SSC-ICT
Ministerie van Binnenlandse Zaken en Koninkrijksrelaties

Innovatie SSC-ICT

RAG prototype

TLP GREEN

Innovatie team SSC-ICT

Martha Romkes

Daan Wijnhorst

Lara Mutsaers

Nicky Ju

Mark Heijnekamp

Ramon Fiedler

Victor Gevers

Waarom Learning Lion?

- GenAI verboden wanneer er juridische risico's aan verbonden zijn.
- Gebruik ChatGPT door rijksambtenaren -> actief datalek
- Behoefte aan inzet genAI binnen rijksoverheid

Meerderheid gemeenten gebruikt ChatGPT en bijna de helft weet niet wat hun medewerkers ermee doen - EenVandaag (avrotros.nl)

Use Case Kamervragen

Het beantwoorden van deze vragen is een tijdrovend proces. Wij willen onderzoeken in hoeverre generatieve AI-modellen zoals Copilot en LearningLion kamervragen effectief kunnen analyseren en beantwoorden.

Wat is Learning Lion

- **R**etrieval **A**ugmented **G**eneration
- Doelstelling van de PoC : experimenteren.
- Transparant
- On-premise

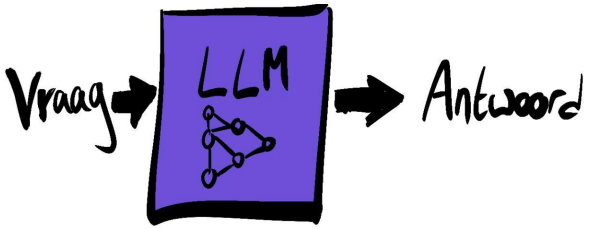
AI GPU rekenkracht

- On premise GPU server
- On premise AI Laptop

- Ervaring opdoen met beheer AI hardware en services ->
- AI infrastructuur als dienst door SCC-ICT

