**Eindverslag Jitse Goutbeek 25-4-2024**

**Inleiding**

Voortkomend uit de opdracht 'AI en spraakbesturing' (urencode 201202044.002), geaccordeerd door het SPB en gebaseerd op de strategische Service Roadmap van SSC-ICT zijn wij[[1]](#footnote-0) als innovatieteam van SSC-ICT de afgelopen maanden bezig geweest met onderzoek naar generatieve AI. Wat kan het ons bieden? Welke potentie zien we voor SSC-ICT? Hoe verhoudt dit zich tot de rest van het kabinetsbeleid?

Generatieve AI is een breed onderwerp. Om onszelf af te bakenen hebben we eerst gezocht naar een vorm van generatieve AI waar we veel potentie in zagen voor SSC-ICT. Deze hebben we gevonden in 'retrieval augmented generation’ (RAG). Dit komt dit neer op een taalmodel[[2]](#footnote-1) verbinden aan je eigen documenten zodat als je een vraag stelt de informatie uit de eigen documenten gebruikt kan worden om hem te beantwoorden. In dit [Visiestuk](https://docs.google.com/document/d/1J6k0HdlY5NMrOtSI0UhMYBZSoIlaqGwMx-FXzhio3Vk/edit?usp=sharing) staat uitgebreid uitgelegd waarom de keuze hierop is gevallen, maar de korte versie is dat de overheid veel documenten heeft en veel ambtenaren de informatie daaruit graag beter willen gebruiken. RAG kan daarbij helpen.

In 2023 waren we, samen met onze zeer gewaardeerde collega Jelle van Elburg, ook al met dit project bezig en aan het eind van dat jaar is een rapportage geschreven over wat er in 2023 gedaan is [Eindrapportage genAI 2023 v0.2](https://docs.google.com/document/d/1t8toY01jxR7LQdoAXlBOMoEdQ5GiYduCf4E3Xj9le3Q/edit). Sindsdien zijn een aantal grote zaken veranderd, beleidstechnisch is er meer duidelijk geworden door bijvoorbeeld het verschijnen van de kabinetsvisie op generatieve AI. Het onderwerp heeft extra interesse binnen de overheid opgepikt en verschillende mensen hebben elkaar beter kunnen vinden, door nieuwe plekken waar kennis samenkomt. Daarnaast hebben we zelf nieuwe lessen geleerd en onderzoek gedaan.

Nu, 4 maanden later, is het ook tijd voor mij, Jitse Goutbeek, om door te stromen naar een volgende opdracht. In dit document beschrijf ik de stappen die we de afgelopen 4 maanden gezet hebben, en de belangrijkste inzichten die we daaruit gehaald hebben.

Dit doe ik in drie delen. In het eerste deel schets ik de context van ons onderzoek, hoe de applicatie werkt en wat er nog meer in de overheid en maatschappij te zien is aan ontwikkelingen. Dit is relevant om uit het tweede deel, waarin onze eigen resultaten besproken worden, de juiste conclusies te kunnen trekken. In het derde deel bespreek ik wat dit betekent voor wat er voor het innovatieteam van SSC-ICT allemaal op de agenda staat.

**Deel 1: context**

**De werking van de applicatie**

Alvorens we allerlei zaken omtrent RAG bespreken is het goed eerst iets uitgebreider toe te lichten hoe RAG eigenlijk werkt. Het basisidee van RAG is dat je (interne) documenten kan koppelen aan een groot taalmodel (een LLM). Grote taalmodellen zijn modellen zoals Chat GPT, Claude, LLama en Mistral die tekst kunnen genereren op basis van de input van de gebruiker.

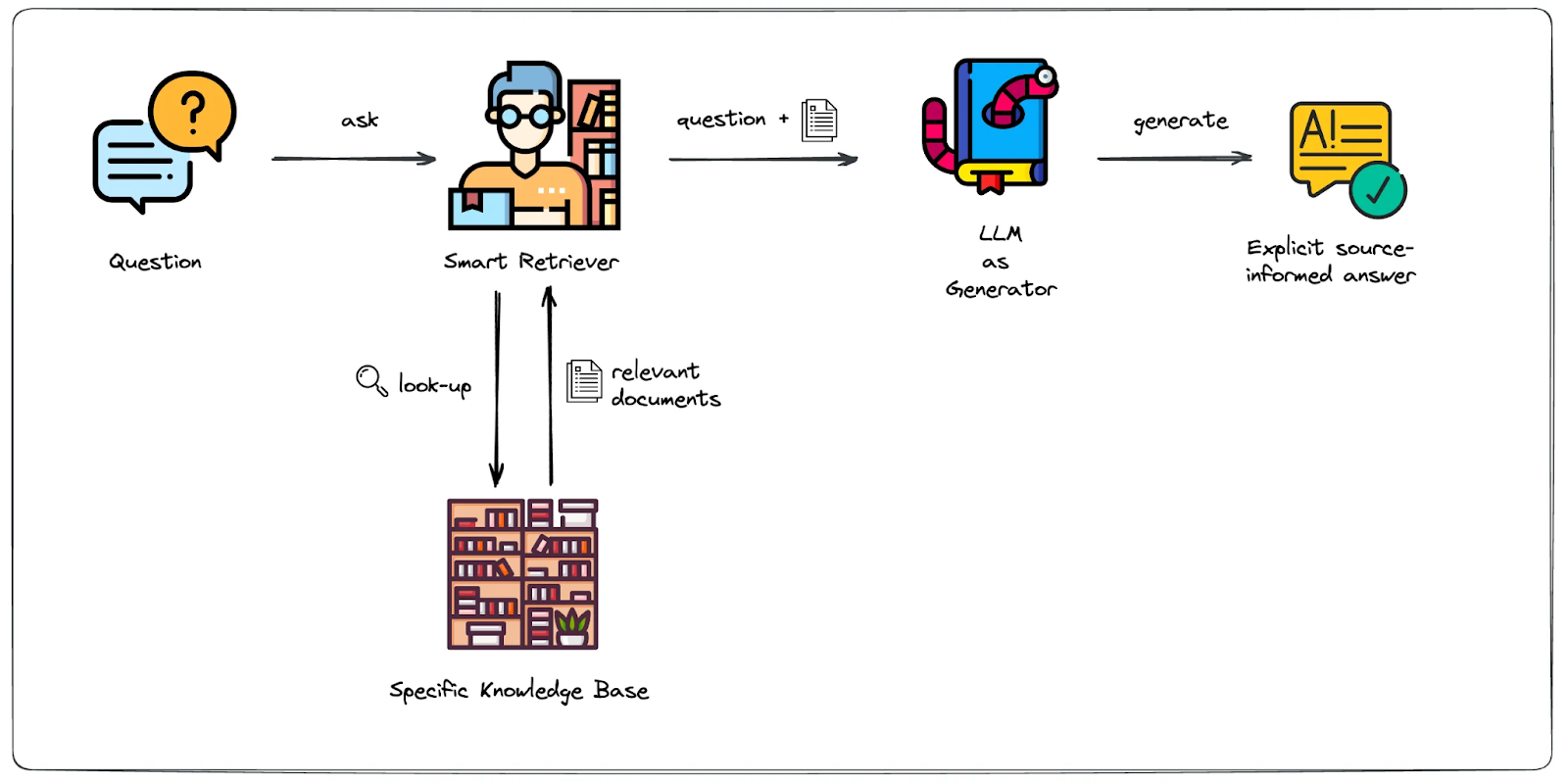
[Onze RAG-applicatie](https://github.com/SSC-ICT-Innovatie/LearningLion) werkt (grofweg) als volgt:

