**RAG testen met inkoopmedewerkers I&W**

Wij hebben vanuit zowel het inkoopteam van SSC-ICT als het inkoop-uitvoeringscentrum van I&W de vraag gekregen of we vragen van leveranciers die in een NVI gesteld worden kunnen beantwoorden op basis van eerdere Nota van Inlichtingen en andere documentatie.

Hiervoor heb ik vanuit het inkoopteam van SSC-ICT nota van inlichtingen gekregen, en vanuit het team vanuit I&W een breed scala aan documenten over de beleidsadvies en ingenieursdiensten. De opdracht hierbij was om deze lokaal te draaien.

In dit document beschrijf ik de verschillende experimenten en ervaringen die wij hebben gehad tijdens het testen van deze casus. We hebben hier ook steeds verschillende instellingen getest, in de hoop dat met andere instellingen de resultaten beter zouden worden.

Dit deden we door telkens dertien vragen tegelijk aan het model te stellen, dan iets in de instellingen aan te passen en dit nogmaals te doen, om vervolgens de resultaten met behulp van de inkoopmedewerkers van I&W te vergelijken.

De eerste experimenten waren mislukt omdat de vragen niet goed opgeknipt waren, waardoor halve zinnen als losse vragen geduid werden waar uiteraard geen goed antwoord op kon komen.

Vervolgens zijn er een aantal experimenten geweest waarbij sprake was van zeer vreemde antwoorden, vaak in het Engels en de opdracht en context duidelijk niet goed begrepen werd. Toen heb ik na wat verschillende dingen te testen de volgende prompt toegevoegd:

**Prompt #1**

### OBJECTIVE ###

Je bent een assistent voor inkoopmedewerkers van de rijksoverheid. Jouw taak is om vragen van leveranciers te beantwoorden in het Nederlands. Zorg ervoor dat je alleen antwoord geeft op basis van de beschikbare context en dat je daar ook naar verwijst in je antwoord.

### AUDIENCE ###

De doelgroep van jouw antwoorden zijn leveranciers van de aanbestede dienst. Geef alle relevante informatie uit de context, antwoord in het Nederlands leg in maximaal 100 woorden zoveel mogelijk uit.

### GUARDRAILS ###

Indien de context onvoldoende informatie bevat om de vraag te beantwoorden, verzin dan geen informatie maar geef aan dat er onvoldoende informatie beschikbaar is.

### INSTRUCTIONS ###

- Beantwoord de vraag altijd in het Nederlands, zelfs als de context in het Engels is gesteld.

- Vermijd het herhalen van de vraag in het antwoord en het herhalen van de instructies. Voer de instructies uit en geef een concreet antwoord op de gestelde vraag.

- Geef een stapsgewijze redenering bij het beantwoorden van de vraag en refereer naar specifieke zinnen uit de context die hebben bijgedragen aan het antwoord.

- Houd je antwoord nauw verbonden met de context en vermijd het toevoegen van informatie die niet expliciet in de context wordt vermeld.

### QUESTION ###

Dit zorgde ervoor dat ik antwoorden in het Nederlands kreeg, en de opmaak beter was. Het functioneerde acceptabel genoeg om eerste testjes te draaien.

Ik ben begonnen met een 4-tal variaties te testen, waarbij ik iedere keer dezelfde 14 vragen stelde. Om deze variaties toe te lichten moeten we eerst bespreken wat er zoal te kiezen valt.

Ten eerste is er een **large language model** (of taalmodel), dit is het model dat uiteindelijk de vraag beantwoordt. We hebben eerder een [analyse](https://docs.google.com/document/u/0/d/1U_0FFPa-3OQnJ1r6bwaljcNCusrRCz4o/edit) gemaakt van welke open-source taalmodellen er zijn, en Mixtral 8x7B lijkt de beste te zijn maar die was ook te zwaar om te draaien. Daarom hebben we gekozen om te experimenteren met 2 gefinetunede versies van Mistral 7b, namelijk Mistral-open orca en Zephyr. Gezien de beschikbare modellen zich snel ontwikkelen is dit niet het belangrijkste gedeelte om perfect te hebben, maar een beter model betekent uiteraard een betere indicatie voor wat het taalmodel zelf wel en niet kan.

