# **EcoSage**

### **LLM‑Powered Chatbot for Sustainable Policy**



**Publishers:**

* Nima Agha Ghafar Hamedani
* Jort Akershoek
* Vince Ammerlaan
* Tommi Lander

**Client**:

* Matis (De Haagse Hogeschool)

[**Inleiding 3**](#_57i3w3oplw9n)

[1.1 Probleemdefinitie 3](#_5s7v36z5oyyp)

[1.2 Doel van het onderzoek 3](#_a6pkeibzsa0v)

[1.3 Onderzoeksvraag en deelvragen 3](#_mos5o5ngass2)

[Hoofdvraag 3](#_27smc4gcm6jz)

[Deelvragen 3](#_4iooh2pmmqwl)

[1.4 Relevantie en context 4](#_ov5bx7meilpt)

[1.5 Aanpak en methoden 5](#_z58qsndsmwsc)

[1.6 Stakeholdersanalyse 5](#_8bug5899xx1r)

[**Literatuuronderzoek 7**](#_m8ywc4v55ggg)

[2.1 Inleiding 7](#_rhmyge2004yx)

[2.2 Energietransities en beleid in Europa 7](#_fsfedc9k8xn6)

[2.3 Toepassing van AI en LLM in beleid 8](#_1pg3vd32t363)

[**Methode 10**](#_ohrl66foqerk)

[3.1 Dataverzameling en selectie van documenten 10](#_la6rejhgasht)

[3.2 Exploratory Data Analysis (EDA) 10](#_162sdcekz16n)

[3.3 Data opschoning en preprocessing 11](#_346dhkzexbno)

[3.4 Architectuur RAG + LLM 12](#_sl5p1rgoo1pm)

[3.5 Prototype front end 14](#_47twn4owiso4)

[**Resultaten 16**](#_rvfo4y9n4e6i)

[**Discussie 17**](#_az1n3ljclgby)

[5.1 Reflectie op methodologische keuzes en gemaakte afwegingen 17](#_intplufmvv4t)

[5.2 Beperkingen van het onderzoek 19](#_g1tagmjyy0dc)

[5.3 Vergelijking met bestaande literatuur 20](#_bzu2128acp7a)

[5.4 Juridische en ethische implicaties 20](#_riavwkl969hc)

[**Literatuurlijst 21**](#_s21laxjarlkt)

# 

# Inleiding

## 1.1 Probleemdefinitie

Veel burgers ervaren een informatiebarrière als het gaat om de energietransitie. Beleidsdocumenten zijn vaak geschreven in een juridische en technische taal die moeilijk te begrijpen is voor mensen zonder specialistische kennis. Hierdoor blijft de betrokkenheid van burgers bij de energietransitie beperkt, wat een obstakel vormt voor de implementatie van duurzame oplossingen. Dit gebrek aan toegankelijkheid verminderd de effectiviteit van beleid en vermindert de kans op participatie van burgers.

Daarnaast bestaat het risico dat verkeerde of misleidende informatie zich verspreidt, omdat burgers uit onbetrouwbare bronnen informatie zoeken. Een tool die betrouwbare beleidsinformatie in begrijpelijke taal presenteert, kan bijdragen aan informatiedeling en een actievere rol van burgers in de energietransitie.

## 1.2 Doel van het onderzoek

Het doel van dit onderzoek is het ontwerpen, ontwikkelen en evalueren van een prototype van een Large Language Model (LLM)-gebaseerde tool die burgers helpt om complexe energie beleidsdocumenten te begrijpen. De tool moet in staat zijn om beleidsinformatie op een betrouwbare, toegankelijke en begrijpelijke manier te presenteren, zodat burgers beter geïnformeerde beslissingen kunnen nemen en actiever kunnen deelnemen aan de energietransitie. Het onderzoek richt zich op het toepassen van technieken zoals Retrieval-Augmented Generation (RAG) om relevante informatie op te halen en te herformuleren in begrijpelijke taal, met behoud van de inhoudelijke juistheid.

## 1.3 Onderzoeksvraag en deelvragen

### **Hoofdvraag**

Hoe kunnen we een LLM-gebaseerde tool ontwikkelen die beleidsdocumenten over de energietransitie analyseert en samenvat tot begrijpelijke en betrouwbare informatie voor burgers?

### **Deelvragen**

1. **Modelkeuze**Welk taalmodel sluit het best aan bij dit probleem, en welke rol speelt de taal van het pre trained model in de prestaties?
2. **Data-acquisitie en preprocessing**Hoe verzamelen en preprocessen we relevante beleidsdocumenten zodat ze geschikt zijn voor input in een LLM?
3. **Modelontwikkeling**Welke methoden en technieken zijn geschikt om het LLM te fine-tunen of aan te passen voor deze taak?
4. **Evaluatie**Hoe beoordelen we de kwaliteit, nauwkeurigheid en begrijpelijkheid van de gegenereerde samenvattingen voor burgers?

## 1.4 Relevantie en context

Veel mensen vinden het lastig om beleid over de energietransitie te begrijpen. Beleidsdocumenten van de overheid zijn vaak geschreven in moeilijke taal, met veel technische en juridische termen. Hierdoor weten burgers niet goed wat er mogelijk is of wat er van hen wordt verwacht. Dit zorgt ervoor dat minder mensen meedoen aan de energietransitie, terwijl hun betrokkenheid juist hard nodig is.

Daarnaast kan verwarring of onbegrip leiden tot het zoeken van informatie op onbetrouwbare plekken, wat het risico vergroot op verkeerde keuzes of desinformatie. Door beleidsinformatie op een duidelijke en betrouwbare manier aan te bieden, kunnen burgers betere beslissingen nemen en actiever meedoen aan de overgang naar duurzame energie.

Ook op economisch gebied is dit belangrijk. Als mensen sneller en beter toegang hebben tot de juiste informatie, kunnen duurzame oplossingen sneller worden toegepast. Dat helpt niet alleen bij het halen van klimaatdoelen, maar ook bij het besparen van kosten en energie.

Dit onderzoek vindt plaats aan De Haagse Hogeschool. In dit project werken we aan een tool die gebruikmaakt van een groot taalmodel (LLM). Deze tool zal moeilijke beleidsdocumenten over de energietransitie automatisch samenvatten en uitleggen in begrijpelijke taal.

Het project duurt 18 weken. Aan het eind willen we een werkende tool hebben: een getraind taalmodel én een simpele webpagina waarop mensen vragen kunnen stellen en duidelijke antwoorden krijgen. Zo willen we bijdragen aan een betere en eerlijkere energietransitie voor iedereen.

## 1.5 Aanpak en methoden

Om dit onderzoek goed uit te voeren, volgen we een aantal stappen. We beginnen met een literatuurstudie. We bekijken bestaande grote taalmodellen (LLM’s) zoals die van Hugging Face en OpenAI, en onderzoeken hoe teksten automatisch samengevat kunnen worden. Ook kijken we naar hoe overheden nu al proberen informatie toegankelijk te maken. Daarnaast voeren we gesprekken met onze opdrachtgever om goed te begrijpen wat zijn verwachtingen zijn. Zo zorgen we ervoor dat de tool goed aansluit bij de praktijk.