*Voor de gebruiker een vraag stelt:*

1. Je laadt de documenten die je wilt bevragen in, deze documenten worden opgeknipt in stukjes (chunks) van een grofweg vaste, van tevoren bepaalde lengte.
2. Deze chunks worden omgezet in een numerieke weergave (ookwel vector) door een embedding model, dit model is erop getraind om 2 stukken tekst met een vergelijkbare **betekenis** een vergelijkbare numerieke weergave te geven. Dit noemen we het embedded of vectorizer. Deze numerieke weergaven worden opgeslagen, de plek waar deze opgeslagen worden noemen we de vector store.

*Wanneer de gebruiker een vraag stelt:*

1. De vraag van de gebruiker wordt door hetzelfde embedding model omgezet tot numerieke weergave
2. De numerieke weergave van de vraag wordt vergeleken met de numerieke weergave van alle chunks via een similarity search. Vervolgens wordt er berekend welke chunks de laagste afstand tot de vraag hebben. Gezien het embedding model erop getraind is om 2 stukken tekst met een vergelijkbare betekenis een vergelijkbare numerieke weergave te geven, zou dit tot de chunks moeten leiden die qua betekenis dicht bij de vraag liggen en dus relevant zijn voor de vraag.
3. Deze chunks worden toegevoegd aan de input van de gebruiker, en aan het LLM wordt gevraagd de vraag van de gebruiker te beantwoorden op basis van de gegeven chunks. Deze chunks waar het antwoord op gebaseerd wordt krijgt de gebruiker ook te zien.



*Eenvoudige visuele weergave*

**Wat er in de rest van de overheid gebeurd**

Wij zijn echter niet de enige binnen de overheid die ons met dit soort technologie bezighouden. De interesse in generatieve AI in het algemeen en RAG in het specifiek is enorm. Veel organisaties hebben documenten die ze maar al te graag beter zouden willen doorzoeken, en doen daarom onderzoek of RAG daarvoor een geschikt middel zou zijn.

We zien initiatieven in deze richting bij onder andere het innovatielab van de kamer van koophandel, het rijksvastgoedbedrijf, datalab van infrastructuur & waterstaat, de belastingdienst, inspectie gezondheid en jeugd, het planbureau voor de leefomgeving, rijkswaterstaat en het datalab van J&V. Ik leef niet in de illusie dat dit een uitputtende lijst is, er zijn vast vele vergelijkbare initiatieven waarvan ik geen weet heb, maar het laat zien dat we lang niet de enige zijn met interesse in RAG. De meeste van deze initiatieven zijn er vooral op gericht te bekijken wat RAG voor hun organisatie kan betekenen. Er zijn wel verschillen tussen hoe deze initiatieven ingericht zijn, het datalab van J&V werkt bijvoorbeeld via een beschermde cloud omgeving terwijl wij in basis alles lokaal doen[[3]](#footnote-2).

Het bestaan van vele initiatieven betekent dat het risico op dubbel werk groot is. Om dat te mitigeren is het belangrijk contact met elkaar te zoeken, kennis, informatie en code te delen en vooral die dingen te doen die anderen nog niet doen. Voor dat eerste is er vanuit het CIO-rijk de community of practice rond generatieve AI opgericht. Deze community bestaat uit meerdere werkgroepen (publieke waarde & compliance, handreiking & standpunt, toepassing en als laatste techniek) waarin mensen die met generatieve AI bezig zijn elkaar kunnen opzoeken en kennis kunnen delen. Binnen SSC-ICT is een AI-gilde om iets vergelijkbaars te doen met een bredere thematische scope (AI in het algemeen, in plaats van specifiek generatieve AI) en een nauwere organisationele (puur SSC-ICT ipv overheidsbreed).

