De kabinetsvisie op generatieve AI en retrieval augmented generation

*Door Jitse Goutbeek 6-2-2024*

Inleiding

Deze week is een generatieve AI toepassing uitgekomen die helpt om de kabinetsvisie op generatieve AI te verduidelijken[[1]](#footnote-0).



Op moment dat ik een vraag stel, gaat de applicatie zoeken in de kabinetsvisie naar stukken tekst die mogelijk relevant zijn om mijn vraag te beantwoorden, dit zoeken naar relevante stukken tekst wordt ook wel Retrieval genoemd en werkt een beetje als een slimme zoekmachine. Zodra relevante stukken tekst gevonden zijn worden die aan mijn vraag toegevoegd om een nieuwe vraag te formuleren, die ziet er ongeveer als volgt uit:

“Beantwoord op basis van de volgende context:

{Stukken tekst uit de visie op generatieve AI die de zoekmachine/retrieval gevonden heeft bij mijn vraag}

De volgende vraag:

{Hoe wordt generatieve AI in deze visie uitgelegd?}”

Het toevoegen van de relevante context wordt ook wel augmentation genoemd. Deze vraag wordt vervolgens aan Chat-GPT gegeven die hem beantwoordt, dit wordt generation genoemd. Dit hele proces heet ook wel Retrieval Augmented Generation oftewel RAG.

RAG helpt je bij het doorzoeken van documenten en het beantwoorden van vragen op basis van die documenten. De overheid heeft heel veel documenten, van beleidsevaluaties en wetsartikelen tot procesbeschrijvingen en contracten met leveranciers. Daarom wordt er binnen de overheid volop onderzoek gedaan of RAG de overheid kan helpen in het verwezenlijken van haar doelen.

Die doelen worden democratisch bepaald, door kabinet en kamer. Sinds 1 januari 2024 hebben wij een overheidsbrede kabinetsvisie op generatieve AI[[2]](#footnote-1), en RAG is een vorm van generatieve AI. Daarom leek het me goed om in dit document stil te staan bij de vraag hoe de onderzoeken die nu binnen de overheid naar RAG gedaan worden, bijdragen aan de visie van die overheid.

Ik zal beargumenteren dat RAG veel potentie heeft om bij te dragen aan de visie zoals die er ligt, maar het ook bepaalde zorgen raakt die voor een groot deel nog op waarde geschat moeten worden en waarbij het regelmatig een open vraag is hoe ze het best gemitigeerd kunnen worden. Dit brengt mij tot de conclusie dat de experimenten die nu plaatsvinden een mooie bijdrage aan de visie geven, ook omdat experimenteren op zichzelf als waardevol gezien wordt. Daarnaast zou ik concluderen dat er nog een aantal plekken voor nader onderzoek zijn om te bepalen of, hoe en wanneer het in productie nemen van RAG applicaties bijdraagt aan de visie.

Dit doe ik door eerst een nadere beschrijving te geven van wat er binnen de overheid aan RAG onderzoek loopt, en hoe de gebruikte RAG applicaties werken. Vervolgens beargumenteer ik de meerwaarde van (onderzoek naar) RAG. Tot slot kijk ik welke risico’s RAG potentieel raakt, en welke onzekerheden er op dat gebied nog zijn, om tot de conclusie te komen dat nader onderzoek gewenst is. Ik zal hierbij regelmatig verwijzen naar de kabinetsvisie, door in de voetnoten te refereren aan de tekst uit de visie waaruit een bepaald standpunt blijkt.

Deel 1: RAG onderzoek binnen de Nederlandse overheid

Er zijn tal van RAG-projecten binnen de overheid. Zo ken ik projecten bij o.a. I&W, J&V, Financiën, de Belastingdienst, het PBL en werk ik zelf bij het innovatieteam van SSC-ICT. Dat er zoveel projecten zijn laat zien dat de interesse hoog is, wat logisch is want iedereen binnen de overheid heeft wel informatie die ze makkelijker zouden willen kunnen doorzoeken. De meeste van deze projecten hebben ambtenaren als doelgroep, en zijn erop gericht om ambtenaren beter toegang te geven tot de kennis die in hun organisatie aanwezig is.