Voor de data gebruiken we bestaande beleidsrapporten in PDF-formaat. Deze rapporten moeten worden voorbewerkt. Denk hierbij aan het opsplitsen van lange teksten, het verwijderen van overbodige woorden (zoals stopwoorden), het omzetten naar kleine letters (lowercasing) en het opschonen van dubbele of irrelevante data. Dit alles gebeurt met technieken uit de Natural Language Processing (NLP).

Daarna kiezen we een geschikt model. Omdat we met Nederlandstalige teksten werken, kijken we naar LLM's die Nederlands begrijpen of meertalig zijn. Afhankelijk van wat er beschikbaar is, passen we het model aan via fine-tuning of prompt engineering. De resultaten van het model testen we met meetmethoden zoals ROUGE of BERTScore. We kijken ook of de samenvattingen begrijpelijk zijn voor gewone burgers.

Gedurende het project voeren we gebruikers- en leestoetsen uit, optimaliseren we het model en bouwen we een simpele interface. Het hele traject is opgedeeld in fasen met duidelijke mijlpalen, zodat we stapsgewijs naar een werkende en bruikbare tool toewerken.

## 1.6 Stakeholdersanalyse

Bij de ontwikkeling van een LLM-model voor de energietransitie zijn er verschillende stakeholders. Deze belanghebbenden hebben verschillende soorten van invloed en betrokkenheid, wat er belangrijk is om ervoor te zorgen dat alle verwachten in de juiste manier worden uitgevoerd

De stakeholders kunnen worden ingedeeld op basis van hun macht en invloed op het project:

- Hoge macht, hoge invloed

De opdrachtgever en wij de studenten hebben een grote invloed op de ontwikkeling en implementatie van het model. Wij bepalen de strategie, de technische implementatie en hoe het model wordt geïntegreerd in het energiebeleid en de toepassing ervan bij gebruikers. Regelmatige communicatie en goede samenwerking zijn erg van belang om het project succesvol te maken.

- Hoge macht, lage invloed

De beoordelaar oftewel de opleiding en De Haagse Hogeschool hebben een belangrijke invloed op de kwaliteit en beoordeling van het project. Zij spelen een belangrijke rol in de beoordeling en het behalen van projectdoelen, maar nemen geen beslissingen in de keuzes en de uitvoering. Het is belangrijk om goede communicatie te hebben en ze goed op de hoogte te houden en te voldoen aan de gestelde eisen.

- Lage macht, hoge invloed

Gebruikers (huishoudens en bedrijven) vormen de uiteindelijke doelgroep. Hoewel zij geen directe beslissingen nemen over de ontwikkeling, is hun invloed groot doordat hun feedback en adoptie van het model bepalen of het een succes wordt. Transparante communicatie is essentieel om ervoor te zorgen dat meer mensen het zullen gaan gebruiken.

- Lage macht, lage invloed

Overheidsdatabronnen en Hugging Face leveren de data en modellen voor het project. Zij hebben weinig invloed op de dagelijkse beslissingen, maar blijven relevant voor het onderhouden van een betrouwbare en up-to-date dataset.

# Literatuuronderzoek

## **2.1 Inleiding**

Op dit moment zijn er in heel Europa diverse beleidsmaatregelen die de overstap van fossiele brandstoffen (olie, gas) naar duurzame energiebronnen (zon, wind) stimuleren. Deze zogenoemde energietransitie is nodig om klimaatverandering te beperken en de energievoorziening op lange termijn te verduurzamen. Overheden, zowel op Europees niveau (bijvoorbeeld via de Europese Green Deal) als in Nederland (via het Klimaatakkoord), werken aan plannen en regels om deze transitie te realiseren.

Het probleem is echter dat deze beleidsdocumenten vaak technisch, omvangrijk en lastig te doorgronden zijn voor de gemiddelde burger. Daardoor weten mensen niet altijd welke subsidies of regelingen beschikbaar zijn, hoe zij kunnen deelnemen aan lokale projecten, of aan welke (wettelijke) voorwaarden zij moeten voldoen bij bijvoorbeeld de installatie van zonnepanelen of warmtepompen.

Uit onderzoek van het Rijksinstituut voor Volksgezondheid en Milieu (RIVM) blijkt dat burgers eerder geneigd zijn actie te ondernemen als zij weten welke mogelijkheden er zijn, en dat meer burgerbetrokkenheid de energietransitie kan versnellen (Successen En Valkuilen Bij Burgerparticipatie in De Energietransitie | RIVM, n.d.). Een Large Language Model (LLM) – een AI-systeem dat (grote hoeveelheden) tekst kan begrijpen en samenvatten – kan hier een bijdrage aan leveren. Het kan ingewikkelde beleidsstukken in begrijpelijkere taal vertalen en burgers via een vraag-en-antwoordsysteem van de juiste informatie voorzien (What Is an LLM (Large Language Model)? | Cloudflare, n.d.).

## **2.2 Energietransities en beleid in Europa**

In 2016 heeft de Europese Unie het Klimaatakkoord van Parijs ondertekend, met als doel de opwarming van de aarde ruim onder de 2°C te houden. De EU streeft er daarom naar om de CO₂-uitstoot tegen 2030 met 55% te reduceren (t.o.v. 1990) en in 2050 volledig klimaatneutraal te zijn (Ministerie van Infrastructuur en Waterstaat, 2024). De betrouwbaarheid van deze gegevens is hoog, omdat ze afkomstig zijn van de Rijksoverheid, die officieel verantwoordelijk is voor de uitvoering en rapportage van klimaatbeleid in Nederland.

Nederland werkt in stappen toe naar een volledig duurzame energievoorziening in 2050. Concreet betekent dit:

• De aanleg van windparken (op land en op zee).

• De uitbreiding van het hoogspanningsnet.

• De stimulering van zonne-energie.

• Investeringen in energiebesparingsmaatregelen.

(Ministerie van Algemene Zaken, 2024)

Daarnaast hanteert de EU de zogenoemde Hernieuwbare Energie Richtlijn (RED), die voorschrijft dat in 2030 minimaal 42,5% van de energie in de hele EU uit hernieuwbare bronnen moet komen. Voor Nederland is dit doel gesteld op 27%. Dit kan onder meer gerealiseerd worden via wind- en zonne-energie, maar ook via biogas uit rioolzuiveringsinstallaties (Van Hulst, n.d.). Van Hulst publiceert bij Europa Decentraal, een kenniscentrum gespecialiseerd in Europees recht en beleid voor lokale overheden, wat de expertise en betrouwbaarheid van de bron onderstreept.

**Financiële ondersteuning en draagvlak**

Om de energietransitie te stimuleren, biedt de Nederlandse overheid diverse subsidies en leningen aan. Zo kunnen zowel particulieren als bedrijven financieel worden ondersteund om duurzame maatregelen te nemen, zoals de aanschaf van zonnepanelen of het isoleren van gebouwen (Beleid Duurzame Energie, 2024). Deze informatie is afkomstig van de Rijksdienst voor Ondernemend Nederland (RVO), die gespecialiseerd is in het verstrekken en beheren van subsidies.

Naast overheidssteun blijkt uit onderzoek van de Vertegenwoordiging van de Europese Commissie in Nederland dat een groot deel van de Nederlandse bevolking (82%) klimaatneutraliteit belangrijk vindt. Bovendien vindt 68% dat de inzet van hernieuwbare energiebronnen sneller moet worden opgevoerd en de energie-efficiëntie verbeterd moet worden (Nederland En Het Klimaat: De Europese Green Deal, n.d.). Deze statistieken bevestigen het draagvlak voor de energietransitie.

