de LLM-toepassing. Zoek naar de juiste balans tussen domeinexpertise en gebruikservaring. ▪ Ga na of op voorhand opgestelde prestatievereisten worden behaald. Ga bij significante wijzigingen altijd na of deze invloed hebben op de prestaties van de LLM-toepassing.
a	Domeinexpert	▪ Evalueer gegenereerde content met domeinexperts om kwaliteit, juistheid en toepasbaarheid te waarborgen. Ga na of de domeinspecifieke risico's uit DO.4 voldoende zijn gemitigeerd.
b	Eindgebruiker	▪ Doe gebruikerstesten om te meten of de toepassing voor de gebruiker werkt, zie ook E.

Monitoring (M)

Zodra de LLM-toepassing in productie is, is continue Monitoring (M) vereist om de operationele prestaties, veiligheid en kwaliteit te waarborgen. Dit omvat toezicht op de operationele werking van het systeem, risicobeheersmaatregelen, en gebruikersinteracties.

Identificeren

Doel: Bepaal doelen van de monitoring in lijn met de doelen gedefinieerd in D en AO. Denk hierbij aan:

- **Validatie van eisen:** Vaststellen dat de LLM-toepassing voldoet aan de vooraf gestelde veiligheids-, snelheids- en prestatievereisten.
- **Verfijnen van vereisten:** Identificeren van verbeterpunten op basis van gebruikersinteracties.
- **Klachten en incidenten:** Gepaste omgang met klachten en incidenten.

Risico's: Bepaal risico's van (inadequate) monitoring. Denk in ieder geval aan:

- **Niet halen van prestatievereisten:** Het niet behalen van prestatievereisten gerelateerd aan privacy, kwaliteit, veiligheid, en operationele prestaties.
- **Missen van signalen:** Onvoldoende incidenten-mechanisme en/of exitstrategie.

Monitoring (M)

Evalueren

#	Titel	Beschrijving
M.1	**Monitoringsstrategie**	LLM-toepassingen vereisen constant toezicht op prestaties en gebruikersinteracties.
a	Monitoringsplan	▪ Leg vast wat (zie **M.2** – **M.5**), wanneer, door wie (zie **GO.3**), en hoe (handmatig, geautomatiseerd) monitoring plaatsvindt. Plan in ieder geval continue monitoring (dagelijks of wekelijks) en vaste periodieke evaluatiemomenten (ieder kwartaal of (half)jaar).
b	Incidentenstrategie	▪ Bepaal hoe klachten en incidenten worden opgevolgd, door wie (zie **GO.3**), en binnen welke termijn. Zorg hierbij voor een heldere koppeling met de exitstrategie (**GO.4**). ▪ Ontwikkel een helder klachten- en incidentenmechanisme, waarbij ieder incident wordt opgevolgd door een incident review, waarna mogelijk prestatievereisten en het evaluatieproces (zoals in **E** opgesteld).
c	Monitoringsplatform	▪ Ontwikkel een monitoringsplatform waarin de prestaties van de toepassing snel en overzichtelijk kunnen worden gemonitord.
M.2	**Operationele prestaties en gebruik**	▪ Monitor technische systeemprestaties, zoals snelheid (latency), capaciteit (tokengebruik) en gebruik van resources (geheugen/servers). Registreer tevens (software)fouten die optreden in het systeem.
M.3	**Monitor risicobeheersing**	▪ Monitor de guardrails (**GR**) die voor de LLM-toepassing zijn toegepast. Denk aan de frequentie van het afgaan van contentfilters, het (in)correct maskeren van persoonsgegevens en context-specifieke taken die de toepassing niet behoort uit te voeren. ▪ Bespreek toppers en floppers van de LLM-toepassing met relevante betrokkenen.
M.4	**Gebruikersinteractie met taalmodel**	▪ Monitor, indien gewenst, de kwaliteit van de toepassing op basis van analyses van chattranscripties, gebruikersfeedback en automatische evaluaties. Identificeer verbeterpunten met betrekking tot kwaliteit en gebruiksvriendelijkheid.
M.5	**Gebruikersinteractie met user interface**	▪ Monitor hoe gebruikers de gebruikersinterface (**UI**) ervaren. Analyseer impliciet gedrag: welke functies en knoppen worden gebruikt, waar haken gebruikers af (drop-off points) en wat zijn de meest voorkomende navigatiepaden. Verzamel ook expliciete feedback (bijvoorbeeld via een feedbackformulier).