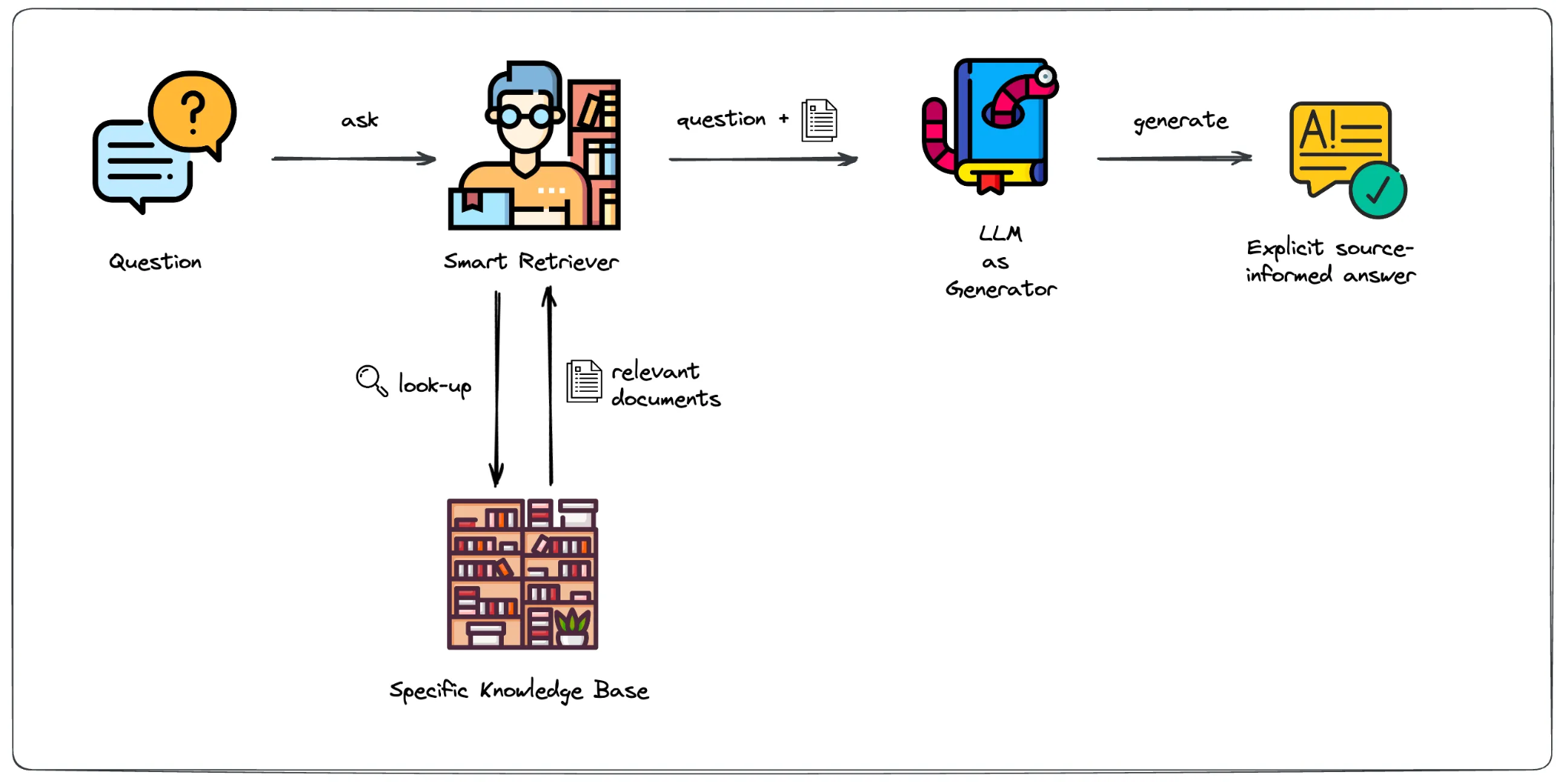
Deze projecten zijn niet allemaal identiek, ze proberen vaak andere problemen op te lossen en andere documenten te doorzoeken, Bovendien maakt niet iedereen dezelfde keuzes in het bouwen van hun test-applicatie, zo maakt de een gebruik van (Chat) GPT terwijl de ander dat niet doet. Deze verschillen hebben invloed op de kansen en risico’s. Maar hoewel er verschillen zijn, is er nog veel meer dat overeenkomstig is. Alle applicaties moeten grofweg hetzelfde doen:

*Voor de gebruiker een vraag stelt:*

1. De documenten waar een ambtenaar kennis uit wil halen worden ingeladen en opgeknipt in stukjes ookwel chunks genoemd (vaak van een paar honderd woorden)
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 vectoriseren. 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 een generatief taalmodel (ook wel Large Language Model of 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 meestal ook te zien.



*Versimpelde weergave van hoe RAG werkt*

Iedereen moet een aantal keuzes maken bij het bouwen van een dergelijke applicatie, dit zijn onder andere:

1. Hoe zoek je op een goede manier naar de meest relevante stukken tekst, welke verbeteringen zijn hierin door te voeren?
2. Welk taalmodel gebruik je om de vraag te laten beantwoorden?
3. Het door een taalmodel laten beantwoorden van een vraag kost best veel rekenkracht, hoe richt je dat het beste in?

De vraag is of het logisch is als iedereen die keuzes afzonderlijk maakt. Is de auteursrechtschending[[3]](#footnote-2) die mogelijk bij het trainen van een specifieke LLM heeft plaatsgevonden minder erg als we die LLM gebruiken om vragen over onderwerp A tov onderwerp B te laten beantwoorden? Uiteraard hangt het van de specifieke gegevens af die gebruikt worden hoe goed ze beschermd moeten worden. Welke gegevens we gebruiken bepaalt echter niet hoe goed ze beschermd zijn, dat hangt af van de inrichting van de applicatie.

De gemeenschappelijkheid van deze vragen, en mogelijk ook de antwoorden, wordt breed gezien. Daarom is er sinds begin februari ook een ‘community of practice’[[4]](#footnote-3) waar projecten die met dezelfde vragen worstelen daarover met elkaar kunnen praten en van elkaar kunnen leren. Dit stuk is bedoeld om relevant te zijn voor zoveel mogelijk van die projecten, en zal zich dus niet focussen op de specifieke toepassingen die wij bij SSC-ICT voor ogen hebben maar de gemeenschappelijke doelen en zorgen bespreken. Waar afwegingen afhangen van de specifieke toepassing zal ik dat zoveel mogelijk benoemen.

Deel 2: de meerwaarde van (onderzoek naar) RAG

**Ambtenaren beter hun werk laten doen**

In december 2023 heeft Van Huffelen een brief naar de kamer gestuurd waarin duidelijk staat dat het niet langer de bedoeling is dat ambtenaren gebruik maken van ‘niet-gecontracteerde’ generatieve AI toepassingen[[5]](#footnote-4). Dit zijn toepassingen als Chat-GPT, waarbij er niets afgesproken is met de leverancier om de risico’s te verminderen. In de brief wordt uitvoerig besproken dat dit een heel scala aan risico’s met zich meebrengt, die optellen tot een onacceptabele som.