Daarnaast is er ook een **embeddingsmodel** nodig, dit is het model dat stukken tekst omzet naar een numerieke weergave zodat de meest relevante stukken tekst numerieke weergaven hebben die op elkaar lijken. Dit is dus relevant voor het terugvinden van de juiste stukken tekst. Er zijn slechts een paar embeddingsmodelen die geschikt zijn voor RAG applicaties en specifiek op het Nederlands getraind zijn (we hebben al eerder waargenomen dat embeddingsmodellen die niet op het Nederlands getraind waren heel slecht presteerde) dat zijn de volgende:

****

De 2 GroNLP modellen zijn praktisch hetzelfde, en op het eerste gezicht leken de prestaties van het model van het forensisch instituut en todai zeer ondermaats. We hebben daarom gekozen om eerst verder te onderzoeken met jegormeister/robbert-v2-dutch-mqa-finetuned en textgain-allnli-GroNLP-bert-base-dutch-cased, vanaf nu afgekort als respectievelijk jegormeister en GroNLP.

We hebben bovendien op 2 verschillende manieren gezocht naar de meest relevante stukjes tekst. Enerzijds door ieder beschikbare stukje tekst met de vraag te vergelijken en te kijken welke stukken tekst er het meest op leken, maar ook door een techniek toe te passen die HyDE (hypothetical document embedding) heet[[1]](#footnote-0) waarbij we een hypothetisch antwoord op de vraag bedenken en gaan zoeken naar stukken tekst die daarop lijken. We noemen dit respectievelijk zoeken op de vraag en het antwoord, waarbij er vaak een mix is gebruikt die ik antwoord en vraag noem.

Tot slot kan je kiezen hoe groot je de stukken wil hebben waarin je de tekst opdeelt, dit heet de chunk-size en gaat uit van het aantal karakters. Die is voor nu steeds op 512 gekozen, dat is een redelijk gebruikelijke omvang en het leek me geen goed idee op dit gebied direct ook variatie toe te passen. Het is echter mogelijk dat een andere chunk-size optimaal is.

Dit alles uitgelegd hebbende waren de gebruikte variaties bij onze eerste test:

Taalmodel: mistral-openorca

Embeddingmodel: GroNLP

Zoekfunctie: antwoord en vraag

Chunk-size: 512

Taalmodel: mistral-openorca

Embeddingmodel: jegormeister

Zoekfunctie: vraag

chunk-size: 512

Taalmodel: Zephyr

Embeddingmodel: GroNLP

Zoekfunctie: vraag

chunk-size: 512

Taalmodel: Zephyr

Embeddingmodel: jegormeister

zoekfunctie: antwoord en vraag

chunk-size: 512

Helaas is bij de variaties die alleen op de vraag zochten iets fout gegaan en kregen we heel vaak dezelfde bronnen terug bij een breed palet aan vragen. Waarom dit precies fout gegaan is weten we niet, maar dat het niet klopt is evident. Een theorie die ik heb is dat er in de code daarvoor iets verkeerd stond waardoor die niet alleen op de vraag zelf maar ook de volledige uitleg, die veel langer is dan de specifieke vraag. Door het overlap in de uitleg kwamen misschien steeds dezelfde bronnen boven drijven.

Van de overgebleven tests is het volgende te zeggen:

* Bij beide zijn de juiste stukken tekst gevonden om de eerste paar vragen te beantwoorden, de antwoorden zelf waren echter tegenvallend maar dit ligt aan het taalmodel met een beter taalmodel komt dit goed
* Soms werden de vragen gevonden maar waren de chunks te klein om ook het antwoord mee te geven, wat suboptimaal was
* Ik zag geen significant verschil in de prestaties van Jegormeister en GroNLP
* Er waren een heel aantal hallucinaties, eerder kreeg ik heel vaak te horen dat het juiste antwoord niet in de bronnen te vinden was dus had hem nu iets minder nadrukkelijk meegegeven dat die moet zeggen dat die het niet wist als de informatie ontbrak, en dat was te merken (zelfs al stond dit alsnog ergens in de system prompt)
* Er was vaker niet dan wel de relevante tekst gevonden, maar af en toe ging het goed
* Zephyr lijkt wat meer te hallucineren, terwijl mistral-openorca meer terug lijkt te leggen bij de aanbestedende dienst
* Beide hebben hun taak niet altijd begrepen, er wordt met enige regelmaat de vraag herhaald of beantwoord vanuit het perspectief van de leverancier ipv de inkoopmedewerkers. Dit vraagt om verbeteringen in de systeem prompt.

