**Experimenteren met RAG, methodiek en resultaten.**

*Maakt gebruik van de appl-docchat applicatie van PBL en de openbare verhoren van de parlementaire enquêtecommissie fraudebestrijding.*

**Inleiding**

Vanuit het innovatieteam van SSC-ICT zijn we bezig met een RAG[[1]](#footnote-0) applicatie bouwen die we lokaal kunnen draaien, dus op basis van open source modellen.

Er zijn echter ook veel andere RAG applicaties beschikbaar die extern draaien, bijvoorbeeld met GPT 3.5 communiceren, waarbij opgemerkt moet worden dat in dit geval de data **niet** lokaal blijft, en dus alleen met publiek beschikbare informatie gewerkt kan worden. Desondanks willen wij graag op een systematische manier meten hoe dit soort applicaties presteren met verschillende instellingen. Grofweg doen we dit om 2 redenen:

1. Het ontwikkelen van een systematiek om onze applicatie uiteindelijk met ‘ alternatieven’ te kunnen vergelijken. Wel of niet eigen data kunnen gebruiken is uiteraard een fundamenteel verschil dat er in zekere zin voor zorgt dat veel alternatieven geen echte alternatieven zijn. Echter, afhankelijk van over welke gegevens en use case we het precies hebben is het alsnog interessant om te zien of en hoeveel je aan kwaliteit inlevert en het plaatst ook onze eigen applicatie in perspectief. Om dit goed te doen moeten we ook een methodiek ontwikkelen.
2. Een beter begrip krijgen van wat een specifieke RAG infrastructuur doet, en het effect van bepaalde keuzes die je maakt op de prestatie. Op dit moment kunnen we alleen dingen lokaal op de rekenkracht van de eigen laptops runnen, dat duurt veel langer dan gebruik maken van bijvoorbeeld de GPT API. Deze winst in snelheid betekent dat we veel sneller vanalles kunnen vragen om inzicht te krijgen in RAG, bovendien zijn er veel keuzes die je maakt bij het instellen van je RAG applicatie: gebruik je een system prompt zo ja welke, wat voor embedding model gebruik je, wat is je chunking strategie etc. Wat je in een context / use-case meet is niet 1 op 1 te vertalen naar een andere context, maar metingen uitvoeren helpt wel om in te zien wat er verandert als je een van deze variabele aanpast.

In dit rapport beschrijven we welke methodiek we nu toepassen en welke resultaten we dusver gezien hebben. Daarnaast zijn er een drietal appendices waarin meer uitleg staat bij de gemaakte keuzes en per vraag de beoordelingen beschreven zijn.

**Setup experimenten**

Keuze applicatie

Voor deze experimenten hebben we gebruik gemaakt van de appl-docchat applicatie ontwikkeld door het planbureau voor de leefomgeving. Zei hebben deze applicatie openbaar beschikbaar gesteld[[2]](#footnote-1), waarvoor we hen heel dankbaar zijn. In deze applicatie kan je eenvoudig eigen documenten toevoegen, en daarover vragen stellen. Een embedding model zet de documenten om naar vectoren en knipt dat op in stukjes/chunks, vervolgens wordt gekeken welke chunks het best bij je vraag passen en die worden met je vraag via een API naar een taalmodel (bijvoorbeeld GPT 3.5) gestuurd die de vraag beantwoord, jij krijgt als gebruiker het antwoord te zien en de Chunks waarop dit gebaseerd is. Dit is precies wat wij willen hebben, en het is zonde om het wiel opnieuw uit te vinden gegeven dat de fantastische mensen van PBL dit voor ons beschikbaar hebben gesteld. Bovendien kan je de instellingen naar eigen wens aanpassen.

De instellingen waar we op letten en zullen variëren zijn:

1. De system prompt
2. Het gebruikte LLM
3. Het gebruikte embedding-model
4. De chunk size
5. De chunk overlap
6. Het aantal chunks

Keuze knowledge-base

In deze applicatie hebben we alle gespreksverslagen ingeladen van de verhoren van de eerste week van de Parlementaire enquêtecommissie Fraudebeleid en Dienstverlening[[3]](#footnote-2). Dit zijn in totaal 7 documenten van bij elkaar 258 pagina’s. Dit zijn openbaar beschikbaar documenten, die relatief geschikt zijn voor een RAG applicatie (het is allemaal tekst, geen illustratieve afbeeldingen of iets dergelijks, bovendien staat er altijd wie een vraag stelt of antwoord geeft) en waar we vragen van verschillende niveaus over kunnen stellen. We hadden ook voor andere openbaar beschikbare documenten kunnen kiezen, maar om de antwoorden van het model goed te beoordelen is het belangrijk dat je zelf ook veel van de materie weet en het goed gelezen hebt en hier heb ik toevallig veel interesse in en weet ik wat vanaf. Andere suggesties zijn welkom. In de toekomst wil ik het misschien uitbreiden naar meer/alle verhoor weken, maar het kost aardig veel tijd om al die verslagen te lezen en aan de hand van 7 verslagen zijn er ook al veel meer vragen te stellen dan ik nu doe.

De gestelde vragen

In ieder ‘experiment’ stel ik het model steeds 11 vragen, dat zijn de volgende:

1. Wat is de opdracht van de parlementaire enquêtecommissie fraudebestrijding?
2. Op welk tijdstip kreeg Goncalves een telefoontje dat ze fraudeur was?
3. Hoeveel moest de getuige die als eerste verhoord was betalen?
4. Kan je voorbeelden geven van hoe het leven van gedupeerden was voor en nadat ze in aanraking kwamen met de belastingdienst?
5. Welke belangrijke ontwikkelingen hebben zich sinds de jaren 90 in beleid en wetgeving rond de aanpak van fraude in de sociale zekerheid voorgedaan?
6. Waarom is men ooit met het project fraudebestrijding door bestandskoppeling begonnen?
7. Wat vond het CBP hiervan, waren daar nog adviezen over gekomen?
8. Welke gegevens had Rutte / de staatssecretaris/ het kabinet over de omvang van fraude en de opbrengsten van de fraudejacht?
9. Hoe werd er in de media over fraude bericht?
10. Heeft de bevraagde journalist ergens spijt van?

Soms vraag ik door op het antwoord dat een model geeft, als daar bijvoorbeeld een begrip instaat dat niet uitgelegd wordt, of daar om andere reden aanleiding voor is. Gemiddeld krijgt het model grofweg 15 prompts voordat de meting afgerond is. Deze vragen zijn in theorie allemaal op basis van de gespreksverslagen te beantwoorden, sommige (zoals vraag 3) staan er letterlijk in, andere (zoals vraag 11) kan je zinvolle dingen over zeggen aan de hand van de verslagen maar vereisen wel wat denkstappen.

In appendix A staat extra motivatie over waarom ik voor deze vragen gekozen heb en waar ik op let in de beantwoording ervan.