Ondertussen gebruikte in Juni 2023 al 18% van de rijksambtenaren al regelmatig of vaak dit soort toepassingen omdat ze de meerwaarde zagen in hun werk[[6]](#footnote-5). Bovendien worden deze systemen alleen maar beter, en vaak sneller dan gedacht. In 2022 werden AI-experts gevraagd in te schatten wanneer AI in staat zou zijn een hele lijst met verschillende taken goed uit te voeren. Hun gokken: een geschiedenis essay schrijven in 2025, grote lijsten aan nummers sorteren in 2027, goede antwoorden geven op vragen die ambigue zijn (er is geen consensus correct antwoord) in 2032[[7]](#footnote-6). We zitten in 2023 en AI kan dit inmiddels allemaal al.

De overheid wil graag dat ambtenaren generatieve AI, op een verantwoorde manier[[8]](#footnote-7), kunnen gebruiken om hun werk beter en efficiënter uit te voeren[[9]](#footnote-8). Zeker als het aankomt op het doorzoeken van de vele documenten die de overheid heeft[[10]](#footnote-9). Zo is het voor beleidsambtenaren vaak niet eenvoudig om te weten wat er in het verleden al geprobeerd is en wat daar de evaluaties van waren, dit inzichtelijker maken kan de kwaliteit van beleid verbeteren.[[11]](#footnote-10)

Bovendien zijn sommige ambtenaren veel tijd kwijt met dingen opzoeken. Wij werken met mensen die diensten inkopen voor de overheid, en van hun leveranciers krijgen ze vaak vragen die al eerder beantwoord zijn of waarvan het antwoord in eerder opgestelde documenten of contracten staat. Ze zijn veel tijd kwijt met die documenten doorzoeken om er zeker van te zijn dat ze geen inconsistente informatie geven en iedere leverancier gelijk behandelen. Dit is arbeidsintensief (en dus duur) werk dat bovendien niet als erg prettig ervaren wordt, hier ziet de overheid ook een kans dat generatieve AI ervoor zorgt dat de kwaliteit van het werk toeneemt.[[12]](#footnote-11)

In theorie zou het gebruik van RAG door ambtenaren dus bij kunnen dragen aan de doelstellingen van het kabinet. Om te weten of we die meerwaarde ook echt in de praktijk kunnen realiseren, moeten we onderzoeken en experimenteren.

**Het belang van experimenteren**

Experimenteren met generatieve AI wordt door het kabinet een belangrijk punt gevonden, en meermaals aangemoedigd in de documentatie die daarover geschreven is. Ze willen dat we “via pilots beproeven van verantwoorde generatieve AI-toepassingen in concrete (proactieve) dienstverlening bij de overheid.”

Dit heeft ook grote meerwaarde. Het helpt voor ambtenaren als ze zelf zien waar AI goed functioneert en waar het (nog) tekortschiet, om een goede inschatting te maken waar het voor ingezet kan worden en waar de maatschappelijke risico’s liggen. AI gebruiken en ermee experimenteren is een manier om beter zicht te krijgen op de kansen en risico’s.

Daarnaast heeft experimenteren met RAG specifiek extra meerwaarde als het gaat om uitlegbaarheid. Een groot vraagstuk rond AI modellen is de uitlegbaarheid van de resultaten, omdat het eindresultaat bepaald wordt door vele, soms miljarden, berekeningen, is het lastig te herleiden waarom specifieke input tot specifieke output leidt.

Willen we met AI dienstverlening verbeteren[[13]](#footnote-12), dan zal de AI de dienstverlening moeten veranderen en dus invloed hebben op welke handelingen ambtenaren uitvoeren. Immers als er niets verandert, verbetert (of verslechtert) er ook niets. Als we echter niet weten waarom het AI systeem bepaalde output geeft, bestaat het risico dat er foutief op gehandeld wordt.

Een bekend voorbeeld hiervan is AI die in een ziekenhuis toegepast was, en een voorspelling maakte van welke patiënten een hoog en laag risico hadden op complicaties na hun behandeling. Dit gebeurde op tal van factoren en het was lastig te herleiden welke factoren welke impact hadden, maar de voorspellingen waren wel vrij accuraat. Op basis van de voorspellingen waren dokters geneigd patiënten met een hoog risico langer op te nemen en met een laag risico eerder naar huis te sturen, maar bij sommige patiënten ging dit tegen normaal beleid in. Toen een uitgebreidere analyse was gemaakt van waar de voorspellende waarde vandaan kwam bleek dat er een hoge correlatie was tussen welke patiënten standaard opgenomen werden en wie juist een laag risico toebedeeld hadden. Het model voorspelde dat sommige mensen een laag risico hadden, juist omdat ze standaard extra behandeld worden en daarmee hun risico verlaagd werd. Hadden de dokters laag risico patiënten eerder naar huis gestuurd, dan was daarmee hun risico drastisch toegenomen en had dit desastreus uitgepakt.[[14]](#footnote-13)

Dit voorbeeld laat zien dat uitlegbaarheid niet alleen maar gaat om verantwoording af kunnen leggen, maar ook om adequaat kunnen omgaan met de risico’s. Het belang van uitlegbaarheid wordt ook in de visie benoemd.[[15]](#footnote-14)