Na overleg zijn besloten de volgende aanpassingen te maken:

* Proberen te kijken of we met een betere systeemprompt betere resultaten kunnen behalen:
  + Wat dikwijls fout ging was rolverwarring, spreken vanuit de leverancier ipv, hier expliciet op ingaan in de nieuwe prompt
* Alles samen in 1 excel zetten, ipv meerdere verschillende
* Mistral-open orca gaf fijnere antwoorden in de optiek van de inkoopmedewerkers

Op basis van deze observaties ging de volgende vergelijking zich richten op de systeemprompt, ik ging veel experimenteren met verschillende prompts en had afgesproken resultaten bij een aantal verschillende prompts samen in 1 excel document te zetten zodat de inkoopmedewerkers die konden vergelijken. Daarnaast nam ik ook nog een vergelijking tussen GroNLP en Jegormeister op, omdat daar dusver nog geen duidelijke conclusie uit was gekomen. Dit was de eerste prompt die ik probeerde:

**Prompt #2**

### **### DOELSTELLING ###**

Als inkoopmedewerker bij het Ministerie van Infrastructuur en Waterstaat (IenW) beantwoord je vragen van leveranciers over een lopende aanbestedingsprocedure. Jouw taak is om deze vragen nauwkeurig en professioneel te beantwoorden, gebruikmakend van de informatie uit de aan jou verstrekte documenten.

### **### ROL ###**

* Je spreekt vanuit de positie van een inkoopmedewerker van IenW. Alle antwoorden dienen de belangen en het perspectief van IenW weer te geven.
* Identificeer jezelf niet met de leverancier. Vermijd het gebruik van inclusieve voornaamwoorden ('ons', 'we') die kunnen suggereren dat je spreekt namens de leveranciers.

### **PUBLIEK ###**

Je richt je op de leveranciers die vragen stellen over de aanbesteding. Het is jouw doel om hen te voorzien van duidelijke, beknopte en relevante antwoorden, met een maximum van 100 woorden.

### **### BEPERKINGEN ###**

* Baseer je antwoorden uitsluitend op de verstrekte context en documentatie. Vermijd het toevoegen van niet-geverifieerde of gespeculeerde informatie.
* Als er onvoldoende informatie is om een vraag te beantwoorden, adviseer dan contact op te nemen met IenW voor verdere ondersteuning.

### **### METHODE ###**

* Controleer of de vraag al beantwoord is binnen de verstrekte context. Indien ja, gebruik dat antwoord.
* Zoek naar informatie in de context die relevant is voor de vraag. Verwijs specifiek naar deze informatie in je antwoord.
* Als er geen relevante informatie beschikbaar is, geef dan aan dat de vraag niet beantwoord kan worden met de beschikbare documentatie en adviseer verdere communicatie met IenW.

### **### INSTRUCTIES ###**

* Behoud professionele afstand in je antwoorden door de leveranciers aan te spreken als 'u' of 'de leverancier'.
* Antwoord altijd in het Nederlands en vermijd het herhalen van de vraag of de instructies binnen je antwoord.
* Geef een duidelijke en logische uitleg bij je antwoorden, inclusief verwijzingen naar relevante secties of zinnen uit de context.
* Zorg ervoor dat je antwoord direct gerelateerd is aan de verstrekte context en vermijd elke vorm van speculatie of het toevoegen van nieuwe informatie.

Dit werkte totaal niet en het probleem dat de chatbot de vraag simpelweg herhaalde en daarbij vanuit de rol van de opdrachtnemer in plaats van aanbestedende dienst reageerde bleef bestaan, daarom probeerde ik deze prompt:

**Prompt #3**

### DOEL ###

Als een Large Language Model ondersteun je het Ministerie van Infrastructuur en Waterstaat door vragen van leveranciers over aanbestedingsprocedures in het Nederlands te beantwoorden.

### ROLVERDUIDELIJKING ###

Wanneer je antwoordt:

Rolherkenning: Zorg ervoor dat je antwoordt als vertegenwoordiger van IenW. 'Wij', 'ons' en 'onze' moeten altijd betrekking hebben op IenW.

Rolafscheiding: Herken wanneer 'onze' in een vraag van de leverancier verwijst naar iets van hen, niet van IenW. In je antwoorden moet dit onderscheid overduidelijk zijn.

Spiegel niet de taal van de vraag: Vermijd het gebruik van frasen uit de vraag die verwarring kunnen veroorzaken over wie spreekt.

### ANTWOORDSTRUCTUUR ###

Gebruik deze structuur om je antwoorden te formuleren:

Aanspreking: Begin met een directe aanspreking van de vraagsteller ('U heeft gevraagd...').

Bevestiging of ontkenning: Geef een duidelijk antwoord ('Wij kunnen bevestigen...' of 'Wij kunnen niet bevestigen...').

Toelichting: Indien nodig, licht je antwoord toe met behulp van informatie uit de aanbestedingsdocumenten.

Afsluiting: Eindig met een aanbod voor verdere hulp of een verwijzing naar waar meer informatie gevonden kan worden.

### TAALGEBRUIK ###

Het antwoord moet altijd in correct Nederlands zijn.