Wij delen actief onze code, onderzoeksresultaten en inzichten zodat andere niet hoeven te doen wat wij al gedaan hebben[[4]](#footnote-3). Tegelijk kiezen we er ook bewust voor bepaalde zaken niet te doen waar andere al mee bezig zijn. Zo maken we bij RAG gebruik van een taalmodel, en zijn wij zelf niet bezig met het trainen of fine-tunen van een dergelijk model. TNO, NFI en Surf zijn samen bezig een Nederlandstalig taalmodel te ontwikkelen waarvoor ze 13,5 miljoen euro van EZK krijgen[[5]](#footnote-4). Wij hebben niet de middelen dat ook te doen, en bovendien kunnen we onze applicatie zo inrichten dat we van hun model gebruik kunnen maken als het af is, dus er is weinig reden om zoiets te willen ondernemen. Voor het fine-tunen van een model is vooral veel goede data nodig, en in het Nederlands is dat maar beperkt beschikbaar. Gelukkig is het project leesplankje van het UWV bezig met het creëren van goede nederlandstalige Datasets, en heeft onderzoeker aan de KU Leuven Bram Vanroy daar ook meteen een model op gefinetuned, resultaat is [GEITJE-7B-ultra](https://huggingface.co/BramVanroy/GEITje-7B-ultra).

Wij hebben gekozen ons echt te focussen op het RAG deel zelf, en daarbij extra te letten op waar wij en anderen tegenaan lopen, en waar SSC-ICT meerwaarde zou kunnen leveren.

**De beleidscontext waarin we ons begeven**

Artificiële intelligentie brengt risico's met zich mee, denk aan vragen omtrent dataveiligheid en de betrouwbaarheid en eerlijkheid van de resultaten. Generatieve AI in het algemeen en RAG in het specifiek zijn geen uitzonderingen. Wat niet wegneemt dat experimenteren aangemoedigd wordt, en de risico's daarvan beperkt zijn.

Er is een [voorlopig standpunt](https://www.rijksoverheid.nl/documenten/kamerstukken/2023/12/11/kamerbrief-over-voorlopig-standpunt-voor-rijksorganisaties-bij-het-gebruik-van-generatieve-ai) gekomen over het gebruik van generatieve AI binnen de rijksoverheid. Hierin staat vermeld dat er eerst gekeken moet worden naar of gegevens op een verantwoorde en rechtmatige manier verwerkt worden, middels een [data protection impact assessment](https://www.autoriteitpersoonsgegevens.nl/themas/basis-avg/praktisch-avg/data-protection-impact-assessment-dpia). Bij 'niet gecontracteerde' generatieve AI toepassingen, zoals ChatGPT is dat doorgaans niet het geval, meldt de brief. Naast de DPIA moet er ook een algemene assessment gedaan worden over de risico’s, hier heb je verschillende instrumenten voor zoals de [Impact Assessment Mensenenrechten en Algoritmen](https://www.rijksoverheid.nl/documenten/rapporten/2021/02/25/impact-assessment-mensenrechten-en-algoritmes) ontwikkelt door BZK, de [AI impact assessment](https://ecp.nl/wp-content/uploads/2018/11/Artificial-Intelligence-Impact-Assesment.pdf) ontwikkelt door ECP en de [gelijknamige assessment van I&W.](https://www.rijksoverheid.nl/documenten/rapporten/2022/11/30/ai-impact-assessment-ministerie-van-infrastructuur-en-waterstaat)[[6]](#footnote-5)

Bij een RAG toepassing zijn er twee vormen van gegevensverwerking van toepassing, enerzijds zijn er de gegevens die ooit verwerkt zijn bij het trainen van het taalmodel. Hier kunnen persoonsgegevens of auteursrechtelijk beschermde gegevens in zitten, dit kan problematisch zijn maar juridisch is de ontwikkelaar van het gebruikte taalmodel hier verantwoordelijk voor[[7]](#footnote-6). Daarnaast speelt de verwerking van de interne documenten die gebruikt worden, die **niet** gebruikt worden om het model te trainen maar slechts doorzocht worden. Hier spelen vergelijkbare vragen als bij een zoekmachine, en het vooral belangrijk is dat de verkeerde mensen geen toegang krijgen tot gevoelige informatie[[8]](#footnote-7). Ook voor andere impact assessments kan er een onderscheid gemaakt worden tussen de risico’s van het gebruikte taalmodel, en de risico’s die ontstaan door het daarvan koppelen aan een eigen database. Binnenkort gaat SSC-ICT impact assessments afnemen voor specifieke open source taalmodellen, als er een model is waarbij er een positieve conclusie is kan je op dat moment aparte impact assessments uitvoeren voor het deel wat je er zelf aan toevoegt.

Hiernaast zijn er een aantal rapporten gekomen over de potentiële impact en risico’s van AI, zoals [dit rapport](https://www.rathenau.nl/sites/default/files/2023-12/Scan_Generatieve_AI_Rathenau_Instituut.pdf) van het Rathenau instituut. Deze inzichten hebben zich vertaald in een [strategisch actieplan](https://open.overheid.nl/documenten/ronl-e14cdcee-690c-4995-9870-fa4141319d6f/pdf) en [overheidsbrede visie](https://www.rijksoverheid.nl/documenten/rapporten/2024/01/01/overheidsbrede-visie-generatieve-ai). Wij hebben uitgebreid beschreven hoe RAG zich verhoud tot deze kabinetsvisie [RAG en de Kabinetsvisie](https://docs.google.com/document/d/1Uoob_kSkX-j1dRUB-xS3JCNJ2GpmZwj0ubQLTsC2jIM/edit?usp=sharing).

**Deel 2: RAG voor inkoop**

De eerste 4 maanden van dit onderzoek hadden ons een aantal inzichten gegeven. Door een RAG tool te gebruiken om vragen te stellen over de openbare verhoren van de parlementaire enquêtecommissie fraudebestrijding, zagen we dat je er best goede resultaten mee kunt bereiken. Zie hier ons onderzoek daarover: [Experimenteren met RAG](https://docs.google.com/document/d/1hDoKZ9ujWAn2x8eu2XZFXqhsg7YL6TqBNXtlaiK-hA8/edit?usp=sharing). We zagen ook dat het niet altijd makkelijk is om goed te meten hoe goed de resultaten precies waren, bij dit punt staan we later uitgebreider stil. Dit was echter gedaan met Chat-GPT, maar we hadden ook een applicatie gemaakt die een vergelijkbare taak volledig lokaal met open-source modellen uit kon voeren, zie [hier](https://github.com/SSC-ICT-Innovatie/LearningLion) de code. Het stellen van vragen ging echter heel langzaam omdat we alles op de eigen laptop aan het uitrekenen waren, en de beste taalmodellen vereisten meer rekenkracht dan onze laptop ter beschikking had.

