# **Eindrapportage Generatieve AI**

# Introductie

## Project GenAI

In augustus 2023 begon vanuit innovatiemanagement een project waarbinnen onderzoek gedaan wordt naar de inzet van generatieve AI voor interne procesverbetering. Dit project lag in het verlengde van de door het SPB geaccordeerde opdracht ‘*AI en spraakbesturing*’ (urencode 201202044.002), welke een onderdeel vormt van de Service Roadmap van SSC-ICT.

De keuze om te focussen op generatieve taalmodellen, ook wel large language models (LLM’s), kwam voort uit de synergetische eigenschappen van LLM’s en spraakbesturing en de grote innovaties binnen het veld waardoor generatieve AI toepasbaar wordt binnen een scala aan verschillende use cases.

## Opzet document

In dit document wordt de voortgang van het project beschreven. Dit is een globaal overzicht, waarin naar documenten verwezen wordt die uitgebreider op de verschillende stappen en overwegingen ingaan. We beginnen met een overzicht van de vorderingen van alles dat we in ons plan van aanpak (PvA) voor het project generatieve AI beschreven hebben. Omdat wij niet hebben gewerkt in een vacuüm en er binnen de Rijksoverheid veel partijen bezig zijn met vergelijkbare en aanvullende initiatieven, zal daarna een schets worden gegeven van het huidige generatieve AI landschap en de positionering van ons project daarbinnen. Vervolgens worden de in 2023 behaalde resultaten samengevat. Tenslotte wordt aangegeven welke vervolgstappen er volgend jaar genomen kunnen worden binnen het project.

# 

# Overzicht deelprojecten PvA

| **Deelproject** | **Beschrijving** | **Documenten** | **Status** |
| --- | --- | --- | --- |
| **PvA update** | Een vernieuwde versie van het plan van aanpak geschreven en intern gedeeld op 5 okt 2023. | [PvA Generatieve AI v1.0.docx](https://docs.google.com/document/d/1X2zs8JKFbam8DCLBujGyoLRKHd5gzI4O/edit#heading=h.2jxww914ocr8) | Afgerond |
| **1.1 LLM onderzoek** | In september 2023 zijn een aantal van de best presterende open source LLM’s geanalyseerd. Vanwege het snel evoluerende veld kan het nodig zijn om een dergelijke analyse in de toekomst te herhalen. | [v1.0 - Vergelijking open source modellen.docx](https://docs.google.com/document/d/1U_0FFPa-3OQnJ1r6bwaljcNCusrRCz4o/edit)  Geüpdate versie:  [v2.0 - Vergelijking open source modellen.docx](https://docs.google.com/document/d/12yM6dpWzID3YLx7XDiGf5ds72Nb2NVS-/edit?usp=sharing&ouid=112016392227977150442&rtpof=true&sd=true)  Verbetermogelijkheden:  [Verbetermogelijkheden RAG](https://docs.google.com/document/d/12FnalqzluJwDR_EPZRVDQwPZTU0zeJFuWZOqTQSaR-M/edit?usp=sharing) | Afgerond |
| **1.2 Lo-fi PoC** | Een RAG applicatie op basis van open source modellen ontwikkeld als proof of concept. De werking van de demo is gepresenteerd in het MT CTO op 14 nov 2023. | [AI LLM demo I&S voor MT/CTO 14 nov.pptx](https://docs.google.com/presentation/d/1iKgjNSxWms0PJeJxTDiqyeKwyxoj9FjK/edit)  [Demo Github](https://github.com/SSC-ICT-Innovatie/LearningLion)  [Demo Architectuur](https://drive.google.com/drive/u/0/folders/15G-jv8uAbFQsUh-WmuKPGBybyFUZnhTb) | Afgerond |
| **1.3 Netwerk GPU’s** | Het is niet gelukt om de lokale GPU’s werkend te krijgen. De mogelijkheid wordt onderzocht om cloud GPU’s te gebruiken. |  | Aangepast |
| **1.3 Rapportage geteste modellen** | We hebben met verschillende tests onderzoek gedaan naar de effecten van de RAG applicatie instellingen die effect kunnen hebben op het functioneren van deze applicatie. | [Experimenteren met RAG](https://docs.google.com/document/d/1hDoKZ9ujWAn2x8eu2XZFXqhsg7YL6TqBNXtlaiK-hA8/edit?usp=sharing)  [Documenten bevragen met docchat](https://docs.google.com/document/d/1etHkvjMxHmY1PQFi0PXwsbSdMMY3nszTCpSZ6DktxvQ/edit?usp=sharing) | Afgerond |
| **1.4 Plan voor finetunen** | Het finetunen van een open source LLM is momenteel niet haalbaar binnen de scope en het budget van het project. Er zal gekeken worden naar de mogelijkheid andere applicatie onderdelen te finetunen. |  | Aangepast |
| **1.4 Dataverzameling RLHF en model-**  **verbetering** | Deze werkzaamheden worden al uitgevoerd door TNO, NFI en UWV. Dit deelproject zal vervallen in 2024. |  | Geannuleerd |
| **2. Azure Chatbot PoC** | Er is een start gemaakt met het opzetten van een chatbot in een Azure sandbox. |  | In uitvoering |
| **3. Eindrapportage en visiestuk** | *Huidige document* | [Eindrapportage genAI 2023 v0.2](https://docs.google.com/document/d/1t8toY01jxR7LQdoAXlBOMoEdQ5GiYduCf4E3Xj9le3Q/edit)  [Visiestuk SSC-ICT](https://docs.google.com/document/d/1J6k0HdlY5NMrOtSI0UhMYBZSoIlaqGwMx-FXzhio3Vk/edit?usp=sharing) | Afgerond |