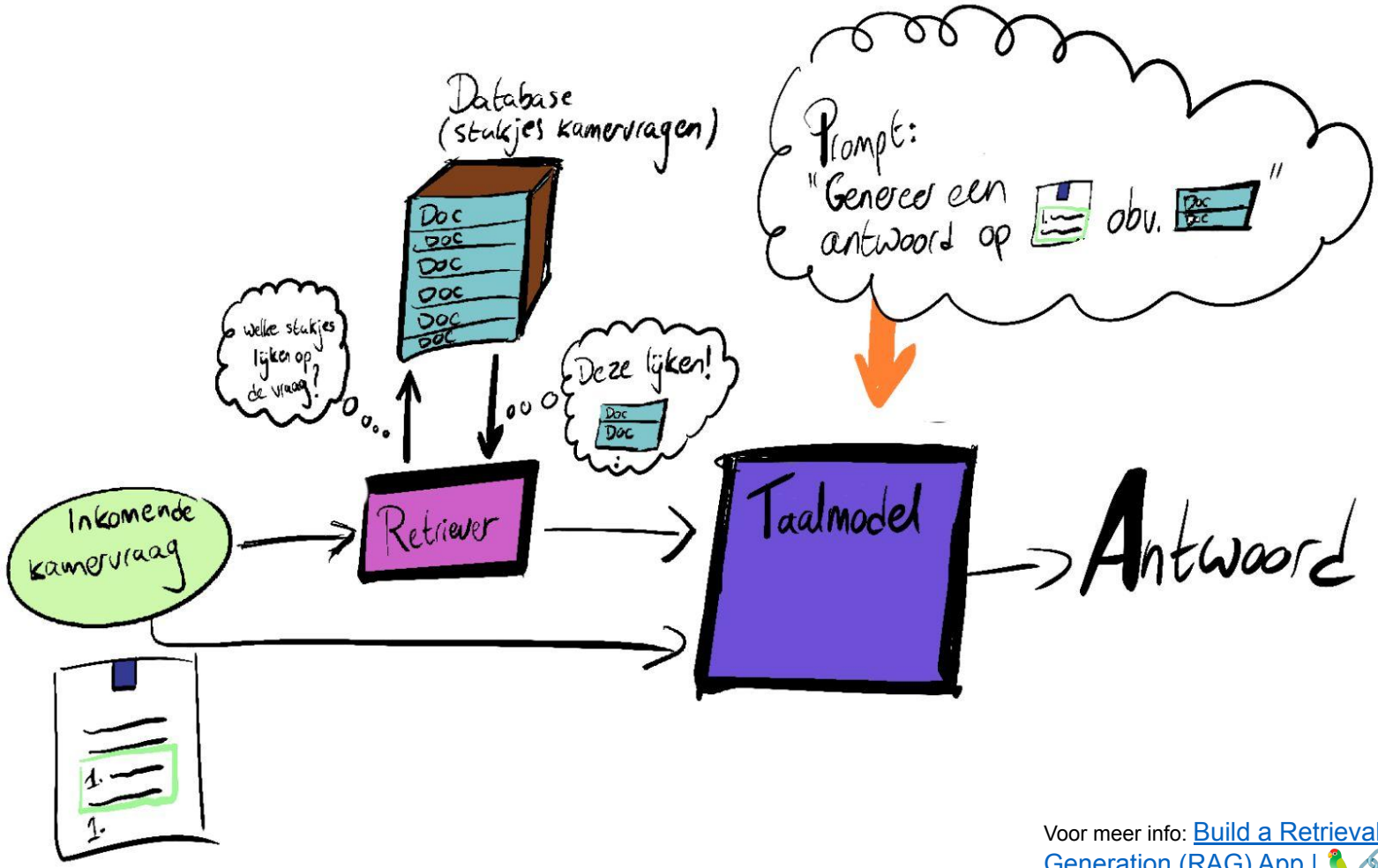
RAG en kamervragen

- **Retrieval:** Zoek (slim) in eigen data naar relevante stukjes tekst (oude kamervragen)
- **Generation:** Genereer obv. die relevante stukjes een antwoord / een respons volgens instructies
- **Omdat:**
  - Niet enkel parametrisch ‘geheugen’
  - Kosteneffectief
  - ‘Simpel’
  - Lokaal implementeerbaar
  - Pogen hallucinaties te verminderen



What is Retrieval Augmented Generation (RAG)? | Databricks

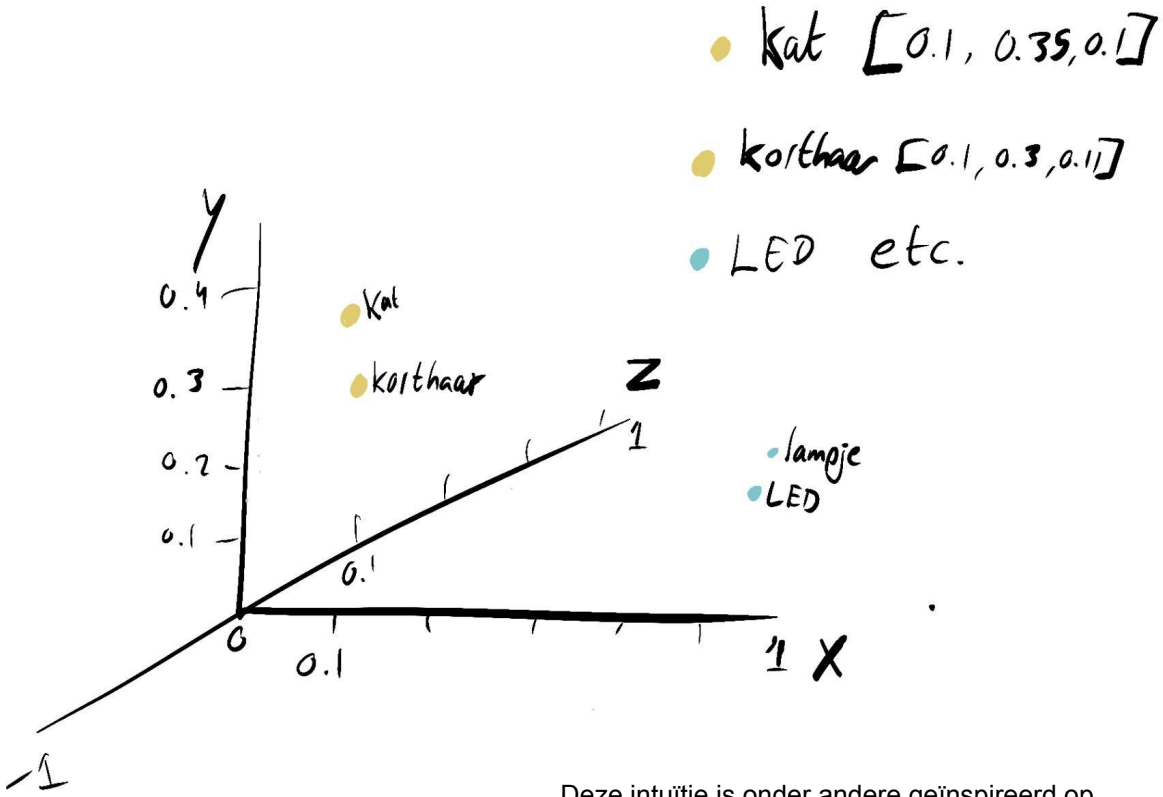
Overview: RAG-pipeline



Voor meer info: [Build a Retrieval Augmented Generation (RAG) App | 🦜🔗 LangChain]