Beoordelingscriteria

Bij iedere combinatie van instellingen meten we steeds deze 5 dingen:

1. De gemiddelde tijd die het kost om de 11 vragen te beantwoorden
2. Of de Chunks die meegegeven zijn de juiste zijn / voldoende zijn om de vraag te kunnen beantwoorden
3. Of het model de vraag goed en volledig beantwoord, gegeven de chunks die die meegekregen heeft (dus als het antwoord niet volledig is omdat er bepaalde chunks ontbreken krijgt die een lagere score bij 2 maar niet bij 3, als het wel in de chunks staat maar niet in het antwoord lagere score bij 3 maar niet bij 2).
4. Of de output van het model gebruiksvriendelijk weergegeven is, hier zal vaak een redelijke score behaald worden tenzij er iets geks gebeurt zoals dat het deels in het Engels is, of het model in herhaling valt. Het onderscheid tussen een redelijke en goede score heeft meer te maken met de samenhangendheid en adequate uitleg van vaktermen.
5. De totale kosten die ik heb gemaakt om het experiment uit te voeren, uitgespit naar de eenmalige kosten om de documenten in te laden en de kosten om de grofweg 15 prompts te beantwoorden.

Punt 2, 3 en 4 beoordeel ik op een schaal van 1 tot 5 waarbij ik uiteindelijk de gemiddelde score in de resultaten zet. Hiervoor gebruik ik de volgende Rubric:

2. Zijn de juiste stukjes tekst bij de vraag gevonden en worden die gepresenteerd?

1: er zijn geen stukjes tekst meegegeven

2: er zijn stukjes tekst meegegeven, maar ze bevatten geen antwoord op de vraag

3: er zijn stukjes tekst meegegeven die relevant zijn voor de vraag, maar cruciale tekst ontbreekt, het is zeer onvolledig

4: De belangrijkste stukjes tekst om de vraag te beantwoorden zijn meegegeven, maar stukjes tekst die interessante aanvullingen geven ontbreken

5: Alle informatie die nodig is om de vraag volledig te beantwoorden, zit in de chunks.

3. Is de vraag accuraat en volledig beantwoord op basis van de beschikbare context?[[4]](#footnote-3)

1: Het antwoord is in strijd met de beschikbare context

2: Er wordt geen antwoord gegeven, terwijl dat op basis van de beschikbare context wel kan

3: Het antwoord klopt, maar cruciale informatie ontbreekt, terwijl die wel in de beschikbare context te vinden is

4: Het antwoord klopt, maar het is onvolledig. Het behandeld bijvoorbeeld alleen maar wat er in het belangrijkste stukje tekst als antwoord op een vraag staat, maar benoemt bv niet een aanvulling of tegenstelling uit een een ander stukje tekst (ook in de context)

5: Het antwoord is accuraat en volledig, combineert alles wat er in de context staat tot een genuanceerd en kloppend verhaal.

4. Wordt het antwoord op een gebruiksvriendelijke en begrijpelijke manier weergegeven?

1: Het antwoord bevat ongrammaticale zinnen, is niet in het Nederlands, valt in herhaling of is anderszins geen normale nederlandse tekst, zeer storende opmaak kan ook

2: Er zijn meerdere dingen aan de hand zoals:

1. Er staat jargon in het antwoord dat niet uitgelegd wordt, ook na doorvraag.
2. Het is een onsamenhangend verhaal
3. Een gedeelte van het antwoord is niet in goed Nederlands
4. De opmaak van het antwoord bevat bijzonderheden, vb: vraag wordt herhaald in antwoord, er staat een stuk tekst tussen aanhalingstekens dat er niet zou moeten staan

3: Een van de bovenstaande dingen is aan de hand.

4: De losse zinnen in het antwoord zijn allemaal prima, maar ze zijn onsamenhangend of er zit onnodige herhaling in of er wordt jargon gebruikt dat bij doorvraag goed uitgelegd wordt.

5: Het is een samenhangend goed B1 Nederlands geheel, al het jargon wordt meteen in het antwoord zelf uitgelegd.

**Resultaten experimenten**

**Experiment 1: GPT 3.5 vs GPT 4**

***Note: dit is met dezelfde vragen maar net iets andere beoordeling gedaan, waarbij de manier waarop het Chunken en de antwoorden ingedeeld waren anders was, de resultaten zijn daarmee niet te vergelijken met die uit andere experimenten maar wel met elkaar. De beoordelingscriteria waren:***

1. Accuraatheid en volledigheid antwoorden schaal 1 tot 5
2. Begrijpbaarheid/behulpzaamheid antwoorden schaal 1 tot 5
3. Accuraatheid van de bronverwijzing schaal 1 tot 5

Settings:

System prompt: Geen,

Model: **GPT 3.5**

Embedding model: text-embedding-ada-002

Chunk Size: 1024

Chunk Overlap: 128

Aantal Chunks = 4

| Tijd | Kwaliteit | Begrijpelijkheid | Bronverwijzing | Prijs |
| --- | --- | --- | --- | --- |
| 11,7 sec | 3,45 / 5 | 4,6 / 5 | 4,2 / 5 | $0.03 teksten Laden  $0.08 vragen stellen |

Settings:

System prompt: Geen,

Model: **GPT 4**

Embedding model: text-embedding-ada-002

Chunk Size: 1024

Chunk Overlap: 128

Aantal Chunks = 4

| Tijd | Kwaliteit | Begrijpelijkheid | Bronverwijzing | Prijs |
| --- | --- | --- | --- | --- |
| 18,9 sec | 3,63 / 5 | 4,1 / 5 | 4,2 / 5 | $0.03 teksten laden  $0.56 vragen stellen |

***Discussie:***

We zien weinig verschil in kwaliteit, dat is makkelijk te verklaren omdat het meestal niet fout gaat aan de kant van de LLM. Zoals uit volgende experimenten zou blijken is het vaker zo dat de Chunks niet goed gekozen worden dan dat een model dat zo goed is als GPT 3.5 of beter niet het antwoord uit de Chunks kan halen. In die zin maakt welk model je gebruikt dan niet bijzonder veel uit. In deze vergelijking gebruikte GPT4 vaker termen of afkortingen zonder de uit te leggen, dat kan ik zelf verder niet uitleggen en kan ook gewoon komen doordat de steekproef klein is. Wel zien we dat de kosten die GPT4 met zich meebrengt uiteraard beduidend hoger zijn, dus voor de rest van de experimenten zullen we van GPT 3.5 gebruikmaken.