De inkoopmedewerkers van I&W waren zo genereus gegevens en tijd beschikbaar te stellen zodat wij een specifieke casus hadden om onze RAG applicatie mee te testen. Dit is niet zo zeer met het idee deze casus ook echt te leveren, maar meer om te zien waar je allemaal tegenaan loopt in een proces met een concrete toepassing. Dusver zijn we al tegen een heel aantal dingen aangelopen, waar lessen uit te trekken zijn. Dit hoofdstuk bevat een beschrijving van wat we gedaan hebben en wat we tegen zijn gekomen sinds januari 2024.

**De kwaliteit die we hadden en de kwaliteit die mogelijk is**

Ons onderzoek vindt plaats onder de aanname dat we alle data lokaal willen houden. De vragen over wat wel en niet met welke mate van gegevensbescherming verwerkt mag worden speelt nog, en andere organisaties hanteren een liberale visie. Onze rol dusver is juist om te kijken wat er kan onder de aanname dat we alles open source en lokaal doen.

Een eerste stap hier was om de code gewoon op een eigen laptop (hp met intel core vpro i5) te runnen, en de resultaten waren voorspelbaar dramatisch. Het duurt minuten om een antwoord te krijgen, en dikwijls bestaat het antwoord dat uiteindelijk gegenereerd wordt uit 0 woorden, oftewel de applicatie denkt minuten na om geen antwoord te geven.

Het duurde me grofweg een kwartier om 5 vragen te stellen, waarbij die slechts op 2 een antwoord gaf. Zeker gezien we de kwaliteit onder veel verschillende omstandigheden willen vergelijken is dit onacceptabel. De applicatie zoals we die hadden op de beschikbare rekenkracht was dus ontoereikend voor de opgave die we mee hadden gekregen.

We begonnen daarom twee paden te bewandelen, enerzijds gingen we kijken of we meer rekenkracht konden organiseren, anderzijds of we alvast konden zien of er überhaupt potentie was door een vergelijkbare casus met openbare data via de cloud of een API key te bestuderen.

**Testen via API en Microsoft cloud**

In afwachting op het werkend krijgen van onze eigen applicatie had ik online een openbare aanbesteding gevonden (goud). Het leek me zinvol vragen uit die nota van inlichtingen alvast via API en cloud tools te stellen, om te kijken of er potentie te zien is en wellicht om een vergelijking mogelijk te maken. [RAG voor inkoop, resultaten en bedenkingen](https://docs.google.com/document/d/1nt97vANjlr4wENu6xhnleWMciVvZzgHLZp1sQwO8gb8/edit?usp=sharing).

We zagen dat zowel in de cloud als met de API het programma in staat was vragen terug te vinden en de antwoorden te reproduceren. Ook als je deze vragen herformuleert. Echter het stellen van vragen die later gesteld werden op basis van vragen die eerder gesteld waren bleek ingewikkeld, wat goed aan beperkingen van de data zelf zou kunnen liggen (misschien stonden de antwoorden er simpelweg niet in) dus lastig om daar meer conclusies uit te trekken.

Het lukte prima om RAG applicaties in de cloud (Azure) te runnen, iedere keer moet je weer allerlei functionaliteiten aanmaken en afbreken maar dat is in basis geen probleem. Mijn methode was voor iedere dag een omgeving aanmaken met de naam van die dag en na gebruik die omgeving en dus alle functionaliteiten daarin af te breken. Een dag bleek ik echter iets aangemaakt te hebben in mijn basisomgeving in plaats van de omgeving voor die dag en had ik dat dus per ongeluk niet afgebroken, in Azure is helaas geen makkelijke manier om te kijken of iets nog draait je ziet de kosten die je maakt namelijk pas na een paar dagen. Dit was vlak voor het weekend en toen ik na het weekend terugkwam zag ik het pas, tegen die tijd had het echter al honderden euro’s gekost. Dit is uiteraard volledig mijn fout, maar het laat wel zien dat het belangrijk is goed te weten wat je doet als je in de cloud werkt, anders kan het snel duur worden.

**Rekenkracht organiseren**

Voor extra rekenkracht hadden wij gelukkig toegang tot wat GPU’s, dat is rekenkracht die heel goed dingen parallel kan uitrekenen, wat voor AI toepassingen zeer nuttig is omdat er veel parallelle berekeningen gemaakt moeten worden. Helaas bleek het gebruik hiervan uitermate complex, en voor een groot deel buiten mijn eigen vaardigheden.

Als eerst ging mijn voormalige collega Jelle van Elburg ermee aan de slag, maar zowel hem als een aantal stagiairs lukte het niet om ze überhaupt aan de praat te krijgen. Daarna kwamen collega’s Yann Rosema en Cas Dekkers ons helpen, Yann was het gelukt alle GPU’s aan elkaar te verbinden en aan te sluiten op het internet. Cas was het na veel uitzoekwerk ook gelukt de GPU’s vanuit een laptop te kunnen aanspreken, via een programma dat slurm heet en de verschillende taken goed door de GPU’s kan laten uitvoeren. Toen we echter onze code gingen runnen bleek dat hij geen gebruik maakte van de GPU rekenkracht, daarvoor moesten we in de code instructies veranderen om wel gebruik te maken van de GPU's. Dit werkte echter ook niet, al snel bleek dat een tool die we gebruikte niet samenkwam met de GPU rekenkracht en we een andere versie daarvan moesten downloaden.Dat doen bleek veel ingewikkelder dan het klinkt.[[9]](#footnote-8) Na vele dagen dingen geprobeerd te hebben en geen steek verder gekomen te zijn moet ik helaas meedelen dat het tot op heden nog steeds niet gelukt is om met de Nvidia GPU’s te werken.