## **2.3 Toepassing van AI en LLM in beleid**

**Voordelen van LLM’s**

LLM’s (Large Language Models) zijn in staat om grote hoeveelheden tekst te analyseren en samen te vatten, en vragen te beantwoorden in natuurlijke taal. In de context van energiebeleid kunnen LLM’s bijvoorbeeld complexe beleidsstukken vertalen naar toegankelijke samenvattingen of burger-specifieke antwoorden genereren (What Is an LLM (Large Language Model)? | Cloudflare, n.d.). Hierdoor wordt het voor burgers laagdrempeliger om in actie te komen en weloverwogen keuzes te maken rondom duurzame energie. Naast het vertalen van complexe beleidsstukken in begrijpelijke taal, kunnen LLM’s ook beleidsmakers ondersteunen door scenario-analyses en simulaties uit te voeren, bijvoorbeeld om de impact van nieuwe duurzaamheidsmaatregelen op energieverbruik en kosten te voorspellen

**Veilig gebruik van AI in de overheid**

Om AI-toepassingen veilig en transparant binnen de overheid te implementeren, wordt gebruikgemaakt van LLMOps (Large Language Model Operations). Dit is een methodische aanpak die bestaat uit:

1. Data voorbereiding: het anonimiseren en structureren van (historische) data.

2. Modelkeuze: een bestaand LLM selecteren en ‘finetunen’ voor een specifieke taak.

3. Implementatie: het model integreren in bestaande overheidsprocessen.

4. Monitoring: continu controleren op nauwkeurigheid en mogelijke vooringenomenheid (bias).

5. Verbetering: het model regelmatig bijstellen op basis van feedback.

(Verloop, 2024)

Deze vijfstapsaanpak wordt bijvoorbeeld gehanteerd door de Nederlandse Belastingdienst, die AI inzet om belastingaangiften te analyseren en fraude sneller te signaleren. Hoewel hetnog niet draait om een grootschalige LLM, laat het wel zien dat de overheid experimenteert met AI in beleids- en uitvoeringsprocessen.

**Internationaal voorbeeld: KemenkeuGPT**

In Indonesië is het ministerie van Financiën een stap verder gegaan met In Indonesië heeft het ministerie van Financiën een AI-model ontwikkeld: ‘KemenkeuGPT’. Dit model analyseert financiële data en regelgeving van het Ministerie van Financiën, Statistics Indonesia en het Internationaal Monetair Fonds (IMF). Onderzoek toont aan dat het de nauwkeurigheid van financiële analyses sterk verbetert, waardoor beleidsmakers betere beslissingen kunnen nemen (Febrian & Figueredo, 2024). Deze studie is te vinden op arXiv.org, een bekende open-access database voor wetenschappelijke preprints. Dit betekent dat het onderzoek nog niet officieel is gecontroleerd door andere wetenschappers voordat het werd gepubliceerd, maar het wordt wel wereldwijd gelezen en van feedback voorzien door experts, wat helpt bij de beoordeling van de inhoud.