Voorbeeld: voorRecht-rechtspraak

Systeemprompt

De systeemprompt van voorRecht-rechtspraak wordt naar aanleiding van monitoring regelmatig geüpdatet. Deze wijzigingen worden bijvoorbeeld doorgevoerd om gebruikers beter van informatie te voorzien. Om te voorkomen dat een nieuwe wijziging zorgt voor de terugkeer van oude problemen is een *benchmark dataset* opgebouwd van testsituaties waarin oude problemen zich voordeden. Op deze manier kan na een aanpassing van de systeemprompt worden nagegaan of oude problemen zich niet opnieuw voordoen.

Een voorbeeld is het onnodig herhalen van een vraag die de gebruiker al beantwoord heeft tijdens het gesprek. Dit kwam aan het licht bij een chat over het schilderen van het appartementencomplex door de vereniging van eigenaren (VvE). De chatbot vroeg toen of de gebruiker in een appartementencomplex woonde, terwijl de gebruiker dit al had aangegeven in een eerder bericht.

voorRecht Rechtspraak

voorRecht-rechtspraak is ontwikkeld door de Rechtspraak in samenwerking met T&T Data Consultancy en Deloitte.

Evalueren en monitoren

Bij voorRecht-rechtspraak is een dashboard ontwikkeld om het gebruik van de LLM-toepassing te monitoren. Dit dashboard is enkel toegankelijk voor medewerkers die daar bevoegd voor zijn (zie **GO.3**).

Daarnaast wordt de *LLM-as-a-judge* evaluatiemethodiek gebruikt om na te gaan of output van de LLM-applicatie juridisch advies bevat (wat niet de bedoeling is), persoonsgegevens correct worden gemaskeerd en of *hallucinaties* voorkomen in de output van de *RAG-applicatie*. Laagscorende prestatiemetrieken worden door het ontwikkelteam aandachtig bestudeerd. Afwijkende chats worden eerst door domeinexperts gelezen om na te gaan in welke context de LLM-applicatie niet optimaal presteert. Relevante casussen worden opgenomen in de *benchmark dataset* om deze situaties in toekomstige versies van de LLM-applicatie te testen.

Praktijkevaluatie (PE)

Tijdens praktijkevaluatie (PE) wordt de impact van de LLM-toepassing in de praktijk beoordeeld. Waar tijdens Pre-Productie Evaluatie (PPE) en Monitoring (M) de technische prestaties van het model worden geëvalueerd (werkt het?), gaat PE de effectiviteit van de oplossing na (helpt het?). Deze evaluatie wordt gespiegeld aan de doelen en de probleemanalyse die in fase D.1 zijn opgesteld.

Identificeren

Doel: Bepaal doelen voor praktijkevaluatie in lijn met de doelen zoals geïdentificeerd in D en AO. Denk hierbij aan:

- **Oplossend vermogen:** Valideren of het kernprobleem (bijvoorbeeld informatie-overload of lange wacht-tijden) daadwerkelijk wordt verholpen door de LLM-toepassing.
- **Proces- en keteneffecten:** Inzichtelijk maken hoe de LLM-toepassing de werkprocessen en klantervaring beïnvloedt (bijvoorbeeld kanaalverschuiving van telefonie naar digitaal of verandering in doorlooptijden).
- **Toegankelijkheid en bereik:** Verifiëren of de oplossing de beoogde doelgroep bereikt, inclusief minder-digitaalvaardigen en of het vertrouwen in de dienstverlening behouden blijft.
- **Duurzaamheid van de oplossing:** Evalueren of de baten op de lange termijn opwegen tegen de structurele beheerlasten en kosten.

Risico's: Bepaal welke risico's optreden bij in gebruikname van de LLM-toepassing in de maatschappelijke en organisatorische context. Denk bij de praktijkevaluatie tenminste aan:

- **Probleemverplaatsing:** Het risico dat de oplossing op één plek leidt tot onvoorziene problemen elders in de keten.
- **Toegankelijkheid:** Het risico dat de LLM-toepassing onbedoeld een drempel opwerpt voor specifieke doelgroepen, waardoor de ongelijkheid in dienstverlening toeneemt.
- **Eenzijdige metrieken:** Het risico dat er gestuurd wordt op oppervlakkige cijfers (zoals 'aantal gebruikers') terwijl het onderliggende probleem (de hulpvraag) niet wordt opgelost.