Ondertussen begonnen er ook andere oplossingen te komen voor dit probleem, er werd een laptop besteld die zelf al GPU’s bevat en mensen vanuit de BU infra kwamen naar ons toe dat ze misschien wel iets voor ons konden regelen. Er werden gesprekken gestart met HP en Fujitsu die wel bereid waren het een en ander uit te lenen, zodat wij konden kijken of we in onze toekomstige infrastructuur meer van hun hardware nodig hadden. Dit gaat uitgewerkt worden tot een proof of concept waarbij gekeken wordt of die hardware voor aanzienlijke verbeteringen zorgt in het experimenteren en tegen welke (energie) kosten. Helaas komt dat voor mij te laat, maar er worden wel mooie stappen in gezet.

**Herijken: toch testen met de applicatie die we hebben**

Aangezien het niet lukte meer rekenkracht te organiseren, moest ik toch maar gaan testen op de eigen laptop. Gelukkig kwam er een nieuwe stagiair bij ons werken genaamd Nicky die een grote verandering in onze code doorvoerde, wat voor een groot deel neerkwam op het kopieëren van de code gemaakt door Stefan Troost van het planbureau voor de leefomgeving en ervoor zorgen dat deze volledig lokaal te draaien was. In deze code zat de mogelijkheid automatisch een reeks aan vragen tegelijk te stellen en de antwoorden op te slaan in een apart bestand. Hoewel dit niet het stellen van een individuele vraag sneller maakte, betekende het wel dat het feit dat het proces van vragen stellen langzaam ging minder schadelijk was. Nu kon ik een paar uur lang de code runnen om een kleine steekproef van vragen te stellen, zeker niet ideaal maar beter dan iedere 10 minuten een losse vraag moeten stellen en hopen dat je een antwoord krijgt. Hoewel de kwaliteit van de antwoorden erg laag was, kreeg ik wel antwoorden, ik weet niet waarom in de transitie van de oude code naar de nieuwe code er minder vaak geen antwoorden uitkwamen.

In dit document: [RAG testen met inkoopmedwerkers I&W](https://docs.google.com/document/d/1CQdw5McQ04oOKoLAit5EjaxkJK5QuXz_n1w7UVZ9BYw/edit?usp=sharing) staan de resultaten van dit onderzoek. In het algemeen waren de resultaten niet geweldig, eigenlijk was in alle experimenten minder dan de helft van de antwoorden naar tevredenheid van de inkoopmedewerkers beantwoord. Wel was er spreiding afhankelijk van voor welke instellingen er precies gekozen was.

Een imperfecte tool als dienst leveren heeft risico’s, zo is er een vergelijkbare applicatie als waar wij aan maken in gebruik door de [politie van New York](https://apnews.com/article/new-york-city-chatbot-misinformation-6ebc71db5b770b9969c906a7ee4fae21). Deze applicatie geeft soms foutief aan dat iets legaal is, wat eigenlijk illegaal is. Een restauranteigenaar kan bijvoorbeeld vragen of het legaal is om een deel van de fooi van werknemers naar het restaurant te laten gaan, en onterecht horen dat dit mag. Wij zien met enige regelmaat dat niet alle relevante informatie gevonden wordt als we een vraag aan onze chatbot stellen, het kan goed zijn dat relevante informatie over het illegaal zijn van een bepaalde handeling niet gevonden wordt en daarom onterecht aangenomen wordt dat het mag. Om dit soort situaties te voorkomen moeten we goed zicht hebben op de volledige spreiding van resultaten die we kunnen krijgen, en wat voor risico dat met zich meebrengt. Wat we nu hebben is evident niet goed genoeg om tot dienst te leveren, gezien slechts een minderheid van de vragen fatsoenlijk beantwoord wordt. De gebruiker heeft alleen iets aan een tool als die antwoorden kan krijgen op vragen waar die zelf het antwoord niet op weet, in welk geval het lastig te onderscheiden is welke antwoorden wel en niet te vertrouwen zijn[[10]](#footnote-9) .

De vraag is echter of dit zo blijft. De ontwikkelingen op het gebied van generatieve AI gaan erg snel, zo is er na het einde van mijn onderzoek [Llama 3 uitgekomen](https://ai.meta.com/blog/meta-llama-3/). Een nieuw open source taalmodel dat mogelijk beter werkt dan de modellen die wij gebruikt hebben, bovendien zijn er ook betere modellen te gebruiken als er meer rekenkracht beschikbaar is. We zien ook dat de maximale hoeveelheid tekst in een vraag snel toeneemt, dit wordt de context window genoemd. Ik heb gewerkt met modellen met een context window van [4096 tokens](https://mistral.ai/assets/Mistral_7B_paper_v_0_1.pdf), wat heel grofweg gemiddeld 5 pagina’s is, genoeg om stukken tekst uit de context mee te geven maar je moet enigszins selectief zijn in wat je meegeeft. Er zijn al [modellen](https://www.anthropic.com/news/claude-3-family) uit die meer dan een miljoen tokens als context goed kunnen verwerken, wat grofweg 125.000 pagina’s is, dit betekent dat je misschien op een hele andere manier kan gaan zoeken in de toekomst - veel minder selectief. Momenteel kan veel context meegeven zorgen voor [slechtere prestaties](https://medium.com/@juanc.olamendy/lost-in-the-middle-a-deep-dive-into-rag-and-langchains-solution-3eccfbe65f49) maar de vraag is wat op termijn beter kan zoeken, het taalmodel zelf of de zoekfunctie die je uitvoert om de juiste voorselectie te maken.

