**RAG met openbare aanbesteding, resultaten en bedenkingen**

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. Gezien we zeer beperkte rekenkracht hadden was lokaal draaien een tijdsintensieve opdracht, daarom hebben we eerst gekeken of het concept werkte met openbare informatie.

Om deze reden hebben we ook gekeken naar mogelijkheden in de cloud of via een API key te experimenteren, met het blik op optimale veiligheid hebben we daarbij gekozen om dat met openbare documentatie te doen. Er is een WOB verzoek geweest rond de inkoop van de Gezamenlijke Ontwikkeling Universele rijksDesktop (afgekort als GOUD). De beschikbare documentatie daarvoor bestaat uit:

* 2 nota’s van inlichtingen
* Aanbestedingsdocument
* Programma van eisen en wensen
* Selectiedocument
* Contra-expertise, reactie op die contra-expertise en reactie op die reactie
* Verslag van verhoren in 2e kamer

Gezien bij de use cases die inkoopmedewerkers voor ogen hebben doorgaans geen kamerdebatten zijn, zijn de verslagen van de 2e kamer verhoren niet gebruikt.

Ook de discussie rond de contra-expertise is vaak buiten beschouwing gelaten, temeer omdat dat niet altijd aanwezig is en voor veel vragen niet relevant was.

Het aanbestedingsdocument en het programma van eisen en wensen zijn niet als tekstbestand maar als afbeelding beschikbaar, de ene applicatie had betere OCR tools dan de andere, dus deze documenten waren niet bij ieder experiment beschikbaar, dit zal erbij vermeld worden.

Verder is de laatste NVI vaak express achtergehouden, om vragen die daarin gesteld worden op basis van de rest van de documentatie te laten beantwoorden.

Bij deze eerste experimenten is vooral globaal gekeken naar de accuraatheid, er is daarom louter beoordeeld of het antwoord grofweg of helemaal overeenkwam met het daadwerkelijke antwoord vanuit de inkopers. Dit is gedaan omdat het hier gaat om een eerste verkenning voor een globaal beeld van kwaliteit om te zien of er potentie is om mee verder te werken, waarbij veel instellingen nog veranderd zullen worden, en geen intentie voor een grondige en definitieve vergelijking.

**Experiment 1**

Als eerste heb ik in de appl-docchat applicatie van PBL (via API-key) vragen gesteld die letterlijk uit de NVI kwamen die als een van de documenten meegegeven was. Dit gaf de volgende resultaten:

Correct: 9

Onvolledig: 4

Incorrect: 0

Onbeantwoord 2

Dit stemt optimistisch, het laat zien dat de opmaak te begrijpen is en veel antwoorden teruggevonden worden, zelfs als het niet voor alle antwoorden geldt en er soms een gedeeltelijk antwoord gegeven wordt. Het geeft iets om mee te werken

Ik heb daarna vragen omgeschreven om de semantische zoekfunctie te testen, hierbij is het idee dat ik zo min mogelijk woorden gebruik die letterlijk in de vragen staan zodat die niet op trefwoorden kan zoeken maar de betekenis van de vraag grofweg gelijk blijft. Als het semantisch zoeken werkt zou die alsnog de juiste vraag en dus antwoord moeten vinden. Dit ging heel goed:

Correct: 12

Onvolledig: 1

Incorrect:

Onbeantwoord: 2

Feitelijk waren de resultaten hier iets beter dan toen ik de vragen letterlijk stelde, maar gegeven de omvang van het experiment is dat waarschijnlijk ruis. Ik zou daar niet te veel in lezen, zoals ik al zei: dit is vormgegeven om een globaal beeld te krijgen en dat is gelijk.

Daarna heb ik 15 vragen uit de 2e Nota van Inlichten laten beantwoorden op basis van de eerste, dit waren daarvan de resultaten:

Correct: 4

Onvolledig: 2

Incorrect: 2

Onbeantwoord: 7

Dat er tweemaal een antwoord verzonnen is, is ongunstig. Verder laat dit vooral zien dat de taak voor veel vragen onmogelijk was, de eerste NVI bevatte vaak onvoldoende informatie om de vragen uit de 2e te beantwoorden. Voor de daadwerkelijke use cases is er vaak veel meer informatie beschikbaar dan hier het geval was, dat die met deze minimale informatie toch een paar vragen goed heeft kunnen beantwoorden stemt mij positief.