# Landschap GenAI

## Initiatieven binnen de overheid

Binnen het publieke domein is er aanzienlijke interesse in kunstmatige intelligentie. Ons project streeft naar helderheid over de verhouding tot andere initiatieven binnen de overheid om dubbel werk te voorkomen en samenwerking te zoeken waar mogelijk. De meest grootschalige projecten zijn 'GPT-NL', waarbij EZK 13,5 miljoen investeert om een Nederlands LLM te laten ontwikkelen door o.a. TNO[[1]](#footnote-0), en ‘Leesplankje’ vanuit UWV, dat zich richt op het ontwikkelen van een Nederlandstalige dataset.

Daarom kiezen wij bewust niet voor het pre-trainen of fine-tunen van een model, we hebben de middelen niet en het gebeurt al op andere plekken. In plaats daarvan richten we ons op het onderzoeken van LLM ondersteunende of versterkende technologieën zoals Retrieval Augmented Generation (RAG). Het is mogelijk om hier applicaties mee te ontwikkelen die eenvoudig kunnen wisselen van model. Wanneer GPT-NL erin slaagt om een Nederlands foundational LLM te ontwikkelen of modellen op basis van de UWV dataset gefinetuned worden, kunnen we deze direct gebruiken.

Wij zijn niet de enige binnen de overheid die hiermee bezig zijn. Bij onder andere de Belastingdienst, ILT (IenW), JenV en het PBL onderzoeken en ontwikkelen teams vergelijkbare applicaties. Met deze organisaties wisselen wij informatie uit om generieke blauwdrukken te maken die af te stellen zijn voor onze eigen use cases en organisaties.

## Zorgen binnen de overheid

Er bestaan ook zorgen over het gebruik van LLM’s door ambtenaren. Deze zorgen zijn in 3 categorieën op te delen:

1. Wat er gebeurt met de door ambtenaren ingevoerde data
2. Afhankelijkheid van specifieke leveranciers
3. De data waarop de modellen getraind zijn

Ons project kan relevant zijn in het ondervangen van de eerste zorg omdat wij systemen lokaal runnen, waarbij de data dus lokaal blijft. Door onze applicatie flexibel te bouwen, verminderen we ook de afhankelijkheid van een specifieke leverancier. De derde zorg valt buiten de scope van dit project. Zorgen (zoals auteursrecht) die te maken hebben met de gegevens waarop een model getraind is, kunnen alleen beantwoord worden door projecten als GPT-NL die een model vanaf nul trainen.

# Resultaten 2023

## LLM Onderzoek

We hebben een literatuuronderzoek uitgevoerd naar de verschillende beschikbare LLM’s[[2]](#footnote-1). Uit dit onderzoek bleek dat verschillende goed functionerende open source modellen beschikbaar zijn. Voor onze toepassingen is het alleen bij veel van deze modellen problematisch dat ze niet zijn getraind op Nederlandse data. Een andere belangrijke bevinding van dit onderzoek was dat de snelheid waarmee nieuwe modellen worden ontwikkeld en gepubliceerd bijna niet bij te benen is. Mede hierdoor is het voor SSC-ICT minder interessant om te focussen op het trainen of finetunen van een LLM.