Bovendien kunnen we zelf ook heel veel doen om de prestaties te verbeteren. Er zijn ontzettend veel mensen die met RAG werken en vertellen over technieken die ze toegepast hebben om hun prestaties te verbeteren, en die wij dus mogelijk over zouden kunnen nemen, daar heb ik dit overzicht van gemaakt: [Verbetermogelijkheden RAG](https://docs.google.com/document/d/12FnalqzluJwDR_EPZRVDQwPZTU0zeJFuWZOqTQSaR-M/edit?usp=sharing). Een van deze suggesties[[11]](#footnote-10) heeft Nicky geïmplementeerd, maar het was lastig vast te stellen of dit tot betere resultaten leidde. Bovendien zijn er heel veel instellingen om uit te kiezen, welk taalmodel gebruik je?, welk embedding model?, hoe zoek je naar de juiste chunks?, hoeveel chunks zoek je?, hoe lang maak je de chunks?, welke opdracht (system prompt) geef je mee? We hebben gezien dat deze instellingen optimaliseren tot betere prestaties kunnen leiden, maar ook dat het niet makkelijk is om te achterhalen welke combinatie van instellingen het beste is.

**Meten is weten?**

Het liefst zou je van iedere combinatie van instellingen willen meten hoe goed de prestaties zijn, en de instellingen kiezen waarbij de prestaties het beste zijn. Net zo zou je verschillende verbeteringen willen uitproberen en willen meten welke wezenlijk voor betere resultaten leiden en welke niet. In beide gevallen moet je vaak de kwaliteit kunnen meten, en dat is voor generatieve AI niet eenvoudig. Bij niet generatieve AI toepassingen is er vaak een trainingsset en een testset, waarbij gekeken kan worden hoe vaak de datapunten in de testset juist voorspeld worden[[12]](#footnote-11). Bij generatieve AI toepassingen is het aantal mogelijke outputs echter enorm groot, waarbij er niet één of een paar outputs correct zijn en iedere andere output incorrect. Als ik vraag wat de rol was van de media in het fraudebeleid zijn daar veel antwoorden op te geven, dezelfde boodschap kan immers op veel manieren geformuleerd worden en soms zijn er veel grijstinten in hoe goed een antwoord is. Bovendien is de variatie in wat voor vragen de gebruiker kan stellen vrijwel oneindig, wat het heel lastig maakt om met vertrouwen vast te stellen dat de dingen die je getest hebt dekkend zijn voor de vragen waar gebruikers mee komen.

Als samenleving zijn we daarom nog zoekende naar hoe we het beste de kwaliteit en risico’s van generatieve AI systemen kunnen bepalen voor we ze in productie nemen. Het [AI index report 2024](https://aiindex.stanford.edu/wp-content/uploads/2024/04/HAI_AI-Index-Report-2024.pdf), een zeer uitvoerige analyse over AI van de universiteit van Stanford, benoemt het ontbreken van goede meetinstrumenten als een van hun belangrijkste observaties. Ook kranten als de [New York Times](https://www.nytimes.com/2024/04/15/technology/ai-models-measurement.html?unlocked_article_code=1.kk0.oMBm.397ORl_r-8Tq&smid=url-share) schrijven hierover en zelfs [vooraanstaande AI labs](https://www.anthropic.com/news/evaluating-ai-systems#entry:148292@1:url) zeggen zelf twijfels te hebben over de manieren waarop ze AI systemen evalueren. We zien binnen de community of practice ook veel mensen hiermee worstelen, en ik ben bij een hackathon generatieve AI van Rijkswaterstaat geweest, bedoeld om zich met dit probleem bezig te houden.

Dit is echter geen hopeloos probleem, voor een deel zijn onze validatie-instrumenten slecht omdat de technologie nieuw is en we nog maar net begonnen zijn met nadenken over hoe we het moeten evalueren. In zowel de overheid, vanuit bijvoorbeeld de Hackathon en de community of practice, wetenschap[[13]](#footnote-12) en markt[[14]](#footnote-13) ontstaan nieuwe ideeën voor evaluatiemethoden. Dit betekent dat we moeten kijken welke evaluatiemethoden voor onze doeleinden werken, willen we automatische evaluaties of met gebruikers in gesprek? Hoe zorgen we ervoor dat we dat op een manier doen die reproduceerbaar is?

Hier komt bovenop dat generatieve AI toepassingen niet altijd hetzelfde antwoord geven, dus de steekproef moet redelijk groot zijn om met vertrouwen conclusies te trekken. Hier komt bovenop dat er een groot aantal permutaties en verschillende zaken waarop je een AI systeem kan willen beoordelen, bijvoorbeeld of de antwoorden begrijpelijk zijn, accuraat, behulpzaam, geen bias bevatten enzovoort. Dit betekent dat voor welke evaluatiemethoden je ook kiest, je een groot aantal vragen moet stellen en antwoorden moet beoordelen. Hoe snel we tot meer kennis komen, een beter beeld hebben van hoe verschillende evaluatiemethodes werken en welke instellingen of aanpassingen tot ‘betere’ resultaten leiden, hangt af van hoe snel we vragen kunnen stellen en beoordelen. Gedurende mijn tijd hier ging dat erg langzaam, waardoor ik vaak genoodzaakt was voor kleine steekproeven te gaan. Sneller vragen kunnen stellen zou daarmee niet alleen helpen voor de kwaliteit van de applicatie, maar ook voor het sneller kunnen onderzoeken en navigeren van alle onzekerheden.