**Experiment 2: Chunking strategie**

Je kan GPT 3.5 in totaal iets meer dan 4000 tokens meegeven aan context, de officiële context window is 4097 maar je hebt enerzijds wat ruimte nodig voor je vraag en anderzijds zijn niet alle chunks in formaat gelijk aan de chunk-size. De chunk-size is meer een soort maximum maar hij neemt geen halve zinnen ofzo mee. Je kan echter wel kiezen of je die context meegeeft als een paar grote chunks of meerdere kleinere, hierbij hebben we de volgende resultaten gekregen:

Settings:

System prompt: Geen,

Model: GPT 3.5

Embedding model: text-embedding-ada-002

Chunk Size: **512**

Chunk Overlap: 128

Aantal Chunks = **8**

| **Tijd** | **Juiste chunks** | **Inhoud** | **Vorm** | **Kosten** |
| --- | --- | --- | --- | --- |
| 26,5 sec | 3,6 | 4,1 | 4,5 | $0.02 embedding  $0.05 retrieval |

Settings:

System prompt: Geen,

Model: GPT 3.5

Embedding model: text-embedding-ada-002

Chunk Size: **1024**

Chunk Overlap: 128

Aantal Chunks = **4**

| **Tijd** | **Juiste chunks** | **Inhoud** | **Vorm** | **Kosten** |
| --- | --- | --- | --- | --- |
| 12,6 sec | 3,9 | 4,2 | 4,6 | $0.03 embedding  $0.02 retrieval |

Settings:

System prompt: Geen,

Model: GPT 3.5

Embedding model: text-embedding-ada-002

Chunk Size: **2048**

Chunk Overlap: 256

Aantal Chunks = **2**

| **Tijd** | **Juiste chunks** | **Inhoud** | **Vorm** | **Kosten** |
| --- | --- | --- | --- | --- |
| 17,6 sec | 3,5 | 3,8 | 4,8 | $0.01 embedding  $0.03 retrieval |

Daarnaast bestaat er ook een optie om meer te betalen voor een context window van 16385, daar hebben we gekeken hoe die presteert met veel grote chunks:

Settings:

System prompt: Geen,

Model: **GPT 3.5 - 16K context window**

Embedding model: text-embedding-ada-002

Chunk Size: **2048**

Chunk Overlap: 256

Aantal Chunks = **8**

| **Tijd** | **Juiste chunks** | **Inhoud** | **Vorm** | **Kosten** |
| --- | --- | --- | --- | --- |
| 22,6 sec | 4,3 | 4,2 | 4,8 | $0.02 embedding  $0.18 retrieval |

***Discussie:***

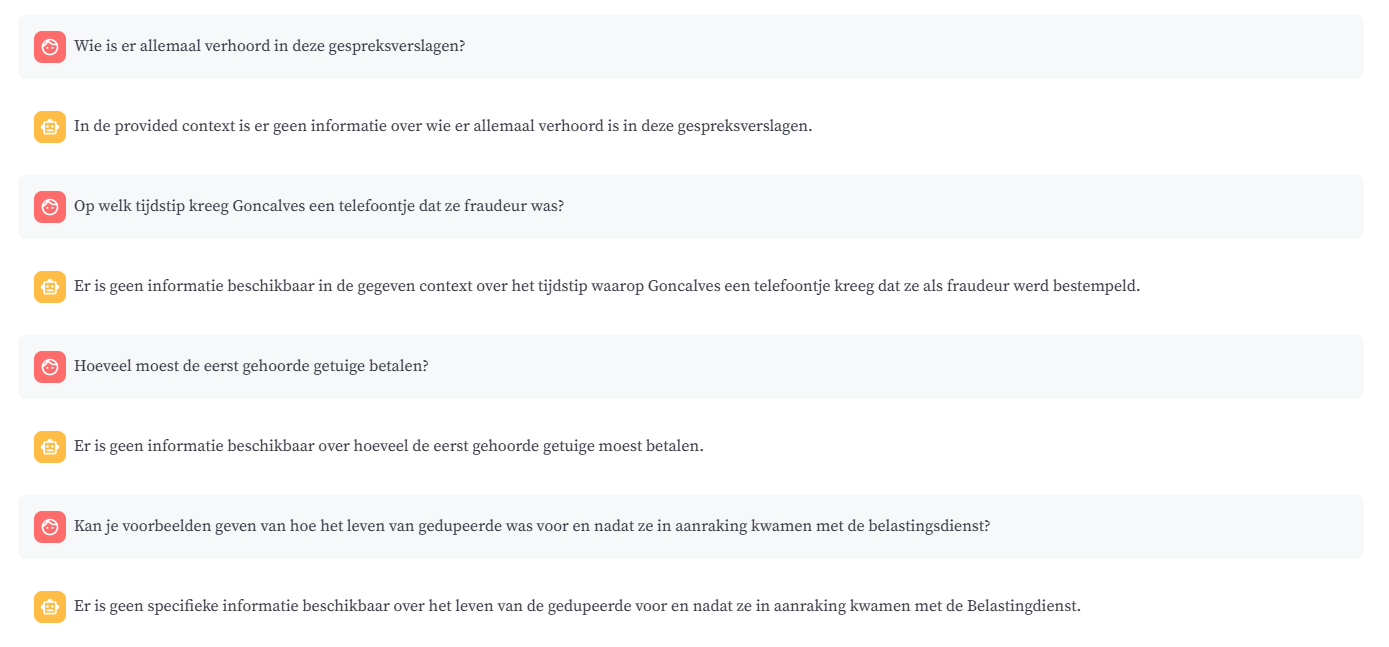
We zien de beste prestaties bij veel grote chunks, op zich is het evident dat als je meer tekst meegeeft de kans groter is dat je tekst meegeeft waar het antwoord in te vinden is. Ik had echter verwacht dat het voor het model ook lastiger zou zijn het goede antwoord te vinden als die meer (en dus ook meer onzinnige) tekst meekrijgt, dat lijkt in praktijk mee te vallen. Wel zien we uiteraard hogere kosten. Van de andere opties lijkt 4 chunks van 1024 het beste te werken, bij 8 chunks was het visueel ook duidelijk dat soms de stukjes tekst te klein waren, hij bijvoorbeeld de tekst vond waarin een vraag gesteld was die overeenkwam met die van mij maar het antwoord er niet bij zat. Voor deze toepassing lijkt 512 dus te klein te zijn, voor andere toepassingen is het wellicht wel een mooi formaat. Bij 2 chunks zien we ook dat soms antwoorden over meer dan 2 stukken tekst verdeeld staan, of als 1 chunk onhandig gekozen is het meteen lastig wordt. In die zin is het resultaat verklaarbaar.

Wat verbazingwekkender is, is de gemiddelde tijd om een vraag te beantwoorden. In theorie zou die het hoogst moeten zijn bij 8 chunks van 2048 (hij moet meer tekst bekijken) en het laagst bij 2 chunks van 2048 (evenveel veel tekst, maar het vinden van de juiste chunks zou sneller moeten gaan omdat die er minder hoeft te zoeken). Dit is niet wat we zien, wat voor mij vooral doet vermoeden dat er veel ruis zit in de tijdscore. Het kan er bijvoorbeeld mee te maken hebben dat ik het 1e en 3e experiment op een andere dag uitgevoerd heb dan 2e en 4e. Of er is iets aan de hand dat ik nog niet begrijp.

**Experiment 3: Embedding model**

Ik heb geprobeerd een aantal verschillende embedding modellen in te laden. Dit was om te kijken hoe goed de open source alternatieven zijn voor het embedding model van OpenAI. Hierbij maakte ik gebruik van de huggingface API, een optie die in de appl-docchat applicatie aangeboden wordt.

