Acceptatie criteria

Learning Lion Kamervragen

Het learning lion project zal vanaf 2 kanten beoordeeld worden; enerzijds is er een evaluatie nodig van de technische aspecten die prestaties meet zoals de **nauwkeurigheid van het model, verwerkingstijd** die het systeem nodig heeft om verzoeken te verwerken, en **consistentie van de output**. Dit heeft als doel om te meten hoe goed het AI-model relevante en correcte documenten ophaalt of accurate antwoorden genereert.

Anderzijds zijn er echter ook kwalitatieve aspecten waar een succesvolle versie van het systeem aan zou moeten voldoen. De criteria waaraan deze aspecten moeten voldoen zijn uitgezet in dit document, en ingedeeld in de categorieën ‘compliance’, ‘Inhoudelijke nauwkeurigheid’, en ‘gebruiksvriendelijkheid’. Deze criteria vanuit SSC-ICT zijn verder aangevuld met de acceptatiecriteria vanuit JenV, de opdrachtgever van de use case ‘Kamervragen’.

Compliance

Naleving van de wettelijke vereisten: Het model voldoet aan alle wettelijke vereisten, zoals de AVG en de AI act, en vertrouwelijke gegevens worden beschermd tijdens het gebruik van de tool. Ook moeten er de juiste impact assessments voor uitgevoerd zijn zoals een DPIA en een AIIA.

Transparantie van de modellen: De ontwikkeling van het project gebeurt in het openbaar met volledige transparantie en inzage in zowel het RAG model zelf als de gekozen modellen binnen dit systeem. Verder is het duidelijk voor gebruikers hoe het systeem werkt en hoe antwoorden tot stand komen.

Ethische normen: De AI wordt ingezet op een ethisch verantwoorde manier. Dit houdt onder andere in dat de bijbehorende risico’s in kaart zijn gebracht en hierop bewust gemaakt is. Een voorbeeld is dat het model actief ontmoedigt de gegenereerde antwoorden te kopiëren zonder deze te controleren op feitelijkheid, en dat er een beargumenteerde afweging gemaakt wordt die de risico’s van pre-trained taalmodellen overzichtelijk maakt. Belangrijk is dat de bijbehorende impact assessments zijn uitgevoerd zoals de IAMA.

Veiligheidsmaatregelen: Er zijn maatregelen genomen om het product te beveiligen tegen kwaadwillige partijen.

Werkt volledig lokaal: Het gehele systeem werkt volledig on-premise zonder enige netwerkverbinding naar externe servers voor data processing. Zo blijft alle data lokaal opgeslagen wat risico’s tot datalekken minimaliseert.

Inhoudelijke nauwkeurigheid

(Subjectieve) kwaliteit van de output: De antwoorden en documenten die het systeem oplevert worden als bruikbaar en waardevol ervaren door de eindgebruikers. (Subjectieve perceptie van gebruikers, anders dus dan de technische evaluatie van de kwaliteit van de output).

Geen hallucinaties: Het model mag geen onjuistheden produceren. Alle gegenereerde antwoorden moeten een feitelijke basis hebben in de brondata, en wanneer dit niet het geval is moet het model aangeven dat er geen informatie over te vinden is.

Gebruiksvriendelijkheid:

Eenvoud: De eindgebruikers van de tool moeten omgang met de interface en de verschillende functies van de tool als eenvoudig ervaren.

Tijdsbesparend: Het model moet ervoor zorgen dat kamervragen gemiddeld sneller beantwoord kunnen worden.

Acceptatiecriteria vanuit JenV

(Subjectieve) kwaliteit van de output

Consistentie: Dezelfde prompt krijgt een zeer vergelijkbaar, liefst identiek antwoord. Antwoorden voldoen altijd op dezelfde wijze aan de andere criteria (bijv. bronvermelding heeft vast format).

Stijl/toon (je/u): Afhankelijk van de vraag wordt er (deels) beantwoord vanuit de eerste persoon. Stijl is in de vorm van tekst die ook gesproken zou kunnen worden als een antwoord in een interview. De stijl is respectvol en gebruikt geen jargon waar dit niet nodig is. Er worden zo min mogelijk afkortingen gebruikt.

Feitelijke juistheid: Antwoorden bevatten enkel feiten, standpunten en beleid uit de brondata. Hierbij wordt het meest recente standpunt/feiten/beleid gebruikt als dit in het verleden gewijzigd is. Antwoorden bieden geen ruimte voor interpretatie anders dan de brondata.

Relevantie: Gegenereerde antwoorden gaan in op (delen van) de vraag. Gegenereerde antwoorden bevatten de gevraagde informatie of bevatten (verwijzingen naar) eerdere standpunten en antwoorden. Als in de brondata geen (deel) van antwoorden te vinden is, wordt geen antwoord gegenereerd.

Snelheid genereren antwoord: Liefst realtime. Maar ten minste moeten antwoorden in orde van minuten worden gegenereerd, zodat een medewerker parlementair een vraag kan stellen

en vrijwel direct door kan gaan met het doorsturen van het antwoord naar de relevante dossierhouder.

Lengte: Antwoorden hebben (per deelvraag) een omvang van 150-250 woorden. Ik verwacht dat we geen voorkeurslengte hebben en dat lengte ondergeschikt is aan de andere criteria. Wel moet een antwoord redelijk kort en bondig geformuleerd worden.

Bronvermelding: De bronvermelding heeft een duidelijke opmaak en verwijst duidelijk welke delen van een genereerd antwoord uit welke (delen van) een bron komen. De bronvermelding is een link die de ambtenaar direct naar het gepubliceerde stuk stuurt en verwijst naar de specifieke vraag/antwoord combinatie in het stuk. Kan in het antwoord een standaardzin komen met iets als ‘bij het vinden van informatie voor dit antwoord is gebruik gemaakt van AI’? Moeten we t.z.t. bekijken welke zin dat moet worden.

Volledigheid: Een gegenereerd antwoord gaat in alle aspecten en deelvragen van een Kamervraag en het gegenereerde antwoord gebruikt de volledige onderbouwing van een bron-antwoord.