**Deel 3: hoe verder**

**Rekenkracht en consultancy leveren als dienst**

Vanuit de observaties die in dit stuk beschreven zijn heb ik een [Advies](https://docs.google.com/document/d/1rfkRz5cdEV5_JWlzTf3NQbUPGc5rSsDoUXF_jPG2sPY/edit?usp=sharing) geschreven aan SSC-ICT. Daarin beschrijf ik hoe belangrijk meer rekenkracht is voor sneller en beter onderzoek kunnen doen, en dat veel klanten van SSC-ICT tegen ditzelfde probleem aanlopen en ook aangeven dat ze graag hulp van SSC-ICT hierbij zouden hebben.

De komende tijd gaat SSC-ICT hier dan ook mee aan de slag, en het innovatieteam gaat daarbij helpen. Vanuit de business unit infrastructuur zijn contacten gelegd met grote leveranciers als HP en Fuijitsu, die ons hardware in bruikleen kunnen geven om mee te experimenteren. Collega’s van ons gaan een experimenteeromgeving inrichten met docker containers en kijken of het ons lukt die te gebruiken voor generatieve AI doeleinden en wat daar dan allemaal bij komt kijken (ook op het gebied van energie en kosten).

Dat er rijksbreed zoveel potentie gezien wordt in generatieve AI, kan SSC-ICT niet onopgemerkt voorbij laten gaan. Klanten hebben zelf vaak hun eigen datalabs of innovatieteams en kennen hun eigen vraagstukken het beste, het lijkt me voor SSC-ICT aantrekkelijk daar gebruik van te maken en ze hun eigen toepassingen te laten onderzoeken. Wel zal niet iedere klant hetzelfde startpunt hebben, bij het Datalab van J&V zijn ze verder met hun generatieve AI onderzoek dan wij zijn. Die hebben vooral toegang tot bijvoorbeeld een experimenteeromgeving nodig. Er zullen ook partijen zijn waarbij we dubbel werk kunnen voorkomen door ze een stukje consultancy te bieden, mee te denken over welke code ze kunnen gebruiken voor hun toepassing en te laten zien hoe ze van bijvoorbeeld een experimenteeromgeving gebruik kunnen maken. In sommige gevallen willen klanten misschien zelfs het onderzoek helemaal aan ons uitbesteden.

Mochten meerdere onderzoeken die grofweg dezelfde technologie gebruiken[[15]](#footnote-14) een positief resultaat bereiken, dan is het misschien niet praktisch voor iedere partij zelf alles te ontwikkelen en in te richten. Als gedeelde dienstverlener is SSC-ICT een logische organisatie om daar dan samen met de mensen achter die onderzoeken een algemene beheerde dienstverlening van te maken.

**Build or buy**

Op basis van ons onderzoek denken we dat het belangrijk is zelf aan de slag te blijven met het bouwen en testen van applicaties, en denk ik dat dit doorgaans de voorkeur geniet boven het inkopen van generatieve AI applicaties. We zien grote verschillen tussen de beste resultaten, die je vaak in demo’s ziet, en de slechtste resultaten, die het meest bepalend zijn voor de risico’s van het in gebruik nemen van een applicatie. Daarnaast zijn we op dit moment nog vrij slecht in het beoordelen van de kwaliteit van applicaties, en er is geen reden om te geloven dat dat automatisch beter wordt als er nieuwe marktoplossingen komen. Dit betekent dat we op zijn minst de kennis binnen de eigen organisatie moeten opbouwen om goed te beoordelen of marktoplossingen passend zijn, en niet afhankelijk te zijn van demo’s en verkooppraatjes. Hiernaast verandert en verbetert de technologie heel snel, wat betekent dat de flexibiliteit om snel van de nieuwste techniek gebruik te kunnen maken cruciaal is. Dit vereist kennis binnen de eigen organisatie om te beoordelen welke ontwikkelingen relevant zijn en wat we ermee moeten, bovendien betekent het dat als je iets inkoopt je heel erg op moet passen dat het niet snel verouderd. Een terugkerend thema hierin is dat in een snel veranderende wereld het hebben van de juiste kennis die je opbouwt door iets te bouwen vaak belangrijker is dan de applicatie zelf. Ons onderzoek is daar een duidelijk voorbeeld van, de RAG applicatie zelf is niet goed genoeg om grote meerwaarde voor SSC-ICT te hebben, maar alle kennis die opgebouwd is tijdens het maken van deze applicatie heeft die meerwaarde wel. Hierbij is wederom de vraag in hoeverre je evenveel kennis opbouwt als je het bouwen van de applicaties uitbesteed.

**De komende maanden voor het innovatieteam van SSC-ICT**

Wij blijven we doorgaan met het inkoop vraagstuk van I&W omdat het ons een heldere manier geeft om prestaties te bekijken. Dit vanuit het perspectief dat er waarschijnlijk een punt komt dat generatieve AI toepassingen mogelijk zijn met meerwaarde voor SSC-ICT, en we genoeg kennis op moeten bouwen om dat punt te identificeren en de voorwaarden te scheppen om er ook gebruik van te maken. Bij ons onderzoek is speciale aandacht vereist voor de manier waarop we resultaten valideren en ook met de meetmethoden zelf zal geëxperimenteerd worden. Martha Romkes gaat als trainee aan de slag bij het innovatieteam en er liggen nog genoeg uitdagingen klaar. Bovendien zullen er ook weer nieuwe modellen uitkomen en nieuwe inzichten ontstaan die vragen om een flexibele houding en om wegen in te slaan waar we nu nog niet aan gedacht hebben.

Prioriteit hierbij blijft dat we kennis opdoen en delen. Het doet ons deugd dat we horen van mensen dat ze gebruikmaken van onze GitHub en er meerdere partijen zijn geweest die contact met ons opgenomen hebben omdat ze met dezelfde vragen worstelde als waar wij mee bezig zijn geweest. Het innovatieteam van SSC-ICT, de overheid en de maatschappij zijn allemaal nog lang niet uitgeleerd als het op generatieve AI aankomt.