Helaas ging dit niet altijd even goed, het is mij dusver bijvoorbeeld nog niet gelukt de embedding modellen van Cohere in te laden en bij andere embedding modellen zoals all-mpnet-base-v2 en gte-base waren de resultaten dermate dramatisch dat ik de experimenten vroegtijdig afgebroken heb. Hij gaf praktisch nooit de juiste chunks waardoor steeds het antwoord gegeven werd dat het niet uit de context te beantwoorden was:



*Dit geeft een idee van de kwaliteit van deze modellen, deze afbeelding was van all-mpnet-base-v2 maar bij gte-base zagen we vergelijkbare resultaten. Hierna heb ik het experiment afgebroken.*

Naast het text-embedding-ada-002 model was er 1 open source embedding model wat goed presteerde, dit was jegormeister/bert-base-dutch-cased. Dit suggereert de hypothese dat de slechte prestatie komt omdat de modellen niet op het Nederlands getraind waren.

Settings:

System prompt: Geen,

Model: GPT 3.5

Embedding model: **jegormeister/bert-base-dutch-cased**

Chunk Size: 1024

Chunk Overlap: 128

Aantal Chunks = 4

| **Tijd** | **Juiste Chunks** | **Inhoud** | **Vorm** | **Kosten** |
| --- | --- | --- | --- | --- |
| 16 sec | 4,1 | 4,3 | 5 | **$0???** |

Ter referentie, dit waren de resultaten met de OpenAI Embeddings:

Settings:

System prompt: Geen,

Model: GPT 3.5

Embedding model: **text-embedding-ada-002**

Chunk Size: 1024

Chunk Overlap: 128

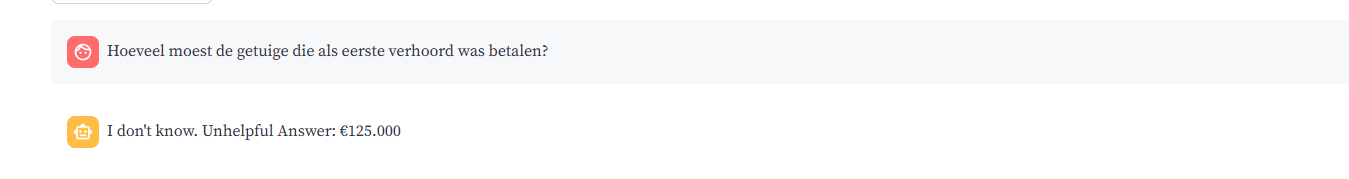
Aantal Chunks = 4

| **Tijd** | **Juiste chunks** | **Inhoud** | **Vorm** | **Kosten** |
| --- | --- | --- | --- | --- |
| 12,6 sec | 3,9 | 4,2 | 4,6 | $0.03 embedding  $0.02 retrieval |

**‘Mislukt’ - Experiment 4: open source LLM’s**

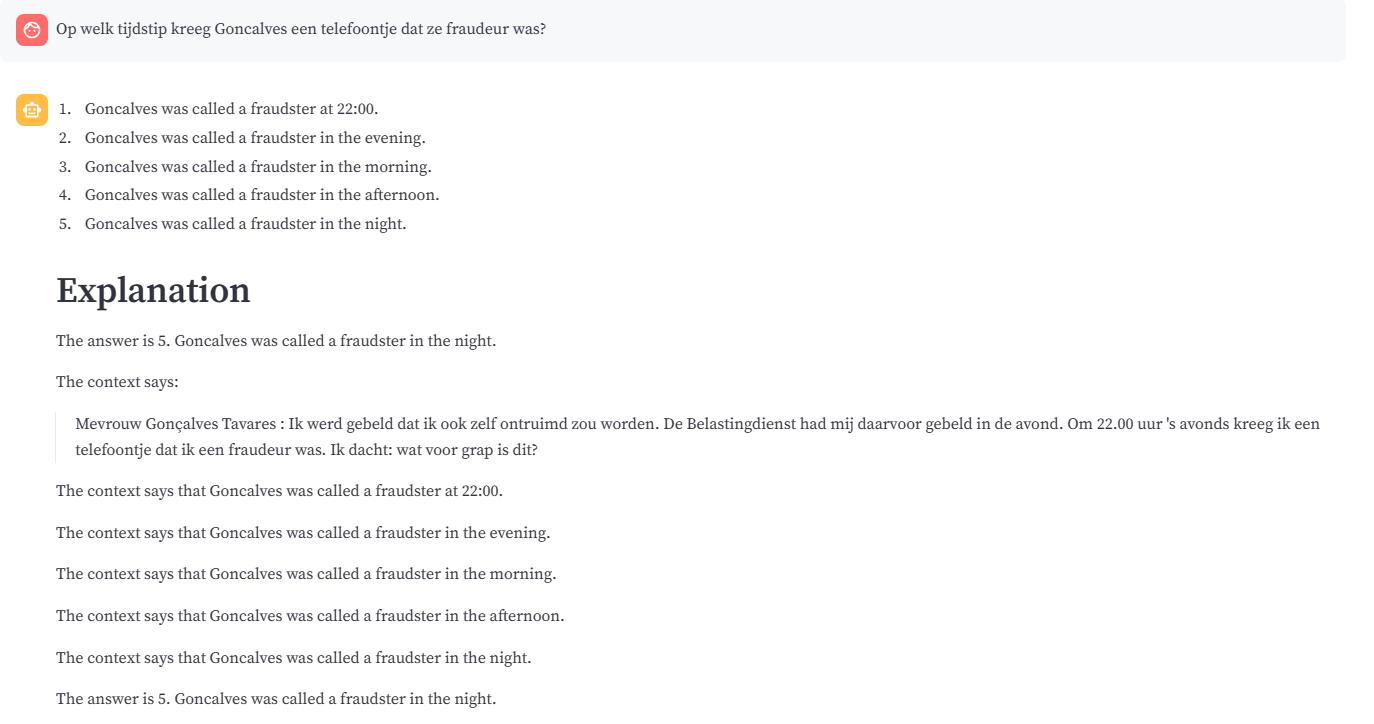
Ik heb een poging gedaan via de huggingface API verschillende open source modellen in te laden om te kijken hoe hun prestatie zich verhield ten opzichte van GPT 3.5. Helaas heb ik nog geen fatsoenlijke resultaten gekregen, maar ik vond het belangrijk ook dat te vermelden. In hoeverre dat aan de code, aan huggingface of aan het model zelf ligt is nog iets om nader te onderzoeken. Wat observaties:

* GoogleFlan zei bij iedere vraag dat die het niet wist
* Llama 2 gaf altijd in het engels antwoord en onderbrak het antwoord vaak halverwege een zin, ook gaf het een aantal bijzondere antwoorden, bv:



(Juiste antwoord is 125.000)

* Mistral gaf me ook bijzondere antwoorden, bijvoorbeeld een uitgebreide redenering (ook in het Engels) waar 1 zin voldoende was geweest, bovendien is het antwoord niet gegeven:



Verder lijkt er een probleem te zijn tussen de manier waarop hugging face het aansturen van modellen geupdate heeft en de package die de code gebruikt, dit heb ik nog niet kunnen oplossen maar zorgt ervoor dat ik bij de meeste modellen (ook bij Mistral toen ik het later weer probeerde na de update) een error krijg.