## Low fidelity proof of concept

Het is voor SSC-ICT wel interessant om al ervaring op te doen met het ontwikkelen van applicaties en dienstverlening die gebruikmaken van een LLM. Om dit aan te tonen hebben we een proof of concept ontwikkeld met Retrieval Augmented Generation (RAG)[[3]](#footnote-2). Dit laat een model gebruikmaken van een (lokale) database van documenten. Hierbij wordt in de documenten gezocht naar de relevante stukken tekst om een vraag te beantwoorden. De gebruiker krijgt ook te zien welke stukken tekst gebruikt zijn om tot dit antwoord te komen. De voornaamste voordelen van RAG zijn dat het nauwkeuriger is, omdat het antwoord gebaseerd is op de documenten die je er zelf aan toevoegt. Het model is minder geneigd om te ‘hallucineren’ en transparanter, omdat je kan zien op welke tekst het antwoord gebaseerd is. Daarbij biedt het een flexibele architectuur omdat zowel het LLM als de database eenvoudig te vervangen zijn. Dit betekent dat de applicatie in te zetten is voor een breed scala aan doeleinden, afhankelijk van welke documenten de gebruiker wilt bevragen.

## Onderzoek RAG

We hebben onder andere onderzoek gedaan naar het effect van het *embeddings model* (model dat de tekst in stukken opdeelt) en de *chunk size* (de grootte van de stukken) op de prestaties van de RAG applicatie.[[4]](#footnote-3) Dit stelt ons enerzijds in staat keuzes daarover beter te maken, anderzijds helpt dit ons een methodiek te ontwikkelen die we kunnen gebruiken om verschillende applicaties met elkaar te vergelijken.

We hebben een aantal resultaten waargenomen. Ten eerste is een groot aantal vragen correct beantwoord met de juiste stukken tekst. Dit laat zien dat RAG potentie heeft om voor meer transparantie in taalmodellen te zorgen (er wordt immers bij vermeld uit welk document en stuk tekst het antwoord komt) en dat je nu al met de GPT modellen redelijke prestaties kan behalen. Daarnaast zien we dat de keuze van het embedding model belangrijk is. Bij een op Nederlands getraind open source embedding model en het embedding model van OpenAI hebben we goede resultaten gezien, maar modellen die enkel op Engels getraind waren, zorgden soms voor een significante reductie in kwaliteit.

# Vervolg 2024

## Aangepaste doelstellingen

We zien dat UWV bezig is met het maken van een Nederlandstalige dataset die gebruikt kan worden om modellen te fine-tunen en TNO met veel meer middelen werkt aan het maken van een beter foundational model, bovendien komen er regelmatig nieuwe modellen uit. Het is voor ons daarom niet rendabel dit ook te doen, dus laten wij deelproject 1.4 deels vervallen.

Het verder uitzoeken en optimaliseren van verschillende RAG onderdelen zien wij het liefst gebeuren in samenwerking met andere overheidsorganisaties. Hoewel iedereen andere documenten en use cases heeft, maakt iedereen grofweg gebruik van dezelfde architectuur. Het is daarom van belang dat we in nauw contact blijven met deze organisaties en samenwerken om te kijken welke vragen al door andere partijen beantwoord zijn en bij welke nog onbeantwoorde vragen we echt meerwaarde kunnen leveren.

We kunnen hierbij ervoor kiezen specifieke onderdelen van de RAG applicatie te verbeteren, en die verbeteringen te delen met andere organisaties. Een voorbeeld hiervan is de ‘*similarity search*'. Dit is het onderdeel waarbij op basis van de gebruikersvraag relevante chunks worden gezocht. Soms vindt het model niet de juiste chunks, waardoor de vraag niet goed beantwoord kan worden. Door data te verzamelen over relevante en irrelevante gekozen chunks, kan dit onderdeel gefinetuned worden en de applicatie in zijn geheel worden verbeterd.

Andere mogelijkheden om RAG effectiever in te zetten zou zijn om het te combineren met LLM agents. Agents kunnen ingezet worden om te bepalen welke vervolgstappen genomen moeten worden wanneer een vraag binnenkomt. Hierbij kunnen ze gebruikmaken van verschillende tools, user input en tussenstappen. Door agents te gebruiken kan een vraag bijvoorbeeld eerst gecategoriseerd worden om vervolgens een specifieke knowledge base te gebruiken om een vraag te beantwoorden.

Waar we uiteindelijk naartoe hopen te werken is dat er binnen een paar jaar een generieke versie is van een RAG applicatie die open source beschikbaar is binnen (en buiten) de Rijksoverheid.