Dit raakt bovendien aan het probleem van hallucinaties, grote taalmodellen generen met enige regelmaat output die onzin is maar betrouwbaar overkomt. Omdat we niet weten waar ze hun informatie vandaan halen en waarom ze specifiek dit antwoord genereren is het ook lastig te achterhalen of wat er gezegd wordt waar of onwaar is, als op onware informatie gehandeld wordt kan dit negatief uitpakken. Temeer omdat automatische systemen vaak juist meer vertrouwd worden.[[16]](#footnote-15)

Retrieval augmented generation wordt gezien als een manier om hallucinaties tegen te gaan[[17]](#footnote-16). Je krijgt te zien welke context gebruikt is om tot het antwoord te komen, en er is vaak een instructie dat er geen antwoord gegeven moet worden als het niet in de context staat. Hieruit volgt het argument dat, aangenomen dat de RAG applicatie werkt zoals zou moeten, er een zekere vorm van uitlegbaarheid aanwezig is. We kunnen uitleggen dat dit antwoord voortkomt uit wat er in een specifiek document staat, dat document vertrouwen we en daarom kunnen we het antwoord vertrouwen.

Tegelijk is het niet te zeggen waarom precies dit antwoord op basis van de documenten gegeven wordt, of waarom specifieke stukken tekst het meest interessant voor de vraag gevonden worden. Bovendien is het heel prettig dat te controleren is of het antwoord overeenkomt met de gevonden context, maar betekent dit niet dat hallucinatie volledig uitgesloten is. Het kan immers dat het taalmodel claims maakt dat voorbij de context gaat, of de context verkeerd geïnterpreteerd wordt.

Dit betekent dat het een open vraag is in hoeverre RAG kan helpen bij het tegengaan van hallucinaties en uitlegbaar maken van gegenereerde tekst. Wanneer wij als overheid generatieve AI gaan gebruiken om ondersteuning te bieden bij belangrijke beslissingen is het cruciaal dit op een uitlegbare manier te doen. Onderzoek naar RAG kan ondersteunen in de zoektocht naar uitlegbare generatieve AI. Door verbeteringen door te voeren en applicaties te testen kunnen we proefondervindelijk kijken of we met een verder geoptimaliseerde vorm van RAG hallucinaties kunnen tegengaan en uitlegbaarheid vergroten. Hierbij is het uiteraard mogelijk RAG te combineren met andere technieken om uitlegbaarheid te vergroten, zoals een meer conceptuele benadering.[[18]](#footnote-17)

Hoe we als overheid uitlegbare generatieve AI aan kunnen bieden is een belangrijk, en nog open, onderzoeksvraagstuk. Onderzoek naar RAG kan hier een rol in spelen.

**De maatschappelijke potentie van RAG**

In RAG applicaties mitigeren we hallucinaties door de bronverwijzing te laten zien en dan kan de gebruiker controleren of er inderdaad overeenstemming is tussen wat in de bronnen staat en in het antwoord. Echter, als de bronnen toch door de gebruiker gelezen worden, waarom zou je dan niet alleen de bronnen laten zien, zonder antwoord? Een vraag die veel opkomt is of het verbinden van de zoekmachine aan een taalmodel voldoende meerwaarde heeft, bovendien een zoekmachine op zichzelf. Het op betekenis zoeken doet vaak al een belangrijk deel van het werk en zorgt ervoor dat je de relevante informatie snel terug kan vinden.

Dat het taalmodel daar dan een lopend antwoord van maakt bespaart misschien wat tijd maar het brengt ook nadelen met zich mee. Zoals we in het volgende hoofdstuk bespreken zijn er namelijk zorgen rond het gebruik van taalmodellen.

Het taalmodel heeft echter de meerwaarde dat het de gevonden tekst kan samenvatten en in **begrijpbare** taal uit kan leggen, jargon toe kan lichten en de gebruiksvriendelijkheid aanzienlijk vergroot. Wanneer ambtenaren dit gebruiken om teksten te doorzoeken die ze zelf geschreven hebben, vergroot dit potentieel het gemak en de efficiëntie maar het is niet van cruciaal belang. De vraag of RAG geschikter is dan een slimme zoekmachine die zoekt op basis van betekenis is daarmee een hele valide vraag, die voor iedere manier om RAG in te zetten nader bestudeerd dient te worden.

Echter, de overheid en het kabinet willen graag ook generatieve AI gebruiken om de overheid voor de overheid toegankelijker te maken[[19]](#footnote-18).

Hier zou RAG enorm bij kunnen helpen en is de koppeling met het generatieve taalmodel cruciaal. Als een burger een vraag stelt kan de zoekfunctie de juiste overheidsdocumentatie vinden, en het generatieve taalmodel net zo lang (vervolg)vragen beantwoorden totdat de burger het begrijpt. Bovendien kan die documenten vertalen of omschrijven naar toegankelijk nederlands.

In dit soort toepassingen zijn de risico’s die kleven aan RAG veel groter dan bij de toepassingen voor ambtenaren. Willen we iets aan burgers leveren dan is het extra belangrijk dat we het kunnen vertrouwen, en het geen niet discriminerende of inaccurate dingen zeggen. Bij ambtenaren kunnen we duidelijk zeggen dat het in een experimentele fase zit, er geen belangrijke beslissingen op gebaseerd moeten worden en als er een keer onwenselijke taal gegenereerd wordt is de schade makkelijker te beperken. Daarom is het cruciaal dat we dit eerst voor de overheid zelf maken, zodat we kunnen zien hoe het werkt en wat de risico’s zijn. Die risico’s kunnen we mitigeren en door ervaring hebben we een beter beeld van welke maatschappelijke toepassingen we wel en niet verantwoord kunnen leveren. RAG toepassingen voor ambtenaren proberen te ontwikkelen, en te kijken hoe goed het gaat, is daarmee cruciaal om de kennis op te doen uiteindelijk ook maatschappelijke toepassingen verantwoord te kunnen leveren.