Dieper: Retrieval (intuïtie)

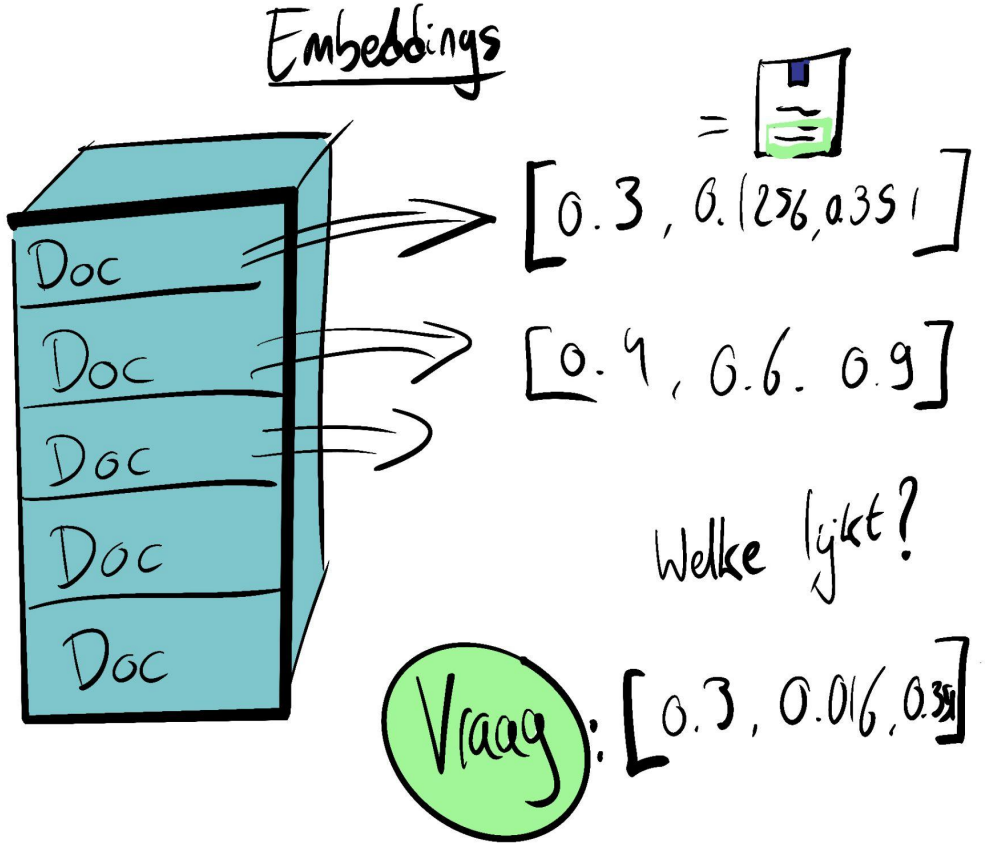
- kat, korthaar, viervoeter?



Deze intuïtie is onder andere geïnspireerd op Deciphering the Dimensions of Embeddings: A Journey into Semantic Spaces | by Anand | Medium.

Dieper: Retrieval (intuïtie)

- ‘Semantiek’ gevangen,
- Zoeken in de archiefkast
- Combineren keyword?
- Stochastisch



Implementatie

- **Vragen**
  - Data? Chunks? Modellen?
  - Architectuur? Toevoegingen?
  - Voorgebakken?

- **Principes:** Transparant, Toegankelijk, Flexibel
  - Streven: codebase simpel en uitlegbaar
  - Modulair: swappen van modellen (up-to-date)

Use Case WOO verzoeken

Wat is het probleem?