1. Als gesproken wordt over ‘wij’ gaat het over het innovatieteam van SSC-ICT, als er gesproken wordt over ‘ik’ gaat dat over Jitse Goutbeek [↑](#footnote-ref-0)
2. ChatGPT is een voorbeeld van een taalmodel, maar wij hebben met andere taalmodellen gewerkt. Specifiek taalmodellen die openbaar te downloaden zijn en dus op de eigen computer te gebruiken, dit wordt ook wel open source genoemd. [↑](#footnote-ref-1)
3. Op dit moment is het lastig te zeggen wat ‘beter’ is en is het bestaan van verschillen dus vooral gunstig omdat het helpt de opties beter te vergelijken als er ervaring met verschillende keuzes is. [↑](#footnote-ref-2)
4. Zie: https://github.com/SSC-ICT-Innovatie/LearningLion [↑](#footnote-ref-3)
5. Dit project heet [GPT-NL](https://www.tno.nl/nl/digitaal/digitale-innovatie/data-sharing/generative-ai/gpt-nl-versterkt-nederlandse-autonomie/) en RAG is een van de toepassingen waarvoor de makers dit model gaan proberen te optimaliseren. [↑](#footnote-ref-4)
6. [Voor onze toepassing zou ik de laatste aanraden omdat de vragen concreet zijn en het meest eenduidig voor onze toepassing.](https://www.rijksoverheid.nl/documenten/rapporten/2022/11/30/ai-impact-assessment-ministerie-van-infrastructuur-en-waterstaat) In de IAMA staan veel vragen over welke voorspellingen er gemaakt worden en welke keuzes daarop gebaseerd worden, en die passen makkelijker op een klassiek algoritme dan een generatieve toepassing zoals deze. In de tool van I&W gaat het meer over betrouwbaarheid en uitlegbaarheid, wat logischere vragen zijn om te stellen over een RAG applicatie. [↑](#footnote-ref-5)
7. Op dit moment is de rechter nog niet in staat geweest hierover te oordelen maar er lopen rechtzaken over. Daarnaast is er een ietswat technische vraag of het gebruik van modellen dat op persoonsgegevens getraind is zelf een inbreuk op privacy kan zijn, doordat die gegevens impliciet verankerd zijn in de gewichten van het model en dus ook de gegenereerde tekst bij gebruik. Hier is de relevante vraag of je met het model de persoonsgegevens waarop het ooit getraind is weer ‘terug’ naar boven kan halen, en bij RAG toepassingen met modellen met miljarden parameters is dit onwaarschijnlijk maar dit soort onderzoeken lopen ook nog. [↑](#footnote-ref-6)
8. Er is ook een security analyse gemaakt van onze code: [Beveiligingsrapport LearningLion](https://docs.google.com/document/d/1dU5LgZ7owJijDgZ_Oj08WkfnIZ9_Z6CPKdxYNWVsjRI/edit?usp=sharing) [↑](#footnote-ref-7)
9. Steeds opnieuw hadden we het gedownload op een manier die compatibel zou moeten zijn, maar bleek het weer niet samen te kunnen werken. We kwamen erachter dat we niet de juiste toestemmingen hadden om de verschillende delen met elkaar samen te laten werken, het programma slurm bleek zelf nooit de rechten te hebben dingen te installeren maar omdat we in slurm moesten werken om de berekeningen op een goede manier te verdelen had slurm wel toegang nodig tot programma’s die erop geïnstalleerd moesten worden. Uiteindelijk hebben we een manier gevonden om slurm de juiste toestemming te geven. Er bleven echter problemen ontstaan met het installeren van de juiste pakketten. [↑](#footnote-ref-8)
10. De bronvermelding helpt hier iets bij, maar soms kan onvolledige bronvermelding mensen op het verkeerde been zetten en het kan vrij veel werk zijn de bronvermelding volledig door te nemen waarbij we er niet automatisch vanuit kunnen gaan dat iedere gebruiker dat ook gaat doen. [↑](#footnote-ref-9)
11. HYDE - hierbij zoek je niet direct op de vraag maar laat je eerst het taalmodel een (incorrect) antwoord op de vraag formuleren zonder de nodige context en ga vervolgens zoeken naar stukken tekst die lijken op dit antwoord in plaats van op de vraag. Dit kan helpen omdat het antwoord dat je zoekt soms in structuur meer lijkt op een incorrect antwoord op dezelfde vraag dan de vraag zelf. [↑](#footnote-ref-10)
12. Bij een AI tool die voorspelt of iemand een bepaalde ziekte heeft, kan bijvoorbeeld gekeken worden naar sensitiviteit - hoeveel mensen die een ziekte hebben correct als ziek aangemerkt worden - en specificiteit - hoeveel mensen die een ziekte niet hebben correct als niet ziek aangemerkt worden. Dit kan gedaan worden door met de trainingset te kijken hoe vaak beide het geval zijn in die trainingset, en aangezien de trainingset als het goed is een goede steekproef is kan dit veralgemeniseerd worden. [↑](#footnote-ref-11)
13. Dit paper: [[2309.15217] RAGAS: Automated Evaluation of Retrieval Augmented Generation (arxiv.org)](https://arxiv.org/abs/2309.15217) beschrijft een veel gebruikt framework om RAG applicaties te evalueren, maar mijn persoonlijke ervaring en ook die van het PBL is dat de automatische beoordelingen hieruit niet kloppen met wat wij zelf belangrijk vinden in een antwoord. [↑](#footnote-ref-12)
14. Het bedrijf Anthropic, de makers van Claude 3, heeft over dit onderwerp bijvoorbeeld meerdere onderzoeken gepubliceerd: [Research \ Anthropic](https://www.anthropic.com/research), in het onderzoek waarin benoemt wordt hoe lastig dit is (challenges in evaluating ai systems) worden ook nieuwe opties genoemd voor evaluatie. [↑](#footnote-ref-13)
15. Bijvoorbeeld meerdere RAG toepassingen voor verschillende use cases [↑](#footnote-ref-14)