De vraag die overblijft is wat verantwoord gebruik is, en wat ervoor nodig is om RAG toepassingen verantwoord te maken. Hier zal het volgende deel van dit stuk over gaan.

Deel 3: verantwoorde RAG toepassingen

Een RAG toepassing bestaat uit meerdere delen, enerzijds heb je het zoek-onderdeel van de applicatie. In de documenten die je zelf hebt toegevoegd wordt gezocht naar relevante stukjes tekst en die worden aan de vraag toegevoegd. Daarnaast heb je het generatieve AI onderdeel, waar een groot taalmodel gebruikt wordt om op basis van de relevante stukjes tekst een goed antwoord op de vraag te formuleren.

De grootste zorgen zijn er vaak over dit laatste onderdeel, het formuleren van het antwoord door het generatieve taalmodel. Dit taalmodel wordt niet door onszelf ontwikkeld, die capaciteit hebben we niet. Vaak zijn voor het trainen van deze modellen webscraping technieken gebruikt, het risico is hier dat (bijzondere) persoonsgegevens en auteursrechtelijk beschermde gegevens in de trainingsdata zitten. Het is een open vraag of grote taalmodellen het auteursrecht schenden[[20]](#footnote-19).

In principe is het verantwoord trainen van het model de verplichting van de ontwikkelaar, niet van organisaties die erop verder bouwen. Echter, we willen graag als overheid het juiste voorbeeld geven[[21]](#footnote-20). Enerzijds maakt een RAG applicatie altijd gebruik van een dergelijk taalmodel, en erf je dus de problemen van dat model. Anderzijds zijn er een tweetal gunstige eigenschappen van RAG applicaties in dit perspectief:

1. In een RAG applicatie beperk je de antwoorden tot dingen die gebaseerd zijn op de toegevoegde informatie, en dus geen kennis uit de trainingsdata te gebruiken. Dit maakt het onwaarschijnlijker in het algemeen dat informatie uit de trainingsdata naar boven komt tijdens het gebruik, ook auteursrechtelijk beschermde informatie of persoonsgegevens.
2. RAG applicaties zijn niet afhankelijk van een specifiek taalmodel, dit betekent dat welk taalmodel je gebruikt vaak helemaal vrij is en makkelijk te veranderen. Dit betekent dat wanneer de conclusie getrokken wordt dat bepaalde modellen op een ondeugdelijke wijze getraind zijn, een ander model gekozen kan worden. Andersom, zodra er modellen beschikbaar zijn waarvan we weten dat ze verantwoord getraind zijn, dan kunnen die vrijwel direct gebruikt worden. Denk hierbij bijvoorbeeld aan een succesvol product vanuit het GPT-NL project.

Dat tweede punt draagt bij aan het bewaren van een zekere gewenste onafhankelijkheid[[22]](#footnote-21) van grote Amerikaanse techbedrijven. Doordat in RAG applicaties het model eenvoudig te verwisselen is, vermindert dat de afhankelijkheid. Er kan expliciet gekozen worden voor modellen die niet van Amerikaanse leveranciers zijn, maar zelfs als wel voor een Amerikaans taalmodel gekozen wordt valt de afhankelijkheid mee gezien die keuze makkelijk verwisseld kan worden zodra het niet meer bevalt.

Het gevaar voor een afhankelijkheidsrelatie zit meer bij de hardware kant. Om een taalmodel een vraag te laten beantwoorden is rekenkracht nodig. Als we deze rekenkracht via de cloud van een Amerikaans techbedrijf halen zal die afhankelijkheid groter zijn dan de afhankelijkheid met de leverancier van het taalmodel. Een cloud-oplossing heeft echter weer voordelen als het op andere doelen aankomt.

**To cloud or not to cloud**

Een van de belangrijkste afwegingen bij het in productie nemen van een RAG-applicatie is hoe je de rekenkracht organiseert. Deze afweging goed maken vereist meer kennis en onderzoek dan ik momenteel ter beschikking heb. Dit is geen puur technische afweging, en ik zal proberen een grove schets te geven op welke onderdelen van de visie deze keuze invloed heeft en waar er nog veel onzekerheid is.

Een optie, die terug te vinden is in waardevolle onderzoeken en POCs zoals van het planbureau voor de leefomgeving[[23]](#footnote-22) is om een API-key te gebruiken. Dit betekent dat je in feite de vraag opstuurt naar het bedrijf waar je de API-call naar maakt die het vervolgens voor jou uitrekent. Voordeel hiervan is dat het makkelijk in te richten is en vaak heel goedkoop, een vraag kost doorgaans fracties van een cent[[24]](#footnote-23). Bovendien is het doorgaans makkelijk mogelijk om een API-key in te wisselen voor een andere, en ben je dus beperkt afhankelijk van een leverancier. Nadeel is dat de dataveiligheid slecht is, je stuurt immers je vraag en relevante stukjes document op naar een leverancier. In de kamerbrief over het gebruik van generatieve AI door de overheid staat uitgebreid beschreven dat er risico’s zijn aan het gebruik van generatieve AI als het aankomt op dataveiligheid. We willen niet dat OpenAI of een andere leverancier alles van ons weet, de overheid heeft veel documenten die om goede redenen niet openbaar beschikbaar zijn[[25]](#footnote-24). Gebruik maken van een API-key is daarom waarschijnlijk alleen acceptabel in test settings of als er louter openbare data gebruikt wordt, maar in de meeste gevallen zal het geen optie zijn. Het zou de moeite waard zijn verder te onderzoeken hoe onveilig de dataverwerking via een API key is, en of er situaties zijn waarvoor dit acceptabel is.