Praktijkevaluatie (PE)

Evalueren

#	Titel	Beschrijving
PE.1	**Korte termijn praktijkevaluatie**	Evalueer directe, meetbare impact van de LLM-toepassing in de praktijk.
a	Doelrealisatie op toepassingsniveau	▪ Ga na of de individuele informatiebehoefte of het probleem van de gebruiker daadwerkelijk opgelost wordt door gebruik van de LLM-toepassing (zie **D.1d**). ▪ Onderzoek of de toepassing leidt tot de gewenste gevolgactie (bijvoorbeeld: de burger snapt de brief, de lezer leent het boek, of de rechtzoekende onderneemt de juiste juridische stap).
b	Gebruikerservaring en vertrouwen	▪ Ga na of de LLM-toepassing de beoogde doelgroep bereikt (ook de minder-digitaalvaardigen) of dat bepaalde groepen afhaken.
c	Operationele efficiëntie	▪ Onderzoek het effect van de LLM-toepassing op de werkdruk bij medewerkers (bijvoorbeeld: minder telefoontjes, mindere of betere voorbereiding van dossiers). ▪ Onderzoek of de doorlooptijd van het proces voor de burger of professional verkort ten opzichte van de oude situatie (zie nulmeting uit **D.1b**).
PE.2	**Lange termijn praktijkevaluatie**	Evalueer de strategische en structurele impact van de LLM-toepassing over langere tijd.
a	Oplossen kernprobleem	▪ Reflecteer op de probleemanalyse in **D.1a**. Is het gekwantificeerde probleem (bijvoorbeeld: “te hoge werkdruk”, “onbegrijpbare informatie”) significant afgenomen door de inzet van de LLM-toepassing? ▪ Onderzoek of de doelgroep verschuift binnen kanalen (bijvoorbeeld van fysieke balie naar digitale loketten, waardoor experts meer tijd hebben voor complexe zaken).
b	Neveneffecten en systeemimpact	▪ Ga na of de toepassing onvoorziene gevolgen gehad heeft elders in de keten (bijvoorbeeld: leidt betere informatievoorziening tot meer bezwaarschriften omdat burgers mondiger zijn geworden?) ▪ Onderzoek of er een cultuurverandering zichtbaar is in hoe burgers of medewerkers met het onderwerp omgaan (bijvoorbeeld de verschuiving van conflict naar dialoog)?
c	Organisatorische borging	▪ Ga na of de beheerskosten (technische doorontwikkeling, API-kosten, moderatie) opwegen tegen de maatschappelijke en operationele kosten op de lange termijn. ▪ Onderzoek of de oplossing technisch en organisatorisch houdbaar is: wordt de afhankelijkheid van de leverancier/technologie te groot (bijvoorbeeld door een ‘vendor lock-in’)?

AI-Verordening (AIV)

De AI-verordening (AIV) stelt eisen aan verschillende type AI-systemen en -modellen. In de meeste gevallen stelt de AIV naast transparantievereisten geen eisen aan LLM-toepassingen voor publieke informatievoorziening. Toch is het van belang om na te gaan in welke categorie van de AIV de toepassing valt, zodat indien van toepassing aan de wettelijke vereisten voldaan kan worden.

Naam	Voorbeeld
AI-modellen voor algemene doeleinden	GPT5.2 (OpenAI), Sonnet (Anthropic), GPT-NL (TNO, SURF, NFI et al.)
AI-systemen voor algemene doeleinden	ChatGPT (OpenAI), Claude (Anthropic)
AI-systeem	De verkenner van voorRecht-rechtspraak (maakt gebruik van AI-modellen)