**Experiment 2**

Ik heb dit ook in Azure herhaald en via de GitHub copilot omgeving, hierin heb ik de volgende systeem prompt gebruikt:

*### OBJECTIVE ###*

Je bent een assistent voor inkoopmedewerkers van de rijksoverheid. Je krijgt vragen van leveranciers. Probeer zo goed mogelijk te voorspellen hoe een inkoopmedewerkers die vraag zou beantwoorden op basis van de beschikbare context.

*### AUDIENCE ###*

Leveranciers van de aanbestede dienst. Geef uitgebreid antwoord, en leg zoveel mogelijk uit zodat de leverancier geen extra vraag hoeft te stellen.

*### GUARDRAILS ###*

Als de context onvoldoende informatie bevat om de vraag te beantwoorden, verzin dan geen informatie maar zeg dat er onvoldoende informatie beschikbaar is.

*### INSTRUCTIONS ###*

1. Refereer naar de zinnen uit de context die je gebruikt hebt om tot het antwoord te komen.
2. Redeneer stap voor stap

Al snel bleek dat letterlijke vragen waarschijnlijk geen groot probleem in Azure zouden worden, temeer omdat in Azure gebruikgemaakt wordt van een hybride zoekfunctie waar ook op basis van trefwoorden gezocht wordt. Daarom ben ik na een paar vragen doorgegaan naar het omschrijven van vragen. Dit heb ik door GPT-4 laten doen met de volgende prompt:

“Ik heb een opdracht voor je, ik type een vraag in en wil dat jij die vraag zo omschrijft dat de betekenis van de vraag hetzelfde blijft maar er zo min mogelijk dezelfde woorden en zinsconstructies instaan. De bedoeling is dat een zoekmachine die werkt op trefwoorden, het niet terug kan vinden. ”

Wederom heb ik 15 (willekeurige) vragen gesteld uit de eerste Nota van Inlichtingen, dit waren de resultaten:

correct: 10

onvolledig: 3

incorrect:

onbeantwoord: 2

Dit is vergelijkbaar met wat we in de PBL applicatie zagen, wederom het zijn kleine steekproeven dus de exacte cijfers zeggen niet heel veel maar het gaat meer om de orde van grootte.

Ik heb dit tweemaal gedaan en tweemaal zeer andere resultaten gekregen, ik had van de eerste keer deze resultaten genoteerd:

correct: 7

(zeer) onvolledig: 5

incorrect: 3

Ik had echter niet goed gedocumenteerd hoe ik aan deze resultaten gekomen was, en opvallend genoeg is er in de documentatie nooit geweigerd een vraag te beantwoorden. Daarom had ik het opnieuw gedaan, dit keer genoteerd welke 15 vragen er gebruikt zijn, dat zijn:

Gebruikte vragen:

* 6 ™ 11
* 60 ™ 68

Ik wist zelf weinig van het onderwerp af, en soms werden antwoorden gegeven waarvan het voor mij niet evident was wat de kwaliteit van dat antwoord was, dat heb ik als categorie onzeker opgenomen.

Resultaten:

correct: 2

(zeer) onvolledig: 4

(zeker) incorrect: 0

onbeantwoord: 6

onduidelijk: 3

Deze resultaten komen waarschijnlijker over dan de andere, en meer in lijn met wat we ook bij de appl-docchat applicatie gezien hebben. Ik denk dat deze resultaten meer over de data zelf zeggen dan de applicatie, het kan heel goed dat veel vragen uit de tweede nota van inlichtingen niet af te leiden zijn uit de eerste. De Azure applicatie keek vaak naar minder verschillende stukken tekst.

**Conclusie:**

Of je rekenkracht uit Azure gebruikt of een API-key lijkt niet bijzonder veel uit te maken voor de kwaliteit van de resultaten, wat ergens logisch is gezien dezelfde modellen gebruikt worden. Beide applicaties lukt het redelijk goed (maar niet perfect) vragen terug te vinden in een Nota van Inlichtingen, ook als de vraag geherformuleerd is. Daarentegen leek het amper mogelijk vragen uit een 2e Nota van Inlichtingen te beantwoorden op basis van het aanbestedingsdocument en de 1e Nota van Inlichtingen, wat goed aan beperkingen van de data zelf zou kunnen liggen dus lastig om daar meer conclusies uit te trekken.

Wel wil ik opmerken dat ik in het doen van de experimenten de methodiek beter had kunnen structuren, welke vragen waarom gesteld zijn was nu vaak onduidelijk. Bovendien zijn sommige antwoorden lastig te beoordelen, dus ik moet nog verder denken over een gestandaardiseerde manier om daarmee om te gaan. Voor een eerste indruk was de huidige methodiek echter afdoende.