Een andere optie is om gebruik te maken van een cloudleverancier die dan ergens rekenkracht heeft die jij kan gebruiken. Hier zijn contractuele afspraken mee te maken over de bescherming van de gegevens, er zijn vaak omgevingen in te richten met extra garanties dat de gegevens lokaal blijven. Voor zover ik begrijp zijn er ook hier vraagstukken over hoe veilig de data daadwerkelijk is, je moet immers wel de cloudleverancier vertrouwen. Voor hele gevoelige informatie is dus ook dit geen oplossing, maar nu staat er ook al veel overheidsinformatie in de cloud. Weten welke documenten wel en niet verantwoord naar de cloud kunnen is een belangrijke vraag om dit goed in te richten. Bovendien zijn er vraagstukken rond afhankelijkheid van de specifieke leverancier, op dit moment maakt de overheid veel gebruik van Azure (Microsoft) voor cloud oplossingen. De vraag is of dit ook voor RAG applicaties de afhankelijkheid versterkt. Kan een hardware leverancier net zo makkelijk verwisseld worden als een taalmodel of is daar aanzienlijk meer voor nodig? Het zou goed zijn meer onderzoek te doen naar afhankelijkheidsrisico's bij verschillende cloudleveranciers.

Een laatste optie is om het volledig lokaal te doen, en zelf onze rekenkracht in te richten en te organiseren. De grote vraag is hierbij of we als overheid in staat zijn dat goed, efficiënt en veilig te doen. Mensen die ik hierover gesproken heb, hebben uiteenlopende opvattingen over hoe realistisch deze optie is. In feite zou het neerkomen op zelf een datacenter inrichten en onderhouden. Een veelgenoemd risico is dat wij dat niet zo efficiënt gaan kunnen als de grote cloudleveranciers, alleen al omdat er veel meer ambtenaren zijn die dinsdag om 10 uur een vraag voor werk zullen stellen dan zaterdag om 4 uur in de nacht. Dit zou niet alleen extra kosten kunnen betekenen, maar ook groter energieverbruik en dus klimaatimpact. Het kabinet maakt zich zorgen over dit energieverbruik[[26]](#footnote-25).

Hier staat wel tegenover dat energieverbruik ook voor een groot deel bepaald wordt door het aantal berekeningen dat er gemaakt moet worden. Tegenwoordig zijn er steeds meer kleine (open source) modellen die aardig functioneren, zo heeft de vlaamse wetenschapper Bram Vanroy in de eerste week van Februari een model uitgebracht dat tientallen tot honderden keren minder rekenkracht[[27]](#footnote-26) vereist dan de grootste modellen van OpenAI zoals ChatGPT. Uiteindelijk is er veel onzekerheid over hoeveel geld en energieverbruik er precies met de verschillende opties gepaard gaat, en is dat een onderwerp voor verder onderzoek.

**Naar productie**

Wanneer uit onderzoek blijkt dat een RAG toepassing waardevol is en nagedacht is over waar de rekenkracht vandaan gehaald gaat worden zijn er nog een paar randvoorwaarde waar aan voldoen moet worden voor het ook echt in productie genomen kan worden. Ten eerste mag het geen fundamentele rechten schenden.[[28]](#footnote-27) Om daar een goede inschatting van te maken dient er een bepaalde procedure gevolgd te worden:

“ Om vast te stellen welke specifieke vorm van inzet van generatieve AI wel of niet mogelijk is, dient voorafgaand aan het gebruik ervan per unieke casus een risicoanalyse uitgevoerd te worden. Dit zijn een Data Protection Impact Assessment (DPIA) en een algoritme impact assessment (zoals een Impact Assessment Mensenrechten en Algoritmes (IAMA)), waarin de risico’s en risico beperkende maatregelen worden vastgesteld. De uitkomsten hiervan dienen voorafgaand aan de inzet van de toepassing ter advies aan de (departementale) Chief Information Officer en de Functionaris Gegevensbescherming te worden voorgelegd.”

De exacte invulling van de IAMA en DPIA zal per toepassing een beetje verschillen, al zal er ook veel overlap in zitten. Uiteindelijk moeten de Chief Information Officer en Functionaris Gegevensbescherming oordelen welke RAG toepassingen toegestaan zijn, en die zijn (voor zover mij bekend) nog niet in de gelegenheid geweest zijn daar een oordeel over te geven.

Voor sommige toepassing lijkt de kans mij klein dat het stukloopt op het discriminatieverbod.

Antwoorden worden in RAG toepassingen beperkt tot wat er in de documenten van de gebruiker staat, het soort antwoorden dat je krijgt hangt daarmee af van de documentatie die gebruikt wordt. In veel overheidsdocumentatie, denk aan een contract met een aanbestedende dienst of een wettekst, staat geen tekst die betrekking heeft op personen of waar een hoog risico op discriminatie aanwezig is. Validatietechnieken en het AI-validatieteam kunnen een rol spelen in het nader bepalen van het risico op discriminatie.[[29]](#footnote-28)

Het oordeel van de functionaris gegevensbescherming zou waarschijnlijk afhangen van welke documentatie er doorzocht wordt en hoe de hardware is ingericht.