#	Titel	Beschrijving
AIV.1	**Transparantie**	▪ Bij directe gebruikersinteractie met een AI-systeem (zoals een chatbot): Ga na of er duidelijk wordt vermeld dat er sprake is van interactie met een AI-systeem. ▪ Bij contentgeneratie (bijvoorbeeld een samenvatting of tekst): Ga na of duidelijk vermeld is dat het door AI gegenereerde content betreft.
AIV.2	**AI-modellen voor algemene doeleinden (General-Purpose AI)**	▪ Ga na of er sprake is van substantiële wijzigingen aan het onderliggende LLM. Win in dat geval (juridisch) advies in of de organisatie onder de AIV als ontwikkelaar van het model gezien kan worden.
AIV.3	**Hoog risico AI-systeem**	▪ Ga na of de toepassing wordt toegepast voor een hoog risico doeleinde.
AIV.4	**Beoogd doel**	▪ Bij eigen ontwikkeling: Ga na dat in de gebruikersinstructies en in bijbehorende disclaimers duidelijk wordt beschreven wat het beoogde doel is van de toepassing. Zorg dat ook beschreven is waarvoor de toepassing niet bedoeld is, in het bijzonder bij voorzienbaar verkeerd gebruik voor een hoog risico toepassing. ▪ Bij inkoop van het systeem: Ga na of de toepassing gebruikt wordt in lijn met het beoogde doel, zoals door de leverancier is vastgesteld. Bij gebruik voor een hoog risico doeleinde dat niet het beoogde doel van de toepassing is, win (juridisch) advies in of de organisatie op grond van de wet als ontwikkelaar van het model gezien kan worden. **Bijvoorbeeld: "Deze toepassing mag niet worden ingezet in processen rondom de aanvraag, beoordeling, handhaving en toezicht, of terugvordering van uitkeringen, toeslagen of essentiële publieke diensten."**

AI-Verordening (AIV)

1 Transparantievereisten

Het moet altijd duidelijk zijn voor de gebruiker dat hij/zij/die interacteert met een AI-systeem. De ontwikkelaar moet de toepassing zo ontwerpen dat dit voor de gebruiker duidelijk is.

Als door middel van de LLM-toepassing teksten gegenereerd worden die vervolgens gebruikt worden voor publieke informatievoorziening (bijvoorbeeld als tekst op een informerende website) dan moet daarbij vermeld worden dat het automatisch gegenereerde tekst betreft.

Zie voor meer informatie artikel 50 van de AIV.

2 Eisen aan AI-modellen voor algemene doeleinden (GPAI modellen)

Juridisch kader

De AIV stelt eisen aan AI-modellen voor algemene doeleinden (general purpose of GPAI-modellen). Dit zijn AI-modellen die in staat zijn een breed scala aan verschillende taken uit te voeren. In de context van dit kader zijn dit de LLMs waar de toepassing gebruik van maakt (bijvoorbeeld: GPT5 van Open AI, GPT-NL, Sonnet van Anthropic etc.). De AIV stelt alleen eisen aan de leverancier (aanbieder) van het GPAI-model. In de meeste gevallen zijn er geen eisen aan de ontwikkelaar van een toepassing die gebruik maakt van een bestaand model.

Let op: in het geval van aanzienlijke wijzigingen aan het LLM waarbij de capaciteiten of de risico's van het model aanzienlijk veranderen, kan het zijn dat de ontwikkelaar die deze wijzigingen aanbrengt voor de wet als leverancier (aanbieder) van het model wordt gezien. Dit kan vergaande gevolgen hebben voor de eisen die de AIV stelt. Win bij aanpassingen aan het model (juridisch) advies in.

Beschikbare documentatie

Leveranciers van GPAI-modellen moeten documentatie beschikbaar stellen. Zo moet de volgende informatie beschikbaar zijn:

- Technische documentatie van het model, inclusief trainings- en testprocessen, en evaluatieresultaten.
- Documentatie voor andere AI-leveranciers die van plan zijn het GPAI-model te integreren in een toepassing, inclusief details over de specificaties en beperkingen.
- Een gedetailleerde samenvatting van de content en data die zijn gebruikt voor het trainen van het AI-model.
- Voor open source modellen geldt een uitzondering: voor deze modellen hoeft deze documentatie niet beschikbaar te worden gesteld.

Deze documentatie kan gebruikt worden voor de keuze van het model (zie Large LLM). Omdat dit een wettelijke verplichting is, betekent het ontbreken van bovenstaande documentatie dat de leverancier niet aan Europese wetgeving voldoet.

AI-Verordening (AIV)

3 Hoog risico toepassingen

De AIV stelt uitgebreide eisen aan een specifieke categorie AI-systemen: hoog risico AI-toepassingen. Hoog risico toepassingen zijn uitputtend benoemd in de AIV. Dit betekent dat wanneer een toepassing niet gebruikt wordt voor een van de benoemde doelen, er geen sprake is van een hoog risico AI-systeem. Er gelden dan geen hoog risico-vereisten voor de AI-toepassing, maar mogelijk wel transparantieverplichtingen. Zie 1 op pagina 50.