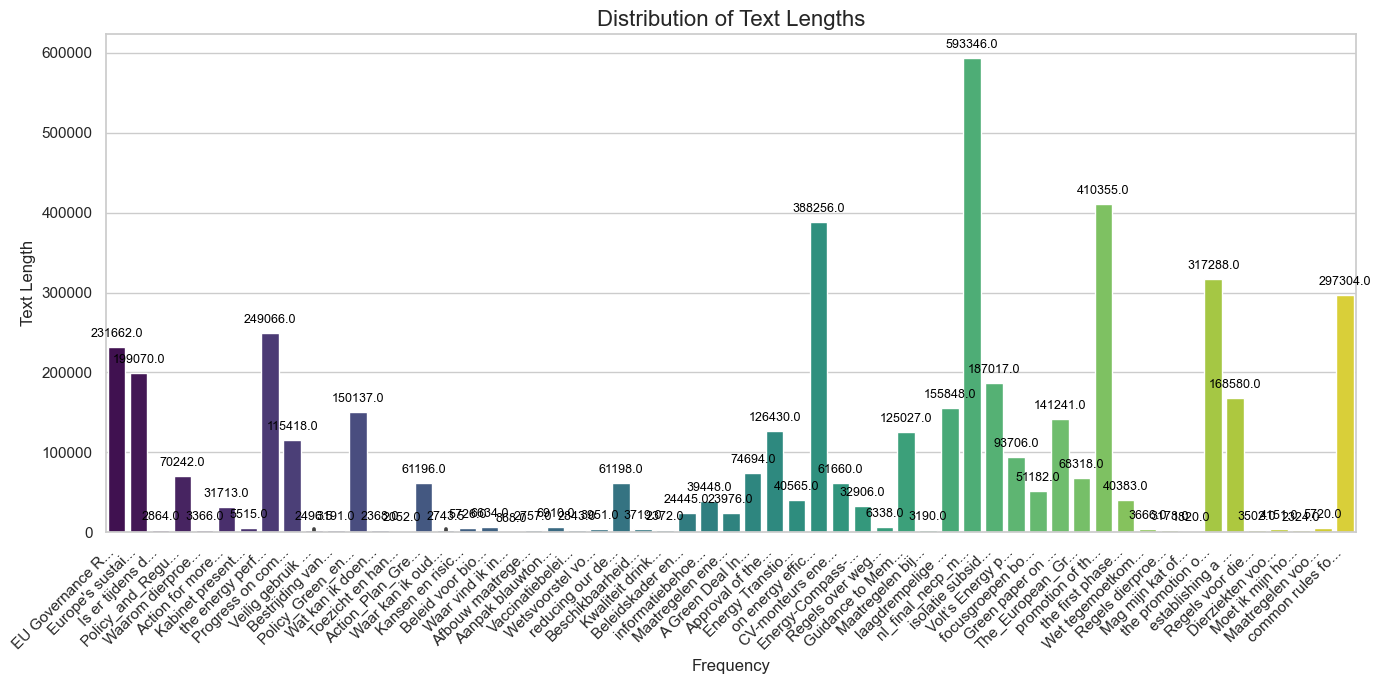
# Methode

## 3.1 Dataverzameling en selectie van documenten

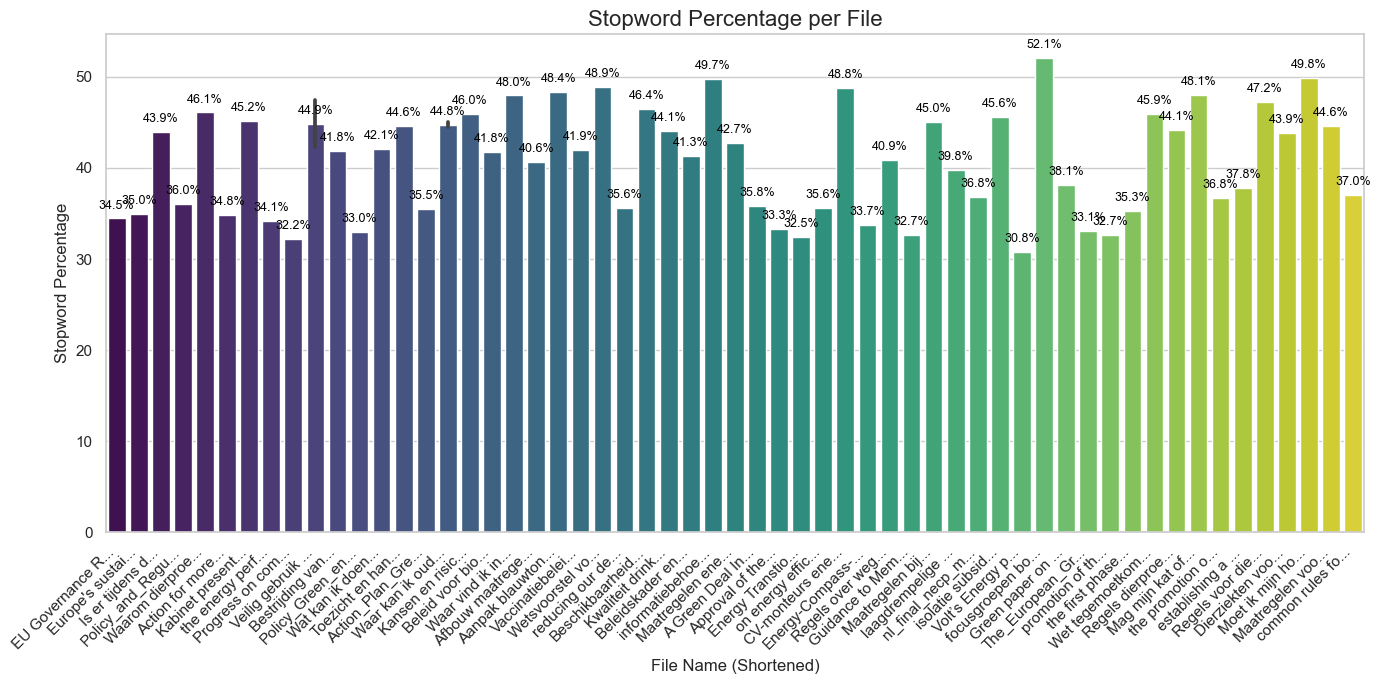
Voor dit project hebben wij van onze opdrachtgever een aantal Engelstalige documenten ontvangen. Deze bevatte informatie over Europese wetgeving omtrent de energietransitie en subsidies. Dit waren ongeveer zes documenten. Verder zijn wij zelf diverse overheidswebsites afgegaan om meer data te werven. Wij hebben documenten geselecteerd op trefwoorden die overeenkomen met de energietransitie. Ook moesten we zowel Engelse als Nederlandse documenten ophalen. Dit is omdat ons model Nederlanders en niet-Nederlandstaligen moet kunnen helpen.

## 3.2 Exploratory Data Analysis (EDA)

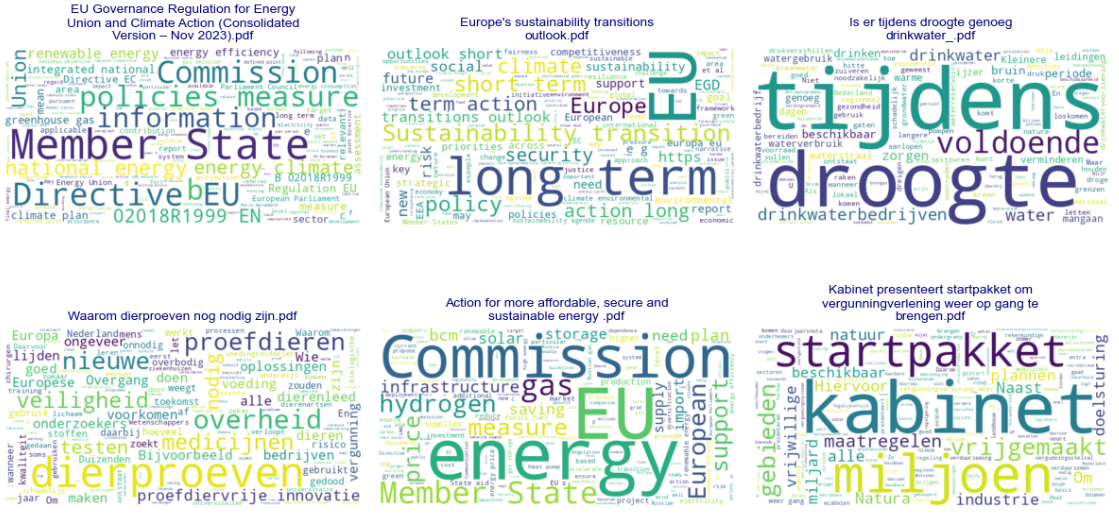
We begonnen met kijken naar de lengte van de tekst in de documenten. We zien dat er veel variatie is in de teksten. Dit is niet zo heel erg aangezien die informatie erin het belangrijkst is.



Daarna keken we naar de hoeveelheid stopwoorden per document als percentage van de totale tekst. Ook zien we hier de documenten ongeveer evenveel percentage hebben.



Verder hebben we ook een aantal “wordclouds” gemaakt voor de documenten. Deze wordclouds helpen met het bepalen of de documenten daadwerkelijk betrekking hebben op onze usecase.

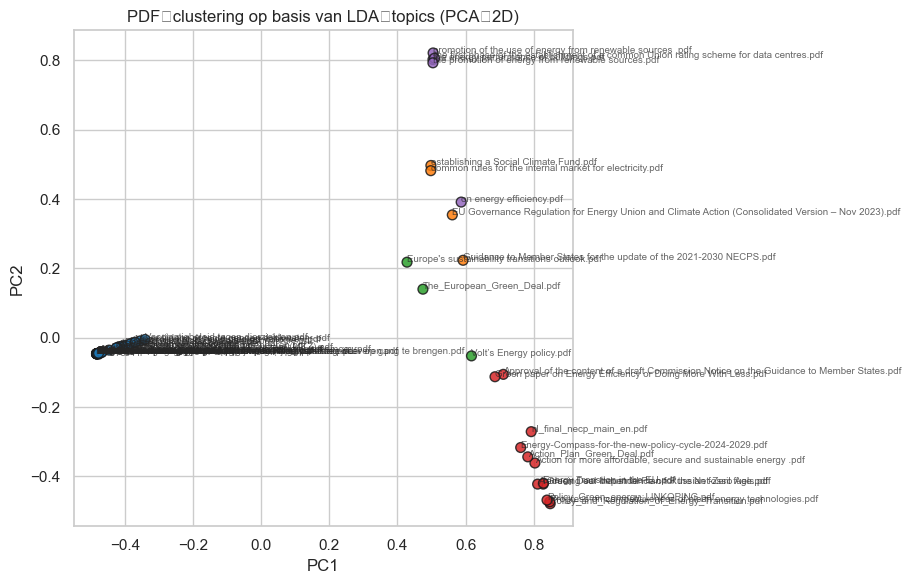


## 

## 3.3 Data opschoning en preprocessing

Moderne LLMs hebben minimale preprocessing nodig. Eerst hadden we alle stopwoorden verwijderd. Echter kwamen we erachter dat moderne LLMs juist beter kunnen presteren met de extra stopwoorden als context. Dus hebben we besloten om deze toch in te houden. (Singh, 2024)

Verder hebben we de data schoongemaakt door extra whitespace te verwijderen, alleen ASCII-letters en standaard leestekens te behouden en de text lowercase maken.