Naast deze oordelen is het belangrijk dat het voor de gebruiker duidelijk is dat ze met een AI interacteren[[30]](#footnote-29) en wat de beperkingen van het model zijn[[31]](#footnote-30). In sommige applicaties staat al standaard informatie over de applicatie op de interface waarmee de gebruiker met de applicatie interacteert. Zolang informatie dus duidelijk en begrijpbaar op het scherm te zien wordt aan deze eisen voldaan. Als er sprake kan zijn dat ambtenaren antwoorden uit de RAG applicatie gebruiken buiten de RAG applicatie is het belangrijk dat dit erbij vermeld wordt. Hierbij dient goed vastgelegd te worden over wie verantwoordelijk is voor deze vermeldingen en de handelingen die ambtenaren uitvoeren op basis van de door een RAG applicatie gegenereerde uitkomsten[[32]](#footnote-31).

Conclusie

RAG heeft veel potentie, zowel om overheidsprocessen efficiënter te maken als om burgers te helpen. Voordat deze potentie gerealiseerd kan worden, moet eerst goed onderzoek plaatsvinden, en kleinschalig met toepassingen geëxperimenteerd worden. Dit onderzoek versterkt bovendien het kennisniveau van de overheid en kan een bijdrage leveren aan vraagstukken rond uitlegbaarheid van generatieve AI toepassingen. Dit soort onderzoek en experimenten dragen daarom in belangrijke mate bij aan de visie op generatieve AI.

Daarnaast zijn er een aantal open vragen die beantwoord moeten worden voordat RAG applicaties verantwoord in productie genomen kunnen worden. Een cruciale vraag is hoe de benodigde rekenkracht gerealiseerd wordt, en wat dit betekent voor onze afhankelijkheidsrelaties, gegevensbescherming en duurzaamheidsambities. Bovendien moeten functionarissen gegevensbescherming en chief information officers nog oordelen of en onder welke voorwaarde RAG applicaties verantwoord ingezet kunnen worden, hier is ook een rol voor validatietechnieken om na te gaan of bias voldoende uitgesloten is.

1. [ChatGPT - Overheidsvisie Generatieve AI Uitgelegd (openai.com)](https://chat.openai.com/g/g-Oi0G6a4lL-overheidsvisie-generatieve-ai-uitgelegd) [↑](#footnote-ref-0)
2. [Overheidsbrede visie Generatieve AI | Rapport | Rijksoverheid.nl](https://www.rijksoverheid.nl/documenten/rapporten/2024/01/01/overheidsbrede-visie-generatieve-ai) [↑](#footnote-ref-1)
3. “Een tweede uitdaging betreft de mogelijke schendingen van rechten op het gebied van privacy, gegevensbescherming en auteursrecht en de daaraan verwante rechten. Zo kan de trainingsdata, veelal verkregen via grootschalige (web)scraping26 van openbare bronnen op internet of andere digitale bronnen, (bijzondere) persoonsgegevens bevatten.” [↑](#footnote-ref-2)
4. [Generatieve AI (pleio.nl)](https://generatieveai.pleio.nl/) [↑](#footnote-ref-3)
5. [Kamerbrief over voorlopig standpunt voor Rijksorganisaties bij het gebruik van generatieve AI | Kamerstuk | Rijksoverheid.nl](https://www.rijksoverheid.nl/documenten/kamerstukken/2023/12/11/kamerbrief-over-voorlopig-standpunt-voor-rijksorganisaties-bij-het-gebruik-van-generatieve-ai) [↑](#footnote-ref-4)
6. [18 procent van rijksambtenaren zegt regelmatig of vaak AI toe te passen - iBestuur](https://ibestuur.nl/artikel/18-procent-van-rijksambtenaren-zegt-regelmatig-of-vaak-ai-toe-te-passen/?tid=TIDP4136285X89728E2F55F04F19B0459B820C379515YI5&utm_campaign=IB_NB_Wekelijks&utm_medium=email&utm_source=ibestuur) [↑](#footnote-ref-5)
7. [Through a Glass Darkly—Asterisk (asteriskmag.com)](https://asteriskmag.com/issues/03/through-a-glass-darkly) [↑](#footnote-ref-6)
8. “wanneer generatieve AI op verantwoorde wijze wordt gebruikt, biedt het allerlei mogelijkheden om productiviteit op de werkvloer te vergroten.” [↑](#footnote-ref-7)
9. “Daarmee kan generatieve AI onder andere zorgen voor verbeterde efficiëntie, kostenbesparing, betere besluitvorming, betere dienstverlening en tal van innovatieve oplossingen.” [↑](#footnote-ref-8)
10. “In het kader van leren en ontwikkelen kan generatieve AI bijvoorbeeld ondersteuning bieden bij het analyseren van (een grote hoeveelheid) documenten” [↑](#footnote-ref-9)
11. “Het gebruiken van generatieve AI voor het analyseren van grote datasets voor beleidsvorming en -evaluatie. Hiermee kan de overheid beter inspelen op maatschappelijke behoeften en de effectiviteit van huidige maatregelen toetsen” [↑](#footnote-ref-10)
12. “Als productietool kan generatieve AI een positief effect hebben op de aard van het werk dat door mensen wordt uitgevoerd. Zo kan de inzet van AI routinematige taken (zoals het maken van notulen, het uitluisteren van audio of het beantwoorden van standaardvragen) van werknemers overnemen. Door de genoemde factoren kan de (gepercipieerde) kwaliteit van werk toenemen.” [↑](#footnote-ref-11)
13. “Ook de overheid kan van deze inzet van generatieve AI profiteren en zo haar dienstverlening proactief maken en optimaliseren.” [↑](#footnote-ref-12)
14. Dit voorbeeld komt uit het boek ‘the alignment problem’ van Brian Christian [↑](#footnote-ref-13)
15. “Moderne generatieve AI-modellen hebben bovendien een significant black box-karakter. Dit betekent dat we niet kunnen voorspellen wanneer een model op onwenselijke of onbetrouwbare wijze te werk zal gaan, en ook niet kunnen verifiëren dat de doelen die we een model meegeven correct in het model hun inbedding hebben gekregen. Dit gebrek aan interpreteerbaarheid en uitlegbaarheid wordt problematischer wanneer toekomstige modellen worden gebruikt om zelfstandig acties te nemen of beslissingen te nemen." [↑](#footnote-ref-14)
16. “Dit zogenoemde ‘hallucineren’ brengt grote risico’s met zich mee als het gaat om waarheidsvinding, vooral omdat veel mensen de neiging hebben tot ‘automation bias’ waarbij ze te veel vertrouwen hebben in de uitkomsten van geautomatiseerde systemen.” [↑](#footnote-ref-15)
17. [2005.11401v4.pdf (arxiv.org)](https://arxiv.org/pdf/2005.11401v4.pdf), [Addressing AI hallucinations with retrieval-augmented generation | InfoWorld](https://www.infoworld.com/article/3708254/addressing-ai-hallucinations-with-retrieval-augmented-generation.html) [↑](#footnote-ref-16)
18. [2310.01405.pdf (arxiv.org)](https://arxiv.org/pdf/2310.01405.pdf) [↑](#footnote-ref-17)
19. “[Generatieve AI] biedt overheden kansen om processen te verbeteren, het algemeen functioneren van de overheid te verbeteren en de dienstverlening aan burgers te optimaliseren. Bijvoorbeeld door bij te dragen aan het beter bereiken van de inwoner. Een andere mogelijkheid waar generatieve AI zich voor leent is het toegankelijker maken van overheidsinformatie voor iedereen, door in aanpassingen van het taalniveau te voorzien.” [↑](#footnote-ref-18)
20. “Op dit moment is de rechter nog niet in de gelegenheid gesteld om hierover te oordelen.” [↑](#footnote-ref-19)
21. “Allereerst zullen we als overheid het juiste voorbeeld geven.” [↑](#footnote-ref-20)
22. “Er is een groeiende afhankelijkheid van een selecte groep techbedrijven. De meest gebruikte generatieve AI-modellen en diensten in Nederland zijn afkomstig van een klein aantal Amerikaanse techbedrijven.” [↑](#footnote-ref-21)
23. [GitHub - pbl-nl/appl-docchat: Chat with your own documents pressure cooker](https://github.com/pbl-nl/appl-docchat) [↑](#footnote-ref-22)
24. [Experimenteren met RAG - Google Documenten](https://docs.google.com/document/d/1hDoKZ9ujWAn2x8eu2XZFXqhsg7YL6TqBNXtlaiK-hA8/edit) [↑](#footnote-ref-23)
25. 2 zorgen uit de kamerbrief met het voorlopig standpunt op generatieve AI raken hieraan:

    1, “De interactievragen van de gebruikers (‘prompts’) kunnen worden hergebruikt om het model te finetunen, waardoor onbedoeld gevoelige informatie openbaar kan worden gemaakt”

    2. “Een generatieve AI-toepassing kan zeer gevoelige informatie afleiden uit de interactie met de gebruiker” [↑](#footnote-ref-24)
26. “Een tweede maatschappelijke uitdaging heeft te maken met de grote energie die nodig is voor het trainen en gebruiken van generatieve AI-modellen. Als deze energie niet afkomstig is van hernieuwbare bronnen kan generatieve AI een onwenselijk effect hebben op klimaatverandering.” [↑](#footnote-ref-25)
27. [BramVanroy/GEITje-7B-ultra · Hugging Face](https://huggingface.co/BramVanroy/GEITje-7B-ultra) dit heeft 7 miljard parameters, ChatGPT zo 175 miljard [↑](#footnote-ref-26)
28. “De fundamentele rechten die geraakt worden door generatieve AI zijn het discriminatieverbod en het recht op privacy en gegevensbescherming.” [↑](#footnote-ref-27)
29. “Om verdere kennis en ervaring op te doen met de validatie van AI, heeft het kabinet een (Rijks) AI-validatieteam opgericht. Dit team buigt zich onder meer over het meetbaar maken van risico’s en kansen van generatieve AI.” [↑](#footnote-ref-28)
30. “Aanbieders van generatieve AI-systemen moeten ervoor zorgen dat het voor mensen dui delijk is dat ze met een AI interacteren of dat content door AI is gemaakt.” [↑](#footnote-ref-29)
31. “Gebruikers kunnen gebaat zijn bij begrijpelijke en overzichtelijke model cards – een soort bijsluiters met technische details en mogelijke beperkingen van het betreffende AI-model.30 Zo’n bijsluiter kan (eind)gebruikers ook helpen bij het bepalen of een model geschikt is, en of er eventuele gevaren bestaan bij de inzet ervan.” [↑](#footnote-ref-30)
32. “Het moet duidelijk zijn wie verantwoordelijkheid draagt voor het goed functioneren van AI-modellen en wie onder welke voorwaarden verantwoordelijk is voor eventuele schadelijke of ongewenste uitkomsten”. [↑](#footnote-ref-31)