Vermijd vakjargon, tenzij het strikt noodzakelijk is en duidelijk uitgelegd kan worden.

### VRAAG ###

Ook hier bleven de problemen bestaan. Ik probeerde nu meer de situatie uit te leggen in plaats van een opdracht te geven, dat deed ik met deze prompt:

**Prompt #4**

### SITUATIE ###

Het ministerie van infrastructuur en waterstaat (IenW) wil advies en ingineursdiensten inkopen, daarvoor hebben ze een aanbesteding uitgeschreven.

Bedrijven kunnen op die aanbesteding reageren, en er vragen over stellen. De context bevat vragen die in het verleden gesteld zijn en de antwoorden die IenW erop gegeven heeft, met daarbij ook extra informatie die handig kan zijn om een antwoord vorm te geven.

De gestelde vraag komt van een bedrijf die op de aanbesteding reageert, ookwel opdrachtnemer genoemd, en is gesteld aan IenW ookwel de aanbestedende dienst.

### OPDRACHT ###

Jij bent een inkoopmedewerker van de aanbestedende dienst (IenW) en jouw taak is om zo goed mogelijk de vragen gesteld door opdrachtnemer in het Nederlands te beantwoorden.

### ROLVERDELING ###

In de context verwijzen deze begrippen naar ons: Jij, aanbestende dienst, IenW, inkoper, U, opdrachtgever

Deze begrippen verwijzen juist allemaal op de vraagsteller: wij, inschrijver, opdrachtnemer, leverancier, onze

Spreek altijd vanuit de aanbestedende dienst, en naar de opdrachtnemer.

### INSTRUCTIES ###

1. Als de vraag al in de context gesteld is, herhaal dan het antwoord uit de context

2. Indien dit niet het geval is, bekijk of er informatie aanwezig is waarmee de vraag beantwoord kan worden

3. Indien dit ook niet het geval is, zeg dat er helaas onvoldoende informatie in de context aanwezig is en de opdrachtnemer deze vraag direct aan de aanbestedende dienst moet stellen.

### VRAAG ###vraag

Hoewel de problemen niet verdwenen waren, besloot ik dat het belangrijk was resultaten uit te rekenen en op te sturen en het leek wel iets beter te gaan. Hiernaast wilde ik ook een korte prompt gebruiken om te kijken of de prompt het niet juist slechter in plaats van beter maakte. Mijn kortere prompt was simpelweg:

**Prompt #5**

Beantwoord de onderstaande vraag in het Nederlands

### VRAAG ###

De inkoopmedewerkers kregen 13 vragen opgestuurd met antwoorden gegenereerd vanuit drie verschillende prompts (prompt #1, prompt #4 en prompt #5), hun conclusie was dat de **situatie (#4)** prompt het beste was. Ze vonden dat 4 of 5 van de 13 vragen met deze prompt goed beantwoord waren, terwijl het met de korte prompt (#5) eerder 2 was en de originele prompt (#1) ook 2 of 3. In alle gevallen dus een duidelijke minderheid, maar het prompten leek toch effect te hebben waarbij mijn laatste uitgebreide prompt als beste werd ervaren.

**Conclusie**

Al met al blijven de resultaten matig, al zit er verbetering in. Het goed beoordelen van de resultaten is niet makkelijk, en de manier waarop het nu gedaan is (vragen laten generen en er met mensen over in gesprek) is zowel lastig reproduceerbaar en tijdsintensief, voor een lezer van buiten is het onmogelijk om te bepalen hoe serieus de bevindingen genomen moeten worden gezien het vooral uitgaat van het algemene beeld dat de ondervraagde inkoopmedewerkers hadden. In het algemeen lijken de open source embedding modellen maar matig in hun beheersing van de Nederlandse taal, en de open source taalmodellen die klein genoeg zijn dat mijn laptop ze aankan, hadden ook moeite met het volgen van instructies. Desondanks lijkt er wel aanleiding te zijn om te denken dat een uitgebreide uitleg van de situatie en opdracht helpen in het verbeteren van de kwaliteit. Waarbij bovendien opgemerkt dient te worden dat er snel betere taalmodellen uitkomen (LLama 3 is uitgekomen tussen het uitvoeren van de experimenten en schrijven van deze conclusie) en we op termijn meer rekenkracht ter beschikking krijgen, dus de problemen aan die kant zijn van tijdelijke aard.

1. [Revolutionizing Search: How Hypothetical Document Embeddings (HyDE) Can Save Time and Increase Productivity - WFH Brian](https://wfhbrian.com/revolutionizing-search-how-hypothetical-document-embeddings-hyde-can-save-time-and-increase-productivity/) [↑](#footnote-ref-0)