Daarna “chunken” we de text met hulp van het “gpt2” model. Dit breekt lange teksten in kleinere stukjes zodat ze in het context window van het LLM passen. Veel LLMs hebben token limits dus chunken zorgt ervoor dat we binnen deze limits blijven.

Daarna voeren we een LDA uit. Een LDA kijkt naar de topics van de chunks en zet deze bij elkaar. Hieronder te zien welke documenten dezelfde topics bespreken. De documenten aan de linkerkant van de grafiek lijken sterk op elkaar en gaan waarschijnlijk over soortgelijke onderwerpen. Ze vormen één grote groep omdat ze veel inhoudelijke overlap hebben. De documenten rechts hebben minder overlap, maar kunnen alsnog relevant zijn, aangezien er bepaalde vragen van de ‘user’ die in die documenten zijn vermeld.

## 3.4 Architectuur RAG + LLM

**Onze keuze: Retrieval-Augmented Generation (RAG)**

In plaats van het model telkens opnieuw te finetunen, wat uren tot dagen kost en een enorme GPU nodig heeft, zorgt Retrieval-Augmented Generation (RAG) ervoor dat het model de data kan gebruiken zonder te finetunen. Bij elke vraag gebeurt het volgende:

1. Embedden: De vraag en alle vooraf gechunkte document stukken zijn al omgezet naar vectoren (getallen­reeksen) die de betekenis bevatten in vectoren.
2. Retrieval: Een razendsnelle vector-zoekopdracht (FAISS) vergelijkt de vraagvector met de document vectoren en levert bijvoorbeeld de top-5 meest gelijkende chunks binnen een aantal seconden.
3. Context: Die chunks worden letterlijk in de prompt geplakt vóór de eigenlijke vraag, zodat de LLM ze “leest” alsof het net ingetypte achtergrondinformatie is.
4. Generatie: De LLM produceert het antwoord, nu met actuele tekst en bronnen uit de verschillende PDF’s.

Dit ontwerp heeft grote voordelen:

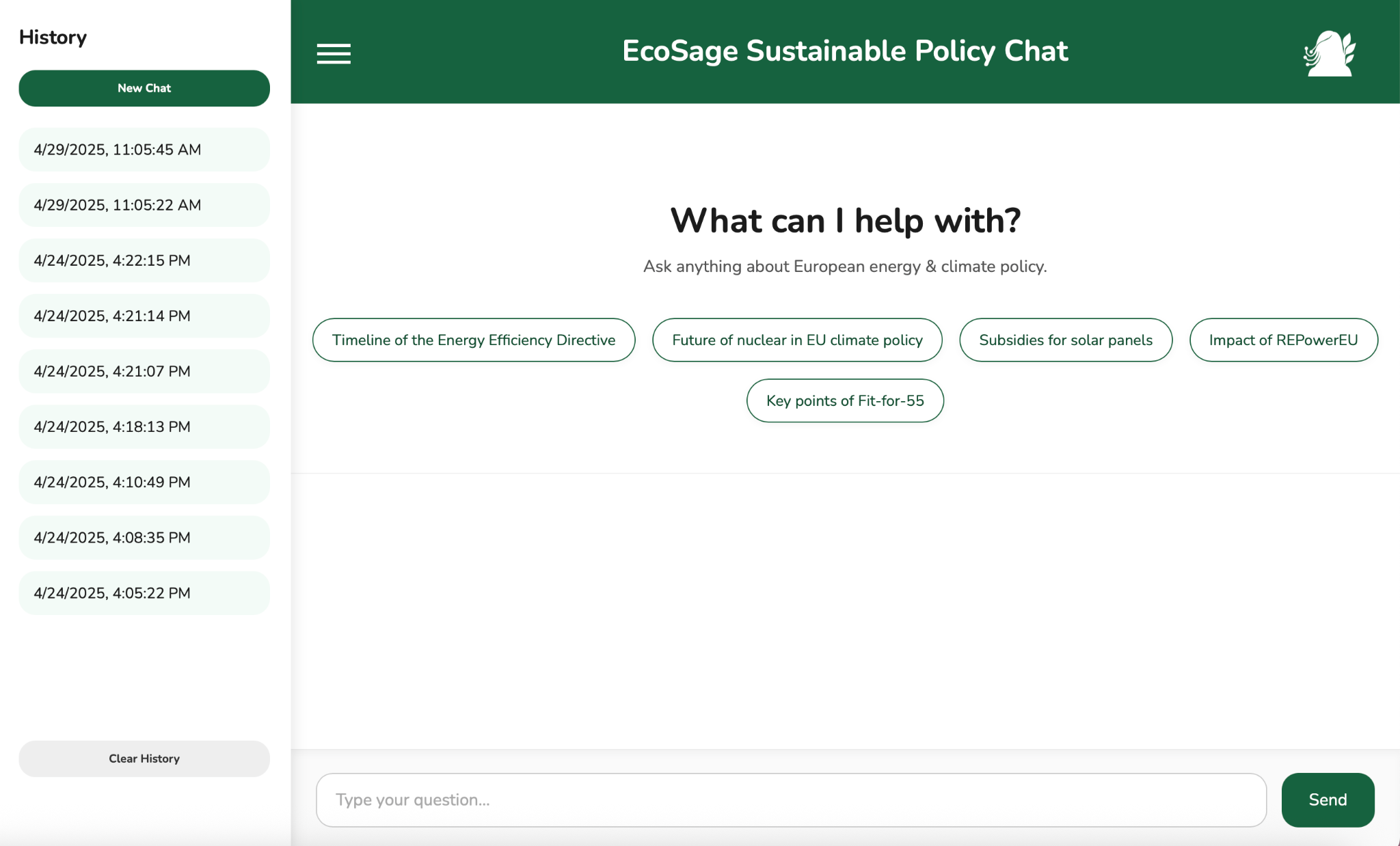
* Altijd up-to-date: Als er een nieuw beleidsstuk (bijv. een nieuwe policy) moet worden toegevoegd, dan hoef je alleen de vector-index te her-embedden; de model­ weights blijven onaangepast.
* Schaal & snelheid: Een data update duurt minuten in plaats van uren, en een inference vraagt hooguit een paar seconden GPU-tijd in plaats van een volledige finetuning van honderden GPU-uren.