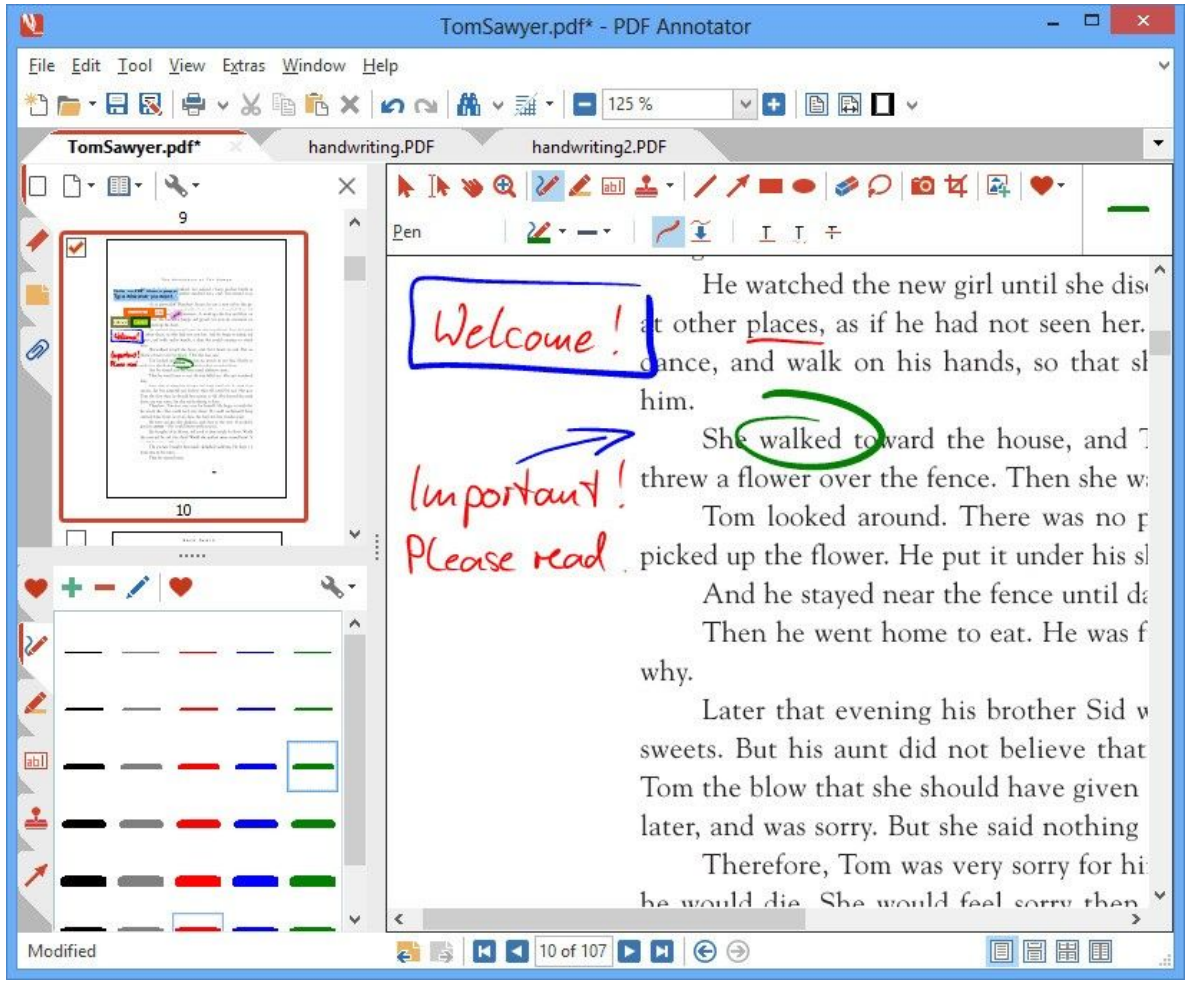
- Een groot aantal Woo-verzoeken wordt niet op tijd afgehandeld.

- Het afhandelen van een Woo-verzoek kost ZEER veel tijd:
  - Handmatig zoeken naar documenten/e-mails.
  - Handmatig controleren welke stukken relevant zijn.

RAG voor WOO

RAG gebruiken voor het versnellen van WOO verzoeken:

- Metadata Filteren met documentindeling, datum, doelgroep
- Rephrasing: onderzoeksvraag herformuleren en probeer het opnieuw
- Onnodige leestekens of stop woorden filteren



Samenvattend: drie basisprincipes

1. Bewaken van soevereiniteit
2. Experimenteel: het opdoen van kennis
3. Transparantie en openheid

Learninglion.nl

GitHub

Referenties

Meerderheid gemeenten gebruikt ChatGPT en bijna de helft weet niet wat hun medewerkers ermee doen - EenVandaag (avrotros.nl)

Overheidsbrede visie Generatieve AI

Introductie | Learning Lion

SSC-ICT-Innovatie/LearningLion-kamervragen: Project opensource generative AI (github.com)

What is Retrieval Augmented Generation (RAG)? | Databricks

Deciphering the Dimensions of Embeddings: A Journey into Semantic Spaces | by Anand | Medium.

Build a Retrieval Augmented Generation (RAG) App | 🦜🔗 LangChain

Fine-tune Embedding models for Retrieval Augmented Generation (RAG) (philschmid.de)