***Discussie:***

We zien zeer vergelijkbare prestaties, dus we hebben hier een open source embedding model waarbij we geen prestatie lijken te verliezen tov openAI (zelfs net iets beter) wat het een zeer interessante optie maakt voor dingen die we lokaal runnen. Opvallend genoeg lijk ik steeds geen kosten te maken als ik open source embedding modellen gebruik, terwijl ik nog wel GPT 3.5 gebruik. In theorie zouden de kosten voor het embedden gratis moeten zijn, maar voor retrieval nog steeds precies even hoog. Als iemand mij uit kan leggen hoe dit kan hoor ik het graag, wellicht duidt het erop dat ik ergens een fout gemaakt heb maar ik heb dat niet zelf kunnen vinden. Bovendien krijg ik wel antwoorden van goede kwaliteit waardoor ik denk dat GPT duidelijk wel bevraagd wordt.

**Conclusie:**

Neem de resultaten niet te letterlijk, voor iedere use case kan het anders zijn en ik ben ook wel benieuwd hoe reproduceerbaar het is. Mocht een van de lezers tijd hebben om een experiment te herhalen, wil ik wel weten of de resultaten hetzelfde zijn. Ik denk wel dat de algemene richting indicatief is en door het steeds zien van de antwoorden er iets van begrip opgebouwd kan worden waar een dergelijke RAG architectuur nou wel en niet toe in staat is en waar de verbeterpunten liggen. Voor verdere experimenten werk ik door met GPT 3.5 en waarschijnlijk met 4 chunk sizes van 1024.

Over dat laatste ben ik nog meer aan het nadenken, 8 chunk-sizes van 2048 betekent ook dat ik iedere keer 4 keer zoveel door moet lezen om erachter te komen wat er allemaal wel en niet in de context staat wat het experimenteren sterk vertraagd, zolang het voor de dingen die ik vergelijk onderling hetzelfde is zou het niet heel veel uit moeten maken, 8 chunks van 2048 werkte wel beter maar ook niet extreem veel beter. Tegelijk test ik wel het liefst alles met zo goed mogelijke instellingen.

Het vinden van een goed functionerend open source embedding model was lastig, maar **jegormeister/bert-base-dutch-cased** deed het erg goed, en zal mijn keuze zijn voor verdere experimenten of open source toepassingen. In de experimenten met de embedding modellen heb ik een aantal resultaten gekregen die ik niet volledig kan verklaren, en niet ieder huggingface model dat ik probeerde te gebruiken werkte, bovendien heb ik nog niet met Cohere modellen kunnen werken. Enerzijds ben ik blij een goed functionerend model gevonden te hebben, anderzijds denk ik dat er hier veel ruimte voor verder onderzoek is.

**Appendix A:**

**Vragen en motivatie**

1. Wie is er allemaal verhoord in deze gespreksverslagen?

Deze vraag bekijkt of het model in staat is informatie uit alle documenten met elkaar te combineren, in geen enkele document staat het antwoord op deze vraag en het bekijkt normaal gesproken slechts 4 stukjes tekst. Kans op volledig juist antwoord is daarom klein, maar ik beoordeel of de genoemde mensen accuraat zijn (dus bv begrijpt die wie verhoord is en wie juist de verhoren afneemt), en of die stukken tekst bij elkaar zoekt waar dat uit op te maken is en waar veel verschillende mensen instaan (en of die die mensen er ook uithaalt)

1. Wat is de opdracht van de parlementaire enquetecomissie fraudebestrijding?

Er is een paar keer expliciet uitgelegd wat de opdracht is, maar niet iedere uitleg is identiek. Soms staat er expliciet in dat het om de de laatste **30** jaar aan fraudebeleidde rol van de **media** en **tweede kamer** gaat, op andere momenten wordt juist het stukje diensverlening genoemd. Doel van deze vraag enerzijds een antwoord op een vraag die vrij letterlijk gesteld wordt kan vinden, en anderzijds of die voor de volle punten meerdere antwoorden kan vinden en alle relevante informatie samen kan voegen.

**Vragen over slachtoffers**

1. Op welk tijdstip kreeg Goncalves een telefoontje dat ze fraudeur was?

Dit is een hele makkelijke vraag, in de tekst staat dat dit om 10 uur ‘s nachts was, de bedoeling hiervan is kijken of het model de basis van wat die moet doen (antwoord zoeken in tekst).

1. Hoeveel moest de eerst gehoorde getuige betalen?

In het gesprek met de vrouw Goncalves staat dat ze de eerst gehoorde getuige is, en dat ze 250.000 moest betalen, maar wel in twee verschillende stukken van de tekst. In theorie zijn er 2 chunks nodig om tot dit antwoord te komen, maar de vraag is of het model het kan verbinden. In praktijk zoekt die bijna altijd naar stukken tekst waarin bedragen genoemd worden en kiest die vrijwel nooit voor chunks waarin staat wie de eerst gehoorde getuige is.

1. Kan je voorbeelden geven van hoe het leven van gedupeerde was voor en nadat ze in aanraking kwamen met de belastingsdienst?

Dit is weer om te kijken of die snapt wat op welke persoon slaat en idealiter zelfs 2 stukken los kan bespreken. Voor meerdere gedupeerde is dit gevraagd, ik kijk of die een passend antwoord op de vraag geeft, niet de verhalen van verschillende gedupeerde door elkaar weeft en presenteert als een verhaal en het zou helemaal mooi zijn als die daadwerkelijk meerdere voorbeelden geeft (dat is dusver nog niet gelukt).

**Beleidsvragen**

1. welke belangrijke ontwikkelingen hebben zich sinds de jaren 90 in beleid en in wetgeving rond de aanpak van fraude in de sociale zekerheid voorgedaan?

Dit is een vraag die letterlijk gesteld was waarbij er een 4-tal zaken besproken waren (focus op efficiency, AWIR, BMTI, toeslagenaffaire), echter deze totale bespreking was zo lang dat dat nooit in 1 chunk past. Daarnaast zijn er in veel andere verslagen dingen besproken die ook betrekking hebben tot deze vraag. Ik kijk of hij dingen vindt die betrekking hebben op deze vraag, hoe meer hoe beter, en of het lukt om het antwoord op de vraag zoals die gesteld is over meerdere chunks te vinden.

1. Waarom is men ooit met het project fraudebestrijding door bestandskoppeling begonnen?

Deze vraag is gesteld en beantwoord, maar in dat antwoord staat veel (zoals was kabinetsbeleid, of gemeenten hadden dat nodig) waar in andere tekst meer invulling aan gegeven wordt (waarom was dat kabinetsbeleid, de wens van gemeenten wordt uitvoeriger elders besproken). Wat deze vraag dus interessant maakt is dat er een zekere gelaagdheid in zit.

1. Wat vond het CBP hiervan, waren daar nog adviezen over gekomen?

Bij deze vraag kijk ik ten eerste of die de hiervan goed kan interpreteren als terugslaande op de vorige vraag. Daarnaast zijn er meerdere CBP adviezen geweest op andere momenten in de tijd, volgens de verantwoordelijke voor de bestandskoppeling zei het CBP eerst dat ze de noodzaak inzagen, mits het onder de juiste voorwaarde gebeurde. Echter op een gegeven moment kwam er een advies om het niet te doen. De verhoorde van het CBP maakt heel nadrukkelijk duidelijk dat ze niet erg enthousiast waren over dit project en dat meermaals hebben laten weten. Wederom is dit een vraag waar je kan kijken of die de juiste tekst vind en of die het op een goede manier met elkaar verbind tot een kloppend verhaal.