**Hoe zit onze pipeline technisch in elkaar?**

| Stap | Wat het doet |
| --- | --- |
| Embeddings | Embedding model is *sentence-transformers/all-MiniLM-L6-v2*. Het is compact, snel en heeft een goed NL-begrip (zie EMBEDDING\_MODEL\_NAME in config.py). |
| Vectors | FAISS lokaal (models/faiss/), 384-dim embeddings; volledig offline dus geen cloud-kosten. |
| Chunking | Elke pdf-tekst wordt gesplitst in blokken van 512 tokens met 50 tokens overlap (CHUNK\_SIZE & OVERLAP in config.py + chunk\_text() in preprocessing.py). |
| Retrieval | Standaard top-5 meest lijkende chunks (K = 5). |
| Generatie model | Het generatie model is: *perplexity-ai/r1-1776-distill-llama-70b* (70 B LLama-variant, half-precision). |
| Device keuze | Automatisch: CUDA → Apple MPS → CPU (logica hiervan in config.py). |
| Prompt-template | “You’re an expert in European policy…” + {context} + {question}, bewaard als PROMPT\_TEMPLATE in [config.py](http://config.py). Dit wordt gebruikt bij elke vraag als prompt. |

## 3.5 Prototype front end

We hebben een front end gemaakt waarin de vragen aan de chatbot gesteld kunnen worden. Hierin zijn verschillende opties toegevoegd. Hieronder staat een screenshot en meer uitleg:



**Belangrijkste functies**

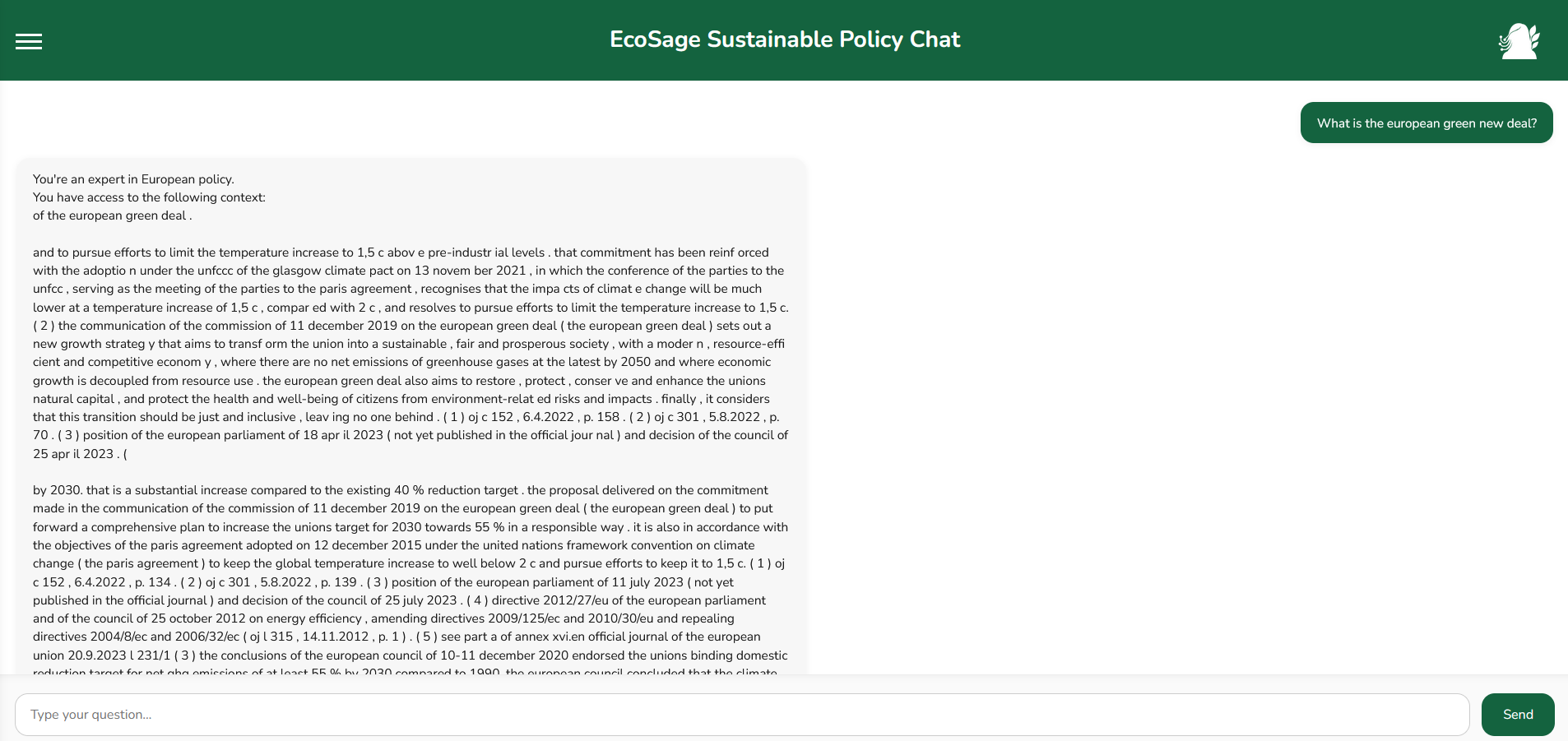
| **Functie** | **Hoe het werkt in de code** | **Wat de gebruiker merkt** |
| --- | --- | --- |
| **Chatvenster** | Bouwt “bubble”-divs met klassen user/bot. | Eigen bericht rechts, antwoord links in grijs. |
| **Typ-indicator** | Voegt drie pulserende stipjes toe tot de backend antwoordt. | Geeft feedback dat het systeem bezig is. |
| **Suggesties** | Toont 5 willekeurige voorbeeld­vragen uit array met 15 verschillende opties. | Eén klik en de vraag wordt gesteld. |
| **Chat-historie** | Sessies in localStorage, sidebar laat lijst van gesprekken zien. | Vorige gesprekken terugzien of wissen. |
| **Sidebar** | Knop voegt/haalt klasse open met schuif animatie. | Menu in-/uitschuiven op mobiel. |

# 

# Resultaten

Tot nu toe hebben we dus een mooie front end waar gebruikers de vragen kunnen stellen. Ook hebben we ons model aan de front end gekoppeld zodat hier ook echt gegenereerde antwoorden op komen. Hier komen al wel resultaten uit. Alleen duurt dit heel erg lang voor we een resultaat krijgen omdat wij dit nog moeten verbeteren. Wel was het resultaat wat we eruit kregen zeer hoopvol. We hebben namelijk gevraagd om een uitleg van de european green deal, vervolgens kwam er een uitleg over hierover, alleen duurde dit dus erg lang. Wij zijn nu bezig met het versnellen en verbeteren van ons model en zullen daarna metrics gaan gebruiken en verder onderzoeken of de antwoorden altijd goed zijn. Alleen hier zijn wij echter nog niet.

Wel hebben we nu dus de front-end gekoppeld aan het model. Hieronder staat een afbeelding van een vraag aan het model. Hier hebben we wel een simpeler model gebruikt, namelijk: *openai-community/gpt2-medium*. Deze zullen we uiteindelijk niet gaan gebruiken. Maar voor deze eerste tests hebben we dit model gebruikt zodat we dit op onze laptops kunnen testen.



Hierin is te zien dat we wel nog veel moeten aanpassen. Zodat bijvoorbeeld de context niet wordt teruggestuurd in het antwoord. Hier zullen we de komende weken op gaan focussen.

