Challenge Brief

Project:

Challenge Owner:

Ministerie van BZK (intern gebruik)

Hoe verbeteren we de kwaliteit van digitale assistent-output in real time?

Achtergrond:

Binnen het ministerie van BZK wordt gewerkt met een interne digitale assistent die medewerkers ondersteunt bij het vinden, interpreteren en toepassen van informatie. De assistent functioneert technisch, maar in de praktijk blijkt dat **de kwaliteit van de output sterk bepalend is voor vertrouwen en adoptie**.

Niet alleen *of* een antwoord correct is, maar ook:

- of het relevant is voor de vraag
- of de toon passend blijft
- of het antwoord voldoet aan beleidsmatige en ethische kaders

maakt het verschil tussen een bruikbaar hulpmiddel en een risico.

Probleemstelling:

Veel huidige AI-assistenten genereren output in één stap. Eventuele kwaliteitseisen worden:

- pas achteraf gesignaleerd
- of helemaal niet actief gecorrigeerd

Dit leidt tot situaties waarin antwoorden:

- onvoldoende aansluiten bij de context of intentie van de gebruiker
- een onwenselijke toon of sentiment aannemen
- niet voldoen aan specifieke eisen (bijv. geen politiek advies, kritisch reflecterend, neutraal geformuleerd)

Alleen het *signaleren* van deze problemen is beperkt waardevol voor de gebruiker. **De echte impact zit in het verbeteren van de output zelf.**

Challenge: **Hoe kunnen we overheidsmedewerkers helpen om digitale assistenten betrouwbaar en verantwoord te gebruiken, door de kwaliteit van AI-output expliciet te definiëren, meten, vergelijken en verbeteren, in plaats van alleen achteraf kwaliteitsproblemen te signaleren?**

Wat onderzoeken we in deze challenge?

Teams werken aan één of meerdere van de volgende kernvragen:

- **Toegevoegde waarde**
  - Hoe maak je concreet welke waarde een digitale assistent levert ten opzichte van:

        - geen assistent?
        - een simpele zoekfunctie?
    - Welke kwaliteitsdimensies zijn daarbij relevant (bijv. relevantie, volledigheid, toon, bruikbaarheid)?
- **Meten en valideren van kwaliteit**
    - Hoe kun je de kwaliteit van AI-output objectief meten?
    - Welke criteria zijn geschikt voor overheidscontexten?
    - Hoe valideer je dat een antwoord “goed genoeg” is?
- **Gouden antwoord-set**
    - Hoe kom je tot een referentieset van “goede antwoorden”?
    - Kun je werken met:
        - experts?
        - voorbeelden?
        - consensus tussen meerdere modellen?
    - Hoe gebruik je zo’n set voor kwaliteitsvergelijking of regressietesten?
- **Tooling en evaluatiecriteria**
    - Welke tooling gebruik je om kwaliteit te meten?
    - Welke evaluatiecriteria zijn geschikt voor:
        - relevantie
        - consistentie
        - toon & sentiment
        - naleving van beleidskaders?
- **Testen en monitoren voor gebruikers**
    - Hoe maak je kwaliteitsmetingen begrijpelijk voor niet-technische gebruikers?
    - Hoe presenteer je scores, signalen of waarschuwingen zonder de gebruiker te overweldigen?
- **LLM’s als beoordelaar**
    - Hoe kun je LLM’s inzetten om andere LLM-output te beoordelen?
    - Wanneer is dit betrouwbaar?
    - Hoe maak je deze beoordeling uitlegbaar en controleerbaar?

💡 *Teams hoeven niet alles op te lossen — focus en onderbouwde keuzes zijn belangrijker dan volledigheid.*

Sustainable development goals (SDG’s):

Criteria

- **Verplicht**
  - Duidelijke definitie van kwaliteit
  - Minstens één meetbaar kwaliteitscriterium
  - Inzicht waar en waarom de assistent ingrijpt
  - Begrijpelijk voor niet-technische gebruikers
- **Bonus**
  - Gouden antwoord-set
  - Human-in-the-loop
  - Transparantie over onzekerheid
  - Governance/beleidskaders in de flow

Data:

Voor deze challenge is een open source starter repository beschikbaar met voorbeelden en structuur om **AI-outputkwaliteit te meten en verbeteren**.

🔗 **https://github.com/mathisvfr/digital-assistant-quality-starter**

De repo biedt een basis om:

- kwaliteitscriteria te definiëren
- output te evalueren en vergelijken
- interventiestappen te ontwerpen

Focus ligt op **kwaliteit**, niet op het bouwen van een volledige assistent.

Resources en inspiratie:

- **OpenEvals: Readymade evaluators voor LLM-apps** (Github) open source toolkit om evaluaties te schrijven voor LLM-toepassingen, met support voor *LLM as a Judge*-patronen en maatwerk criteria zoals conciseness of relevancy.
- **cerai-iitm/AIEvaluationTool** framework voor het testen en benchmarken van conversational AI-systemen; helpt bij het automatiseren van evaluaties rond nauwkeurigheid, betrouwbaarheid en gebruikerservaring.
- **confident-ai/deepeval** open source LLM evaluatie-framework met metrics zoals relevantie, hallucinaties en taak-voltooiing – bruikbaar voor unit-testing van output kwaliteit.
- **SigmaEval/SigmaEval** python-framework voor end-to-end testing van conversational AI-apps; ideaal om *objectieve kwaliteitsscores* te definiëren en regressietesten uit te voeren.
- **Vvkmnn/awesome-ai-eval** gecureerde lijst van tools, methodes en benchmarks rondom AI-evaluatie, met links naar frameworks, datasets en metrics.
- **alopatenko/LLMEvaluation** guide en overzicht van LLM-evaluatiemethoden; helpt teams om te bepalen welke evaluatievormen (bijvoorbeeld relevantie, consistentie, veiligheid) passen bij hun use case.

Wat je oplevert:

- Een prototype of demo van een kwaliteitsbewuste assistent-flow
- Voorbeelden vóór en ná kwaliteitsinterventies
- Een korte toelichting op ontwerpkeuzes en aannames
- Een presentatie over het prototype of demo en jullie visie

DISCLAIMER: …

Casushouder: Monique Neijman, DigiCampus & Jochem Huijps, ADC Consulting

Hackathon Toolkit

Tools die jij kunt gebruiken tijdens de Hackathon. Vraag gerust om hulp aan de crew.

Brainstorm & samenwerken

- Miro – digitale whiteboard voor brainstorms & flowcharts
  👉 https://miro.com

- Jamboard (alternatief) – eenvoudig whiteboard van Google
  👉 https://jamboard.google.com

Design & prototyping

- Figma – ontwerp je app, interface of dashboard
  👉 https://figma.com

- Canva – voor posters, logo's, visuals, presentaties
  👉 https://canva.com

- Draw.io – flowcharts, systemen en logica tekenen
  👉 https://app.diagrams.net

Technologie & code

- GreenPT API *(via hackathon partner in de Hackerpack)*
  👉 Privacy-first AI API, volledig EU-gehost en op open-source modellen.

- GitHub – versiebeheer en samenwerking voor open source code
  👉 https://github.com

- Jupyter Notebook / JupyterLite
  👉 https://jupyter.org

Presentatie & pitch

- Google Slides / Canva / PowerPoint – bouw je pitch

💡 **Tip:**

Werk samen in één tool per team (bijvoorbeeld alles in Figma of Canva), en zorg dat jullie bestand makkelijk te delen is voor de pitch. Hetzelfde geldt voor je code, zorg dat je repository in te zien is door juryleden en creëer een duidelijke README.