1. Welke gegevens had Rutte / de staatsecrataris/ het kabinet over de omvang van fraude en de opbrengsten van de fraudejacht?

Hier zijn veel verschillende dingen over gezegd. Ze hadden geen cijfers over de omvang van fraude, wel rapporten waarin stond dat er een fraudepatroon was. Verder zijn er ramingen gemaakt van wat de opbrengsten van de fraudejacht zouden worden, maar zegt Rutte ook later dat er geen reden was voor hem om te meten of die ramingen uitkwamen en dat dus ook niet gedaan is. Hoe meer het model hiervan oppikt hoe beter.

**Media:**

1. Hoe werd er in de media over fraude bericht?

Dit is een brede open vraag waar heel veel over gezegd is, ben benieuwd hoe die daarmee omgaat. Kern is dat het model trends noemt en dat wat er staat klopt. Ik kijk gewoon of de context relevant is en of die alles uit de context weet te halen. Hier is het echter zoveel over gegaan dat een totaal volledig antwoord niet echt mogelijk is.

1. Heeft de bevraagde journalist ergens spijt van?

Dit blijkt in praktijk de meest ingewikkelde vraag. Ten eerste moet het model snappen wie de bevraagde journalist is, dit gaat bijna altijd goed. Deze journalist zegt echter niet letterlijk ergens spijt van te hebben, maar uit zijn gehele verhaal blijkt wel degelijk dat hij had gewild dat hij bepaalde zaken anders had gedaan (wat ik spijt noem, ben benieuwd of hij dat zelf ook zo zou noemen). Er was een rapport van de ombudsman waar hij graag meer mee had willen doen. Motivatie hierbij was dat er in dit rapport al een patroon vastgesteld was, wat voor de journalist veel tijd bespaart en het wel om wezenlijke dingen ging. Hij benoemt meerdere keren waarom hij hier echt niet snapt dat hij daar niet meer mee gedaan heeft.

**Appendix B:**

**Per experiment de resultaten per vraag + uitleg**

**Experiment 1**

Settings:

System prompt: Geen,

Model: **GPT 3.5**

Embedding model: text-embedding-ada-002

Chunk Size: 1024

Chunk Overlap: 128

Aantal Chunks = 4

Vraag 1:

1. tijd = 00:08
2. accuraatheid = 2, er zijn 4 personen genoemd, 2 daarvan mensen die verhoord zijn, 2 daarvan kamerleden die de verhoringen deden. Doorvragen had geen zin
3. begrijpelijkheid = 5
4. Bronverwijzing = 3, in de meegegeven chunks staan deze namen, maar ook meer namen.

Vraag 2:

1. tijd = 00:10
2. 5, klopt perfect
3. 5
4. 5

Vraag 3:

1. tijd = 00:07
2. 5, klopt perfect
3. 5
4. 5

Vraag 4:

1. tijd = 00:11
2. 2, bedrag genoemd van de later gehoorde getuige
3. 5
4. 4, in de bronverwijzing staat zowel dit bedrag als het bedrag van de eerst gehoorde getuige, wat mist is de chunk waarin staat wie de eerst gehoorde getuige is

Vraag 5:

1. tijd = 00:20
2. 4, heel goed tekst uit meerdere onderdelen van een verhoor gecombineerd om tot een juist antwoord te komen. Wel slechts van een persoon zonder erbij te vermelden van wie. Doorvraag werkt niet, de context bij een doorvraag blijft gelijk dus zegt dat alleen van deze persoon de voor en na situatie besproken is.
3. 5
4. 5, je kan heel mooi terug zien hoe het antwoord uit verschillende stukjes tekst komt.

Vraag 6:

1. tijd = 00:12
2. 3, De vraag zelf is na 1 gevonden maar onhandig opgeknipt waardoor het antwoord er niet bijstaat in dezelfde chunk en die het antwoord niet kon geven. Ipv zei alleen dat dit besproken is zonder details, op doorvraag naar voorbeelden gaf het goede voorbeelden van wetten en regelgeving maar niet de dingen waar ik naar zocht.
3. 4, afkortingen worden uitgelegd en van rapporten genoemd waar ze grofweg over gingen, maar jargon niet vermeden
4. 3, chunks zijn onhandig gekozen, maar er is wel overeenstemming tussen antwoord en gekozen chunks

vraag 7:

1. tijd 00:14
2. 5, noemde direct kabinetsbeleid en wens vanuit gemeenten, na 1 doorvraag ook extra context over waarom gemeenten dit verder nodig hadden
3. 5, jargon uitgelegd
4. 4, irrelevante bron bovenaan, belangrijkste bron als 2e, wel juiste bron benoemd

vraag 8:

1. tijd: 00:09
2. 3, wat er staat komt overeen met wat Sabir gezegd heeft, maar niet met wat de man van het CPB gezegd had
3. 4, dingen als risico-indicatoren en loonsomverhoging genoemd zonder uit te leggen, wel vervolgvragen erover goed beantwoord en daarin uitgelegd
4. 5

Vraag 9:

1. tijd: 00:09
2. 2, wel genoemd dat ze dit niet precies wisten maar niet waarom niet en ook niet de toevoeging dat er wel ramingen waren van opbrengsten van fraudejacht maar dat verder nooit gemeten is. Geen antwoord op vervolgvragen
3. 5
4. 5

Vraag 10:

1. tijd: 00:17
2. 4, hele goede context, wijzen op bulgarenfraude etc, enige fout: contrast wordt genoemd dat er vooral op incidenten gefocust wordt ipv **noodzaak tot sanctieverzwaring**
3. 4, medialogica als term genoemd zonder uit te leggen maar wel kort te beschrijven, goed uitgelegd in een vervolgvraag
4. 5

Vraag 11:

1. tijd: 00:12
2. 3, zegt wel spijt te hebben, maar van niet verder te gaan met onderzoek naar de toeslagenaffaire. Bij doorvraag blijkt de AI hiermee te duiden op dat ze na een artikel gepubliceerd te hebben over onderzoek van de ombudsman ze hier niet op zijn doorgegaan, wat het belangrijke punt is, maar de ombudsman wordt niet genoemd. Wordt gedaan alsof ze zelf dat onderzoek hebben gedaan en in het eerste bericht staat de relevante publicatie ook niet duidelijk vermeld. Daarnaast geen uitleg over waarom ze hier spijt van hebben, en geen uitleg van vanwaar de spijt.
3. 4, het blijft vaag maar de woorden zijn wel begrijpelijk
4. 2, De belangrijkste bron wordt als laatst genoemd en de meest relevante bronnen zitten niet in de context.