# 

# Discussie

## 5.1 Reflectie op methodologische keuzes en gemaakte afwegingen

Tijdens de voorbereiding van dit project hebben we diverse taalmodellen geëvalueerd, waaronder DeepSeek en andere open‑source oplossingen. Ook hebben we nagedacht over een nederlands taal model zoals “[ChocoLlama](https://huggingface.co/ChocoLlama/Llama-3-ChocoLlama-8B-instruct/blob/main/README.md?code=true)”, hier hebben we niet voor gekozen omdat we het niet willen reduceren tot alleen nederlands. Door vergelijkende tests is de keuze gevallen op het model “[perplexity-ai/r1-1776](https://huggingface.co/perplexity-ai/r1-1776)” van Hugging Face, dat eigenlijk een variant van DeepSeek is, maar dan met de Chinese propaganda verwijderd. Deze keuze was gebaseerd op de volgende overwegingen:

* **Multi-language:** Wij hebben zowel Nederlandse als Engelse beleidsdocumenten in onze data, hierdoor is de chatbot in staat om zowel Nederlandse als Engelse vragen beantwoorden.
* **Functionaliteit en prestaties:** het model levert betrouwbare, domeinspecifieke antwoorden over duurzaamheidsbeleid en wet- en regelgeving mits de juiste databronnen.
* **Implementatiewendbaarheid:** updates via de Hugging Face-API vereisen geen hertraining, waardoor nieuwe documenten direct beschikbaar zijn voor retrieval in de RAG-pipeline.
* **Propoganda vrij en licentievoorwaarden:** het model is vrij van ongewenste bias content en kent open-source condities.

De oorspronkelijke dataset van de opdrachtgever bleek onvoldoende representatief voor de duurzaamheidsvraagstukken, omdat het maar bestond uit zes bestanden. Daarom hebben wij aanvullende databronnen verzameld via publieke bronnen (rijksoverheid.nl, milieucentraal.nl, overheid.nl en commission.europe.eu). Om dit allemaal op te halen hebben we een proef-scraper gebouwd om zo automatisch teksten te verzamelen, maar vanwege verschillende sitestructuren hebben we uiteindelijk (voor nu) zestig relevante documenten handmatig gedownload. Deze documenten, beleidsnota’s, subsidies en rapportages over de energietransitie vormden onze database van ons project.

Voor een eerste dataverkenning (EDA) hebben we tekstlengtes en de frequentie van stopwoorden geanalyseerd, waarna we voor elke tekst wordclouds genereerden. Veelvoorkomende verbindingswoorden (zoals "daarnaast") zijn uit de visualisatie gefilterd om patronen duidelijker te zien. Deze woorden waren vaak erg groot aanwezig in de wordclouds, maar vertelden erg weinig.

Op basis van literatuur (o.a. Gelson Mendes, 2022) hebben we ervoor gekozen de originele tekst als input voor het model niet te zuiveren of stopwoorden te verwijderen, omdat dit de uiteindelijke performance nauwelijks beïnvloedt, terwijl context nuances behouden blijven.

Voor clustering van documenten hebben we Latent Dirichlet Allocation (LDA) toegepast. LDA biedt een transparante, ongecodeerde topic‑model dat goed uitlegbaar is aan stakeholders en snel inzicht geeft in de thematische samenhang van de documenten. Alternatieven zoals k-means of hiërarchische clustering hebben we overwogen, maar vereisten vectorberekeningen en leken minder geschikt voor interpretatie in de context van beleidsanalyses.

De gekozen RAG‑architectuur combineert:

* **Embedding-model:** sentence-transformers/all-MiniLM-L6-v2
* **Generatie-model:** perplexity-ai/r1-1776 Z
* **Retriever-configuratie:** TOP\_K = 5, CHUNK\_SIZE = 512 tokens, OVERLAP = 50 tokens. De reden waarom wij 512 hebben gekozen is omdat we zo meer context kunnen behouden tijdens het beantwoorden van de vragen. Waarom exact 512? Dit zagen wij voorbij komen in een artikel van Pinecone (*Chunking Strategies for LLM Applications*, n.d.). Hierbij werd vermeld dat 512 een mooi aantal is om context te behouden. Top\_K = 5 is een veelgebruikte waarde voor RAG om documenten op te halen, dus ons geval 5. Hetzelfde geldt voor Overlap = 5. Een overlap van 50 tokens tussen opeenvolgende chunks zorgt ervoor dat er geen belangrijke informatie verloren gaat bij het splitsen van documenten. Dit is vooral belangrijk bij langere teksten waar de context aan het einde van een chunk relevant kan zijn voor het begin van de volgende.

Daarnaast hebben we de volgende trade‑offs afgewogen:

1. **Retrieval versus fine-tuning** – Fine-tuning zorgt voor betere resultaten op maat, maar Retrieval-Augmented Generation (RAG) maakt het makkelijker om je model steeds bij te werken zonder dat je het opnieuw hoeft te trainen.
2. **Chunk-grootte versus contextbehoud** – Kleinere stukjes tekst zorgen voor meer precisie, maar grotere stukjes houden meer context vast. We hebben een middenweg gekozen.
3. **Latency versus informatiediepte** – Door het aantal retrieval-interacties te beperken, blijft de reactietijd voorspelbaar zonder al te veel details te verliezen.
4. **Kosten versus modelkracht** – We gebruiken een hybride model (een lichter model voor eenvoudige vragen, en een zwaarder model voor ingewikkeldere vragen) om de kosten te beheersen.
5. **Privacy versus datatoegankelijkheid** – We zorgen ervoor dat we voldoen aan privacywetten (zoals de AVG) door data anoniem te loggen, terwijl we tegelijkertijd het model kunnen verbeteren.
6. **Brondiversiteit versus indexcomplexiteit** – Het kiezen van een breed scala aan bronnen helpt om vooringenomenheid (bias) te voorkomen, maar het maakt het onderhoud wel complexer.
7. **Duurzaamheid versus performance** – Door batches van inferenties te gebruiken en energiezuinige servers in te zetten, kunnen we de ecologische impact beperken.
8. **Meertaligheid versus resourcebeheer** – We richten ons op één taal om de kosten en complexiteit van de benodigde infrastructuur te beperken.

## 5.2 Beperkingen van het onderzoek

1. **Handmatige data-verzameling:** handmatig downloaden van documenten kan tot missende updates leiden en is foutgevoelig. Ook duurt het verzamelen van data lang.
2. **Selectiebias:** bronnen zijn vooral officiële en semi‑officiële publicaties; onafhankelijke analyses missen mogelijk.
3. **Stopwoord‑beleid:** behoud van alle woorden kan ruis in de retrieval verhogen bij generieke vragen.
4. **LDA‑clustering:** onderwerpen kunnen zelden volledige semantische overlap vatten, vooral bij meervoudige thema’s.
5. **RAG‑afhankelijkheid:** externe API’s en vector‑databases vormen single points of failure. Single points of failure betekent dat als één onderdeel uitvalt, zoals een API of database, het hele systeem stopt met werken.
6. **Gebruikerstesten:** kleinschalig en binnen één gebruikersgroep uitgevoerd.

## 5.3 Vergelijking met bestaande literatuur

In verhouding tot ander onderzoek valt op dat RAG-studies zich richten op commerciële of wetenschappelijke corpora, terwijl onze toepassing zich specifiek richt op beleidscontext en subsidies. Soudani et al. (2024) vergelijken in *Fine Tuning vs. Retrieval Augmented Generation for Less Popular Knowledge* methoden (–FT–RAG, +FT–RAG, –FT+RAG, +FT+RAG) en concludeert dat RAG vooral uitblinkt bij minder frequente domeinkennis, terwijl de combinatie van fine-tuning en RAG de beste resultaten oplevert voor kleinere modellen. Onze opdracht bouwt hierop voort door de nadruk te leggen op handmatige dataverzameling, clustering met LDA en directe bruikbaarheid voor eindgebruikers binnen het duurzaamheidsdomein.

