OPEN STATE FOUNDATION
open data, transparante samenleving

Belangrijke mededeling: Om de periode voor het indienen van een voorstel te verlengen en in verband met de kerstvakantie is besloten om de deadline te verschuiven naar dinsdag 2 januari 2024.

Aanvullende informatie ter beantwoording van de binnengekomen vragen:

1. **Wat is de relatie tussen het genereren van 'debat' en 'prompt' (de gebruikersinput**)?
   Het genereren van debat willen wij op twee manieren doen.
   Door de twee modellen met verschillende rollen met elkaar te laten debatteren. Daarvoor geven we een input 'prompt' met een situatie, onderwerp en een rol; Bijvoorbeeld: Voer een plenair debat over het tekort aan opvangplaatsen in Ter Apel, als Kamerlid van partij X of als Minister van Ministerie B.
   De tweede wijze is door één model beide (of meerdere) kanten van het debat te genereren. Het prompt is dan gelijk aan de input van hiervoor, behalve dat er twee of meer rollen in beschreven worden.
   Het onderwerp kan een kamerbrief zijn, die niet in de dataset zit, of een nieuwskop.
   De situatie is de setting van het debat, is dat in de plenaire zaal of een commissiedebat.
   We willen hierbij dus ook individuele Kamerleden en bewindspersonen simuleren, de persoon bepaalt, naast de politieke kleur en rol tenslotte de toon en inhoud van het debat.

2. **Wat is de maatstaf waar we de gegenereerde debatten mee beoordelen?**
   De debatten die we genereren zullen we op drie manieren beoordelen.

   Allereerst vergelijken we de debatten met debatten die daadwerkelijk gevoerd zijn. Dit doen we aan de hand van een referentiedataset die bestaat uit data die de LLM niet in zijn trainingsdata heeft gehad. In casu is dat data die recenter is. Daarmee hebben we ook een richtsnoer voor de onderwerpen en debat paren.

   Ten tweede beoordelen we de debatten op aanwezigheid van 'ongepaste uitspraken'. Dit criterium is gebaseerd op juridische grenzen, dus bijvoorbeeld discriminerende uitspraken, racisme, bedreigingen etc.

achten wij ‘ongepast’. De aanwezigheid hiervan zal verder onderzocht moeten worden om te vinden welke ‘vangrails’ we moeten inbouwen om dit te voorkomen bij eventuele doorontwikkeling.

Tot slot beoordelen we de debatten op kwaliteit. Dat betekent dat we doormiddel van sampling een aantal debatten uit de dataset halen om een goede steekproef te doen. Die debatten zullen we beoordelen op de aanwezigheid van gebruikte debattechnieken en conclusies.

3. **Hosting en kosten**
   Het budget van €25.000,- is voor de finetuning inclusief benodigde hostingkosten. De hosting dient op eigen servers, die reeds in bezit zijn, of op gehuurde servers plaats te vinden. Het project heeft een beperkte looptijd, aanschaf van servers is dan ook niet passend. De GPU-tijd die voor de finetuning nodig is dient ook gehuurd te worden.
   Beide kosten zien we dan ook graag als onderdeel van de offerte in de begroting.
   Wat betreft de GPU-tijd die nodig is tijdens de hosting, zien wij graag een schatting van de beschikbare tijd die binnen het budget past. On demand capaciteit heeft de voorkeur omdat we met een klein aantal mensen de modellen zullen testen. De offerte hoeft nog niet te voorzien een open (publieke) test met veel capaciteit en (nu nog) onbekend aantal te verwachten gebruikers.
   Wij hebben geen directe toegang nodig tot de server, maar de dienst moet wel beschikbaar zijn via een API, zodat we vanuit onze frontend ermee kunnen communiceren.

4. **De modellen, Bloom en Alpaca**
   De keuze voor deze twee modellen staat vast om een aantal redenen. Hoewel deze keuze ook beperkingen met zich meebrengt, bijvoorbeeld in het aantal tokens, levert het ook het voordeel op van kortere debatten. Van de referentieset kunnen wij grote debatten altijd opdelen in stukken tussen twee sprekers, hiermee maken we de beoordelingen zoals we die gaan uitvoeren mogelik (zie punt 2).
   Dan is er binnen de modellen nog een keuze te maken uit de verschillende formaten. Voor het antwoord daarop vragen wij jullie hulp. Geef in jouw offerte aan beargumenteert de keuze aan voor een specifieke versie van het model in relatie tot de taak. Het doel is een zo goed mogelijk debat te genereren.

5. **De data(set)**
   Voor de data om het model te finetunen is bewust een ruime set gekozen, met een kleine beperking in tijd. We zijn ons bewust van het feit dat beide modellen niet getraind zijn op Nederlands, daarom is de finetuning met een ruime dataset in het Nederlands essentieel volgens ons. Verder is voor de vergelijkbaarheid van beide modellen nodig dat ze beide gelijke finetuning krijgen, beide op de volledige dataset.

NB We zijn attent gemaakt op het ontbreken van een deel van de documenten in de dataset.
Wij zullen zorgen dat de data die gebruikt moet worden voor het finetunen compleet beschikbaar komt. Daarbij zijn de JSON-bestanden leidend voor het aantal documenten.

Eventuele vragen naar aanleiding hiervan kunnen gesteld worden via contact@openstate.eu