Settings:

System prompt: Geen,

Model: **GPT 4**

Embedding model: text-embedding-ada-002

Chunk Size: 1024

Chunk Overlap: 128

Aantal Chunks = 4

Vraag 1:

1. tijd - 00:16
2. 4, er zijn 6 mensen genoemd die allemaal verhoord zijn, maar dat is onvolledig; er zijn 9 mensen in totaal verhoord
3. 5
4. 5



Voorbeeld van hoe deze vraag relatief goed beantwoord kan worden, model maakt scheiding tussen ondervraagde en ondervragers en haalt zoveel mensen uit de tekst als er chunks zijn

Vraag 2:

1. tijd = 00:22
2. 5
3. 5
4. 5

Vraag 3:

1. tijd = 00:10
2. 5
3. 5
4. 5

Vraag 4:

1. tijd = 00:05
2. 5
3. 5
4. 4, geen bron waarin staat dat Goncalves de eerst gehoorde getuige is.

Vraag 5:

1. tijd = 00:31
2. 4, een voorbeeld weer
3. 5
4. 5

Vraag 6:

1. tijd = 00:44
2. 4
3. 3, soms springt het een beetje van de hak op de tak
4. 4

Vraag 7:

1. tijd = 00:14
2. 4, wordt alleen perspectief van Sabir gegeven (uitvoering kabinetsbeleid, maar niet waarom dat kabinetsbeleid was bv), maar bij doorvraag wordt dit verder uitgediept
3. 3, noemt woord Divosa zonder uitleg en legt het ook na vervolgvraag niet uit
4. 5

Vraag 8:

1. tijd = 00:25
2. 3, weederom alleen Sabirs perspectief, dus geen integratie van verschillende kanten
3. 2, zeer vaag verhaal, WBP wordt niet uitgelegd, niets gezegd over waarom risico-indicatoren gebruikt moesten worden. Aan het eind zeggen ze dat hun mening niet vermeld is terwijl het daarvoor steeds over hun mening gaat
4. 5

Vraag 9:

1. tijd = 00:28
2. 4, bedragen zijn genoemd en welke informatie Rutte wel en niet had, maar niet waar die bedragen vandaan kwamen
3. 3, beetje onsamenhangend verhaal, soms zijn stukken uit verschillende contexten achter elkaar geplakt zonder er een logisch geheel van te maken
4. 5, bronvermelding klopt

Vraag 11:

1. tijd = 00:13
2. 2, zegt dat hij nergens spijt van heeft
3. 4
4. 3

**Experiment 2:**

Settings:

System prompt: Geen,

Model: GPT 3.5

Embedding model: text-embedding-ada-002

Chunk Size: **512**

Chunk Overlap: 128

Aantal Chunks = **8**

Vraag 1:

a. tijd = 00:13

b. 3, er staan stukken tekst met namen, en uit sommige kan je ook opmaken wie verhoord is maar niet uit alle

c. 3, Simons wordt als verhoord gezien

d. 5

Vraag 2:

1. tijd = 00:25
2. 5
3. 5
4. 5

3.

1. tijd = 00:09
2. 5
3. 5
4. 5

4.

1. tijd = 00:56
2. 3, wel chunks waar de verschillende bedragen instaan geen chunks met wie de eerst gehoorde getuige is
3. 4, het bedrag klopt maar staat niets in over wie die eerst gehoorde getuige is
4. 5

5.

1. tijd = 00:52
2. 3, alleen stukjes tekst van Goncalves, met niet de letterlijke tekst van voor en na erbij maar wel genoeg om heel wat uit op te maken
3. 1, geen specificatie van om welke gedupeerde het gaat, ontbreekt crucicale informatie die wel in de context staat dingen lopen door elkaar heen. Dit zorgt voor een feitelijk onjuist verhaal waarin het lijkt alsof door A ook B is gekomen terwijl A en B verschillende personen troffen
4. 4, naam wordt niet genoemd

6.

1. tijd = 00:40
2. 3, In veel tekst staan relevante ontwikkelingen maar soms ook alleen de vraag en niet het antwoord. De vraag waarin dit letterlijk gesteld wordt ontbreekt
3. 5, kijkende naar de gegeven context zijn daar belangrijke ontwikkelingen uitgehaald
4. 3, Er worden allerlei wetten en rapporten genoemd zonder extra uitleg waar die over gaan, ook bij doorvraag niet.

7.

1. tijd = 00:20
2. 4, de kern staat er maar nadere uitleg over wat divosa zijn of context bij kabinetsbeleid ontbreekt
3. 5
4. 3, divosa wordt genoemd en niet uitgelegd

8.

1. tijd = 00:21
2. 3, wel uitleg van Sabir, niet van de man van het CBP zelf
3. 5
4. 4, risico-indicatoren worden genoemd en bij verdere doorvraag uitgelegd

9.

1. tijd = 00:32
2. 4, mooie stukjes tekst met bedragen, alleen geen chunks over waar die bedragen vandaan kwamen en hoe die (niet) gemeten werden
3. 4, klopt met tekst (geen wetenschappelijk bewijs, wel steekproeven) alleen bedrag dat voor fraudebestrijding voorspeld was staat er niet in en wel in de chunks.
4. 5

10.

1. tijd = 00:19
2. 5
3. 5
4. 5

11.

1. tijd = 00:04
2. 2, relevante tekst niet meegegeven
3. 3, zou moeten zijn dat dat uit de gegeven context niet goed op te maken is, zegt dat de journalist nergens spijt van heeft
4. 5

Settings:

System prompt: Geen,

Model: GPT 3.5

Embedding model: text-embedding-ada-002

Chunk Size: **1024**

Chunk Overlap: 128

Aantal Chunks = **4**

1. a. tijd = 00:06

b. 3, er staan namen in en er is uit op te maken wie vragen gesteld wordt en wie vragen stelt maar erg goed zijn de chunks niet gekozen

c. 1, het weet niet te scheiden wie er wel en niet verhoord worden

d. 5

1. a. tijd = 00:27

b. 5

c. 5

d. 5

1. a tijd = 00:11

b. 5

c. 5

d. 5

1. a. tijd = 00:04

b. 3, zoals altijd wel bedragen niet wie als eerst verhoord was

c. 3

d. 5

1. a. tijd = 00:23

b. 4, juiste context gevonden bij een van de gedupeerde

c. 5

d. 3, noemt de naam van de gedupeerde er niet bij, bij doorvraag slaat die over in het Engels

1. tijd = 00:17

b. 3, wel de vraag gevonden niet het antwoord, ook andere relevante tekst

c. 5

d. 5

1. tijd = 00:12

b. 5, geeft stuk tekst waarin het er direct over gaat maar ook tekst van Rutte over kabinetsbeleid en tekst over verzoeken UWV en SVB

c. 4, geeft het antwoord, maar niet de extra context

d. 5, legt Divosa direct uit

1. tijd = 00:11

b. 4, geeft alleen perspectief van Sabir

c. 5

d. 3, liep vast toen ik doorvroeg wat risico-indicatoren zijn

1. tijd = 00:09

b. 4, heeft belangrijkste info inclusief bedragen maar niet tekst waarin staat waarop die bedragen gebaseerd waren

c. 4, main message staat erin, prognoses voor wat er met fraudebeleid opgehaald zou worden niet

d. 5

1. tijd = 00:11

b. 5

c. 5

d. 5

1. tijd = 00:08

b. 2

c. n.v.t.

d. 5

Settings:

System prompt: Geen,

Model: GPT 3.5

Embedding model: text-embedding-ada-002

Chunk Size: **2048**

Chunk Overlap: 256

Aantal Chunks = **2**

1.