## 5.4 Juridische en ethische implicaties

Er is hier nog geen start aan gemaakt, maar hier zijn de mogelijke opties die we kunnen uitvoeren voor onze chatbot.

* **Privacy (AVG-compliance):** We slaan gesprekken alleen op als ze anoniem zijn en verwijderen ze na een bepaalde tijd.
* **Duidelijkheid (Transparantie):** In de chatbot en handleiding leggen we duidelijk uit dat de chatbot alleen informatie geeft, en geen juridisch advies.
* **Eerlijke antwoorden (Bias-monitoring):** We controleren regelmatig of de chatbot geen eenzijdige of verkeerde antwoorden geeft.
* **Verantwoordelijkheid (Aansprakelijkheid):** We plaatsen een waarschuwing dat mensen zelf verantwoordelijk zijn en schakelen een expert in als de chatbot het antwoord niet zeker weet.

# Literatuurlijst

- *Soudani, H., Kanoulas, E., & Hasibi, F. (2024). Fine tuning vs. retrieval augmented generation for less popular knowledge. arXiv.* [*https://arxiv.org/abs/2403.01432*](https://arxiv.org/abs/2403.01432)

- *Successen en valkuilen bij burgerparticipatie in de energietransitie | RIVM. (n.d.).* [*https://www.rivm.nl/publicaties/successen-en-valkuilen-bij-burgerparticipatie-in-energietransitie*](https://www.rivm.nl/publicaties/successen-en-valkuilen-bij-burgerparticipatie-in-energietransitie)

*Een studie van het Rijksinstituut voor Volksgezondheid en Milieu (RIVM), waarin de rol en het belang van burgerparticipatie in de energietransitie wordt belicht. Zoektermen “waarom is burgerparticipatie belangrijk energietransitie”*

- *What is an LLM (large language model)? | Cloudflare. (n.d.).* [*https://www.cloudflare.com/learning/ai/what-is-large-language-model/*](https://www.cloudflare.com/learning/ai/what-is-large-language-model/)

*Cloudflare is een internationaal actief techbedrijf dat veel kennis heeft van AI-modellen en netwerktechnologie; dit artikel geeft een heldere uitleg over LLM’s. Zoektermen “What is a LLM”*

- *Ministerie van Infrastructuur en Waterstaat. (2024, August 22). Klimaatbeleid. Klimaatverandering | Rijksoverheid.nl.* [*https://www.rijksoverheid.nl/onderwerpen/klimaatverandering/klimaatbeleid*](https://www.rijksoverheid.nl/onderwerpen/klimaatverandering/klimaatbeleid)

*Officiële overheidsinformatie over de klimaatdoelen en het beleid dat Nederland voert. Zoekterm “Klimaatbeleid”*

- *Ministerie van Algemene Zaken. (2024, December 6). Nederland gaat stap voor stap over op duurzame energie. Duurzame Energie | Rijksoverheid.nl.* [*https://www.rijksoverheid.nl/onderwerpen/duurzame-energie/nederland-gaat-stap-voor-stap-over-op-duurzame-energie*](https://www.rijksoverheid.nl/onderwerpen/duurzame-energie/nederland-gaat-stap-voor-stap-over-op-duurzame-energie)

*Een publicatie van de Rijksoverheid met specifieke informatie over de stappen richting een duurzame energievoorziening. Zoekterm “Duurzame energie overheid”*

- *Van Hulst, W. (n.d.). Hernieuwbare energie - Europa decentraal. Europa Decentraal.* [*https://europadecentraal.nl/onderwerp/klimaat-en-milieu/energie/hernieuwbare-energie/*](https://europadecentraal.nl/onderwerp/klimaat-en-milieu/energie/hernieuwbare-energie/)

*Europa Decentraal is een kenniscentrum dat zich richt op Europees recht en beleid voor decentrale overheden. Zoektermen “hernieuwbare energie europa”*

-  *Nederland en het klimaat: de Europese Green Deal. (n.d.). Vertegenwoordiging in Nederland Van De Europese Commissie.* [*https://netherlands.representation.ec.europa.eu/strategie-en-prioriteiten/belangrijke-eu-themas-voor-nederland/nederland-en-het-klimaat-de-europese-green-deal\_nl*](https://netherlands.representation.ec.europa.eu/strategie-en-prioriteiten/belangrijke-eu-themas-voor-nederland/nederland-en-het-klimaat-de-europese-green-deal_nl)

*Deze bron levert inzichten in de opvattingen en betrokkenheid van de Nederlandse bevolking bij klimaatbeleid, verzameld door de Europese Commissie. Zoektermen “Europese klimaatwet nederland”*

- *Beleid duurzame energie. (2024, September 5). RVO.nl.* [*https://www.rvo.nl/onderwerpen/beleid-duurzame-energie*](https://www.rvo.nl/onderwerpen/beleid-duurzame-energie)

*Rijksdienst voor Ondernemend Nederland (RVO) publiceert actuele informatie over subsidies en financieringsregelingen voor duurzame initiatieven. Zoektermen “Beleid duurzame energie”*

- *Verloop, D. (2024, August 13). LLMOps voor overheden: de sleutel tot verantwoorde AI-implementatie. CiviQs.* [*https://www.civiqs.nl/llmops-voor-overheden-de-sleutel-tot-verantwoorde-ai-implementatie*](https://www.civiqs.nl/llmops-voor-overheden-de-sleutel-tot-verantwoorde-ai-implementatie)

*Artikel over de methodiek van LLMOps, toegespitst op veiligheids- en ethische waarborgen bij AI-toepassingen in de publieke sector. Zoektermen “llm voor overheden”*

- *Febrian, G. F., & Figueredo, G. (2024, July 31). KemenkeuGPT: Leveraging a Large Language Model on Indonesia’s Government Financial Data and Regulations to Enhance Decision Making. arXiv.org.* [*https://arxiv.org/abs/2407.21459*](https://arxiv.org/abs/2407.21459?utm_source=chatgpt.com)

*Preprint van een wetenschappelijk artikel op arXiv.org. Het beschrijft een AI-case in Indonesië, waarbij een LLM is ingezet voor betere financiële analyses en beleidsbeslissingen. Zoektermen (*[*https://chatgpt.com/share/67b73e1e-4ab8-8013-a16c-d809d037fdfc*](https://chatgpt.com/share/67b73e1e-4ab8-8013-a16c-d809d037fdfc)*)*

*- Singh, M. G. (2024, December 24). To Use or Lose: Stop Words in NLP - Moirangthem Gelson Singh - Medium. Medium.* [*https://medium.com/%40gelsonm/to-use-or-lose-stop-words-in-nlp-de946edaa468*](https://medium.com/%40gelsonm/to-use-or-lose-stop-words-in-nlp-de946edaa468)

*- Chunking Strategies for LLM applications. (n.d.). Pinecone.* [*https://www.pinecone.io/learn/chunking-strategies/*](https://www.pinecone.io/learn/chunking-strategies/)

*Design. (2025) What is Latent Dirichlet allocation | IBM. Retrieved May 08, 2025, from* [*https://www.ibm.com/think/topics/latent-dirichlet-allocation*](https://www.ibm.com/think/topics/latent-dirichlet-allocation)