Een LLM-toepassing die ondersteunt bij publieke informatievoorziening zal in de meeste gevallen niet bedoeld zijn voor inzet in een hoog risico toepassing. Indien dit wel het geval is, moet aan de eisen voor hoog risico AI systemen worden voldaan.

De hoog risico toepassingsgebieden zijn:

- Biometrie
- Kritieke infrastructuur
- Onderwijs en beroepsopleiding
- Werkgelegenheid, personeelsbeheer en toegang tot zelfstandige arbeid
- Toegang tot en gebruik van essentiële overheids-uitkeringen en –diensten
- Beoordelen van kredietwaardigheid
- Verzekeringen
- Rechtshandhaving
- Migratie-, asiel- en grenstoezichtsbeheer
- Rechtsbedeling en democratische processen

Wanneer de toepassing bedoeld is voor inzet in een van deze domeinen is het van belang na te gaan of er sprake is van een van de specifieke use-cases en eventueel uitzonderingen. Zie hiervoor Bijlage III en artikel 6(3) van de AI-verordening.

Ga ook na of er voorzienbaar verkeerd gebruik is, waarbij een gebruiker de toepassing inzet voor een hoog risico doeleinde, terwijl dit niet de bedoeling is.

Let op: de beschrijving van essentiële overheidsdiensten in de AIV is als volgt:

“AI-systemen die bedoeld zijn om door of namens overheidsinstanties te worden gebruikt om te beoordelen of natuurlijke personen in aanmerking komen voor essentiële overheidsuitkeringen en -diensten, waaronder gezondheidsdiensten, of om dergelijke uitkeringen en diensten te verlenen, te beperken, in te trekken of terug te vorderen.”

Als LLM-toepassingen door overheidsinstellingen niet voor een van deze doeleinden worden toegepast, kwalificeert de toepassing op deze grond niet als hoog risico toepassing.

Begrippenlijst

Beschrijving van gebruikte begrippen in het validatiekader. Begrippen uit deze lijst zijn in het kader *cursief* gemarkeerd.

Begrip	Beschrijving
A/B-testing	A/B-testing is een methode waarbij twee versies van de LLM-toepassing worden vergeleken om te bepalen welke beter presteert.
Benchmark dataset	Een benchmark dataset is een verzameling testsituaties, relevant voor de specifieke context waarbinnen de LLM-toepassing wordt gebruikt, die wordt gebruikt om de prestaties van de toepassing te evalueren en te vergelijken.
Chunking strategie	Een chunking strategie is een techniek waarbij grote hoeveelheden tekst worden opgedeeld in kleinere, beter verwerkbare stukken voor efficiëntere verwerking door taalmodellen.
CI/CD	Een werkwijze waarbij software automatisch wordt getest, zodat software sneller en betrouwbaarder in gebruik kan worden genomen.
Classificatie	Het proces waarbij gegevens door een algoritme automatisch worden ingedeeld in vooraf gedefinieerde categorieën op basis van hun kenmerken.
Contentfilter	Een contentfilter is software die de input (wat de gebruiker met de LLM-applicatie deelt) en de output (wat het taalmodel antwoordt) filtert om schadelijke, onveilige of ongewenste inhoud te detecteren en te blokkeren.
Fine tuning	Fine tuning is het proces waarbij een bestaand taalmodel voor een specifieke LLM-toepassing verder wordt getraind op specifieke data om prestaties te verbeteren.
Front-end	Een front-end is het deel van de LLM-toepassing dat gebruikers zien en waarmee ze interacteren.
Jailbreaken	Jailbreaken is het omzeilen van de ingebouwde veiligheidsmaatregelen en contentfilters van een LLM-toepassing om content te genereren die normaal geblokkeerd zouden worden.
Ground truth label	Het juiste, door een expert vastgestelde, label dat wordt gebruikt als referentie om een algoritmische toepassing te trainen of te evalueren.
Guardrails	Risicobeheersmaatregelen die het gedrag van de LLM-toepassing sturen om ongewenste of onveilige output te voorkomen.
Hallucinatie	Een fout waarbij een AI-systeem overtuigend lijkende maar onjuiste of verzonnen informatie genereert.
Kennisbank	Een kennisbank is een gestructureerde verzameling informatie die het taalmodel kan raadplegen om nauwkeurige en contextspecifieke antwoorden te geven.