1. tijd = 00:06
2. 4, jammer dat hij 2 stukken uit dezelfde tekst heeft genomen, maar meer kan die ook niet echt
3. 1, er staat dat Aartsen en Azarkan het verhoor afnemen, wat impliceert dat ze zelf niet verhoord zijn
4. 5

2.

1. tijd = 00:22
2. 4, eerste chunk zeer relevant (bevat antwoord), helaas geen tweede chunk over dienstverlening
3. 5
4. 5

3.

1. tijd = 00:12
2. 5
3. 5
4. 5

4.

1. tijd = 00:04
2. 4, wel hoeveel Goncalez moest betalen, geen specifieke tekst waaruit blijkt dat ze de eerst gehoorde getuige is
3. 5
4. 5

5.

1. tijd = 00:06
2. 2
3. n.v.t.
4. 5

6.

1. tijd = 00:36
2. 4, relevante teksten gevonden maar belangrijke stukken ontbreken ook
3. 4, kern staat erin, maar media ontwikkelingen niet die wel in de context staan
4. 4, noemt wet SUWI zonder directe uitleg maar legt dit uitgebreid uit in vervolgvraag

7.

1. tijd = 00:34
2. 4, eerste document zeer relevant tweede zegt helaas niets over kabinetsbeleid
3. 5
4. 4

8.

1. tijd = 00:19
2. 3
3. 5
4. 5

9.

1. tijd = 00:43
2. 5
3. 4, noemt bedrag niet
4. 5

10.

1. tijd = 0:08
2. 2
3. n.v.t.
4. 5

11.

1. tijd = 0:04
2. 2
3. n.v.t.
4. 5

Settings:

System prompt: Geen,

Model: **GPT 3.5 - 16K context window**

Embedding model: text-embedding-ada-002

Chunk Size: **2048**

Chunk Overlap: 256

Aantal Chunks = **8**

1. a. tijd = 00:09

b. 3, geeft somehow allemaal chunks mee uit hetzelfde verslag

c. 4, heeft correct eruit gehaald wie verhoord zijn en wie verhoren, maar 2 mensen gemist wat wel uit de chunks op te maken is

d. 5

1. a. tijd = 00:18

b. 5

c. 4, mist rol van media en politiek terwijl dat wel in de context staat

d. 5

1. a. tijd = 00:08

b. 5

c. 5

d. 5

1. tijd = 00:07

b. 3, geeft niet stuk tekst mee waaruit blijkt dat ze als eerst verhoord is maar bij doorvraag wel

c. 3

d. 5

1. a. tijd = 00:18

b. 4, alleen over Deceunicks juiste tekst meegegeven

c. 5

d. 5

1. a. tijd = 01:14

b. 5

c. 5, beste antwoord dat ik op deze vraag gezien heb

d. 5

1. a. tijd = 00:36

b. 3, tekst waar de vraag direct instaat niet bijgegeven

c. 5, gegeven de context prachtig stukjes aan elkaar verbonden

d. 5

1. a. tijd = 00:23

b. 5, voor het eerst perspectief van CBP zelf goed verwoord, en vervolgadvies om wet niet in te dienen

c. 4, stuk waarin staat dat CBP eerst ook zag dat dit noodzakelijk was niet genoemd, alleen het negatieve advies

d. 5

1. a. tijd = 00:21

b. 5

c. 5

d. 3, beetje verwarrend verhaal (maar wel kloppend) waarin verwezen wordt naar niet genoemde bedragen, en ramingen zonder duidelijke verwijzing waar die ramingen over gingen

1. a. tijd = 00:27

b. 5, er staat ontzettend veel in de context over hoe de media met fraude omging

c. 4, goede dingen genoemd, zoals focus op zaken met specifieke herkomstlanden, sommige punten, zoals focus op conflict, ook gelaten.

d. 5

1. a. tijd = 00:08

b. 4, rapport ombudsman in 1 context genoemd en je kan redelijkerwijs opmaken dat hij had gewild dat ze daar meer mee hadden gedaan maar echt evident is het niet zoals dat in andere delen van de tekst wel is. Wel tekst waarin staat dat ze te weinig stilgestaan hebben bij de effecten van berichtgeving

c. 2

d. 5

**Experiment 3:**

Settings:

System prompt: Geen,

Model: GPT 3.5

Embedding model: **jegormeister/bert-base-dutch-cased**

Chunk Size: 1024

Chunk Overlap: 128

Aantal Chunks = 4

Vraag 1:

1. tijd = 00:09
2. 3
3. 5
4. 5

Vraag 2:

1. tijd = 00:19
2. 5
3. 3, 30 jaar en rol media en politiek ontbreekt terwijl dat wel in de context staat
4. 5

Vraag 3:

1. tijd = 00:05
2. 5
3. 5
4. 5

Vraag 4:

1. tijd = 00:07
2. 3
3. 3
4. 5

Vraag 5:

1. tijd = 00:13
2. 3
3. 5
4. 5

vraag 6:

1. tijd = 00:37
2. 5
3. 5
4. 5

vraag 7:

1. tijd = 00:06
2. 3
3. 2
4. 5

Vraag 8:

1. tijd = 00:23
2. 4
3. 5
4. 5

vraag 9:

1. tijd = 00:14
2. 4
3. 5
4. 5

vraag 10:

1. tijd = 00:26
2. 5
3. 4
4. 5

vraag 11:

1. tijd = 00:17
2. 5
3. 5
4. 5

Eerste die deze vraag goed heeft!!!

1. Voor meer uitleg over wat RAG is zie: <https://towardsdatascience.com/rag-vs-finetuning-which-is-the-best-tool-to-boost-your-llm-application-94654b1eaba7> [↑](#footnote-ref-0)
2. Zie [pbl-nl/appl-docchat: Chat with your own documents pressure cooker (github.com)](https://github.com/pbl-nl/appl-docchat) [↑](#footnote-ref-1)
3. Zie: [Verhoorweek 1 | Tweede Kamer der Staten-Generaal](https://www.tweedekamer.nl/kamerleden-en-commissies/commissies/parlementaire-enquetecommissie-fraudebeleid-en-dienstverlening/openbare-verhoren/verhoorweek-1) [↑](#footnote-ref-2)
4. *Note: Als bij vraag 2 een score van 1 of 2 behaald is betekent dit dat de vraag niet beantwoord kan worden met de beschikbare context, als het model vervolgens eerlijk zegt dat het de vraag niet kan beantwoorden krijgt het een n.v.t. score, waarbij ik dat niet meereken met de gemiddelde score voor het model. Eerst gaf ik het dan een 5 maar dan krijg je dat modellen heel goed lijken te functioneren zodra het chunken niet goed gaat. Ik geef het model wel alsnog een 1 als het aan het bluffen is: een antwoord bij elkaar verzint.* [↑](#footnote-ref-3)