# 

# Conclusie & aanbevelingen

Vanuit het innovatieteam moeten wij 5 jaar vooruit kijken[[5]](#footnote-4). Binnen het gebied van LLM’s zien wij dat deze modellen zich snel ontwikkelen en deze technologie zich snel uitbreidt. Leveranciers zoals Microsoft hebben al aangegeven dat generatieve AI een meer centrale rol gaat spelen in hun diensten. De kans is groot dat over 5 jaar het gebruikt kan worden voor dingen waar het nu nog niet toe in staat is. Dat betekent dat wij nu al moeten beginnen met het bouwen van de infrastructuur, onderzoek doen en discussies voeren die het mogelijk maken daar als overheid autonoom in te handelen. Wij zien potentie in RAG om de overheid in eerste instantie voor zichzelf en later misschien zelfs voor de burger inzichtelijker te maken.

## Use cases binnen SSC-ICT

Dit jaar is er een werkend prototype gemaakt om aan te tonen wat de mogelijkheden van deze techniek zijn. Nu is het de zaak om aan te tonen waarvoor het binnen SSC-ICT te gebruiken is. Het doel is om één use case verder uit te werken en te testen. Hiervoor zijn intern al verschillende gesprekken gevoerd en presentaties gegeven. Omdat de technologie nog nieuw en experimenteel is, zou het een goed idee zijn om eerst aan de slag te gaan met interne use cases.

## Use Case: Inzicht in beleid

Er zijn vele mogelijke toepassingen, maar één waar wij veel potentie in zien is het creëren van inzicht in beleid door middel van RAG. De reden hiervoor is dat RAG zich nog vooral goed leent voor het doorspitten van grote hoeveelheden tekst. Aan het interpreteren van tabellen en afbeeldingen wordt gewerkt, maar dat werkt voorlopig nog niet voldoende. Verder blijkt uit onderzoek dat kenniswerkers ongeveer 30% van hun dag bezig zijn met het vinden van de juiste informatie. Een ondersteuning hiervan zou daarom een positieve invloed kunnen hebben op productiviteit en interne bedrijfsvoering.

Voor een uitwerking van de use case zou een grote database van niet-sensitieve beleidsbestanden moeten worden gebruikt, hiervoor zou bijvoorbeeld Sharepoint, Confluence of DigiDoc gebruikt kunnen worden. Hiervan wordt een eenvoudig doorzoekbaar bestand gemaakt, een vectorstore. We zouden een portaal maken of een teams-kanaal kunnen inrichten waarop medewerkers van SSC-ICT vragen kunnen stellen over beleid, bijvoorbeed: *“Welke standaarden bestaan er voor het testen van security van niet-kritieke systemen?”*. Op basis van de vraag, gaat de applicatie op zoek naar de meest relevante stukken tekst voor het beantwoorden van de vraag.

Via onze toepassing krijgt de gebruiker dan niet alleen de verschillende documenten waarin het antwoord op haar vraag gevonden kan worden, maar het LLM genereert een antwoord in natuurlijke taal. Bij het antwoord worden ook links gegeven naar de documenten die gebruikt zijn om het antwoord te genereren. Voor een voorbeeld van hoe zoiets eruit kan zien, kan je kijken naar [Postbus 42](https://www.postbus42.nl/).

Om deze use case te realiseren zijn een aantal dingen nodig. Ten eerste een GPU cluster om het LLM en andere onderdelen van de applicatie op te draaien. Ten tweede een database van documenten die gebruikt kan worden als context. De database moet eens in de zoveel tijd beschikbaar gesteld kunnen worden om een geüpdatete versie van de vectorstore te maken zodat recent gecreëerde bestanden ook meegenomen kunnen worden.

1. [Nederland start bouw GPT-NL als eigen AI-taalmodel | TNO](https://www.tno.nl/nl/newsroom/2023/11/nederland-start-bouw-gpt-nl-eigen-ai/) [↑](#footnote-ref-0)
2. Het volledige verslag van dit onderzoek: [v1.0 - Vergelijking open source modellen.docx](https://docs.google.com/document/d/1U_0FFPa-3OQnJ1r6bwaljcNCusrRCz4o/edit) [↑](#footnote-ref-1)
3. De code van deze PoC staat op github: <https://github.com/SSC-ICT-Innovatie/LearningLion> [↑](#footnote-ref-2)
4. Voor een uitgebreide beschrijving van de methode en resultaten zie: [Experimenteren met RAG](https://docs.google.com/document/d/1hDoKZ9ujWAn2x8eu2XZFXqhsg7YL6TqBNXtlaiK-hA8/edit) [↑](#footnote-ref-3)
5. Dit hebben we uitgebreider gedaan in ons [visiestuk](https://docs.google.com/document/u/0/d/1J6k0HdlY5NMrOtSI0UhMYBZSoIlaqGwMx-FXzhio3Vk/edit) [↑](#footnote-ref-4)