Begrippenlijst

Begrip	Beschrijving
LLM-as-a-judge	LLM-as-a-judge is een aanpak waarbij een taalmodel wordt ingezet om de kwaliteit of juistheid van de output van een LLM-toepassing te beoordelen.
Persoonsgegevens	Persoonsgegevens zijn alle gegevens gerelateerd aan een specifieke identificeerbare persoon. Dit is een zeer brede categorie, hieronder vallen ook indirecte gegevens zoals de metadata van een gebruikersinteractie (ook wanneer geen naam verwerkt wordt). Of een persoon geïdentificeerd kan worden hangt af van de combinatie van beschikbare gegevens en de middelen die iemand ter beschikking heeft om een persoon te kunnen identificeren.
Prestatiemetriek	Een numerieke maatstaf die aangeeft hoe goed een algoritmisch systeem een taak uitvoert, zoals nauwkeurigheid (accuracy), precision en recall.
Productie	Het moment waarop de LLM-toepassing daadwerkelijk wordt ingezet en gebruikt in een echte, live omgeving door eindgebruikers.
Prompt -engineering -injecties gebruikers- systeem-	Prompt engineering is het proces van het zorgvuldig ontwerpen en formuleren van prompts om de LLM-toepassing zo effectief en nauwkeurig mogelijk de gewenste output te laten genereren. Promptinjecties zijn aanvallen waarbij kwaadwillende instructies in een prompt worden verstopt om een LLM-toepassing te misleiden of ongewenst gedrag te veroorzaken. Een gebruikersprompt is de instructie of vraag die een gebruiker aan een LLM(-toepassing) geeft om een specifieke reactie te geven of taak uit te voeren. Dit kan ook een chat of query zijn. Een systeemprompt is de invoer of opdracht die aan een taalmodel wordt gegeven om een bepaald antwoord, gedrag of output te verkrijgen.
RAG-applicatie	Een Retrieval Augmented Generation (RAG)-applicatie haalt informatie op uit externe bronnen om een taalmodel te voorzien van context, zodat het accuratere en relevantere antwoorden kan genereren.
Red-teaming	Het systematisch testen van een LLM-applicatie door de rol van een aanvaller aan te nemen om kwetsbaarheden en zwakke plekken te ontdekken.
Voorspelling	Een door een model gegenereerde inschatting van een toekomstige of onbekende uitkomst op basis van beschikbare gegevens.
WCAG	De Web Content Accessibility Guidelines (WCAG) zijn internationale richtlijnen die beschrijven hoe je websites en digitale content toegankelijk maakt voor iedereen, inclusief mensen met een beperking.

Totstandkoming van het validatiekader ‘Verantwoorde inzet Large Language Models (LLMs) voor publieke informatievoorziening’

Dit validatiekader is tot stand gekomen op basis van samenwerking tussen onderstaande partijen in het kader van de call ‘Responsible AI in het praktijk’ van het SIDN Fonds en Topsector ICT.

Algorithm Audit

Stichting Algorithm Audit is een Europees kennisplatform voor verantwoorde AI. Het platform brengt data scientists, juristen en ethici bijeen om oplossingen te formuleren voor waardegedreven vraagstukken die centraal staan bij de inzet van algoritmische systemen.

Deloitte.

Deloitte is een wereldwijd advies- en accountantsbedrijf dat diensten aanbiedt op het gebied van audit, consulting, financiële advisering, risicobeheer en belastingen.

de Rechtspraak

De Rechtspraak is het geheel van rechtbanken en rechters die in Nederland onafhankelijk recht spreken en geschillen beslechten volgens de wet.

T&T Data Consultancy

T&T Data Consultancy is een Nederlands data- en AI-adviesbureau dat organisaties helpt met het opzetten van slimme data-oplossingen, de implementatie van AI en het betrekken van mensen in die transitie.

TU/e EINDHOVEN UNIVERSITY OF TECHNOLOGY

De Technische Universiteit Eindhoven (TU/e) is een Nederlandse universiteit die zich richt op onderzoek en onderwijs in technologie, engineering en innovatie.

Dit project is mede mogelijk gemaakt door:

SIDNfonds

TOPSECTOR ICT dutch digital