



Verantwortung für Deutschland

Koalitionsvertrag zwischen
CDU, CSU und SPD

21. Legislaturperiode

144 pages
4588 lines of text

ALEPH ALPHA – CASE STUDY

Wie RAG den öffentlichen Sektor mit Antworten versorgt

Max Lautenbach · 28. April 2025

WAS IST RAG?

LLMs haben keinen Zugriff auf Unternehmensspezifische Daten

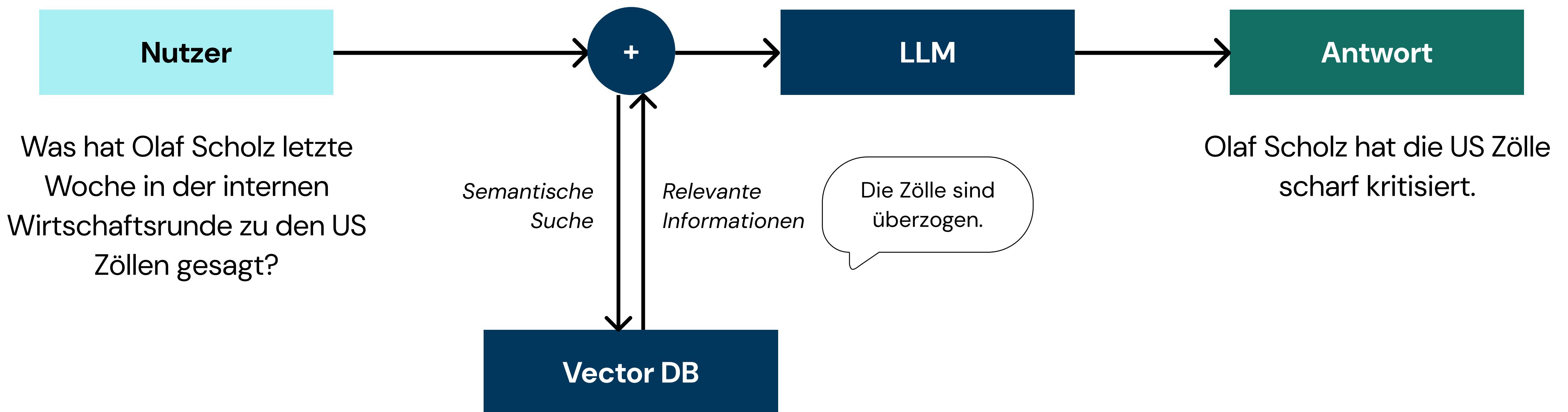


Was hat Olaf Scholz letzte
Woche in der internen
Wirtschaftsrunde zu den US
Zöllen gesagt?

Ich habe keine Informationen
dazu.

WAS IST RAG?

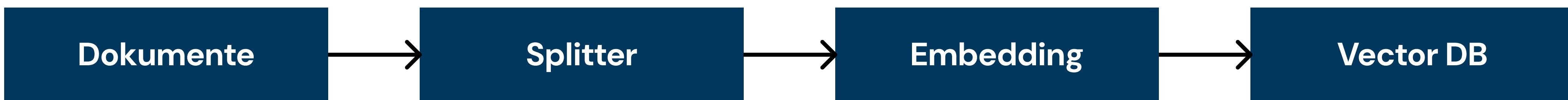
RAG kann LLMs Zugriff zu Unternehmensspezifische Daten herstellen



Demo

TECHNISCHE UMSETZUNG

Dokumente müssen zunächst vorverarbeitet werden



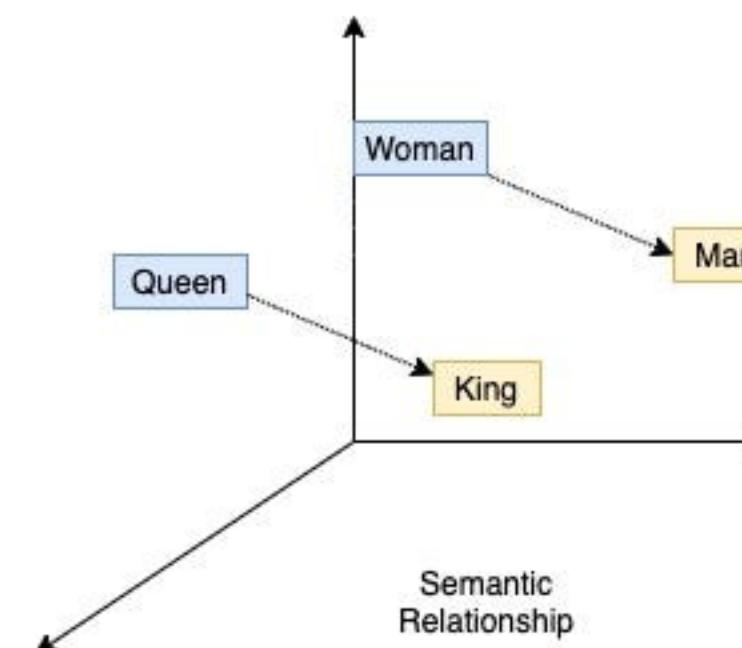
Unterstützte Dokumente

- PDF
- JSON
- Markdown
- Word
- HTML

Character Splitter
Chunk Größe: 512 Chars
Overlap: 50 Chars

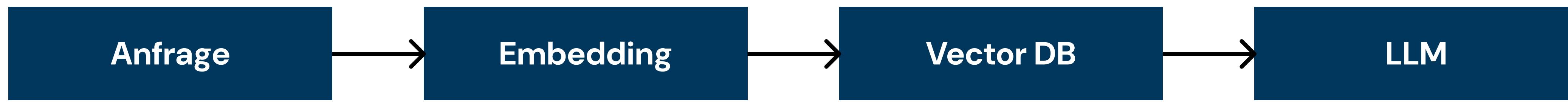
Embedding Model
bge-m3

Vector DB
Qdrant (via Docker)



TECHNISCHE UMSETZUNG

RAG kann end-to-end Open-Source umgesetzt werden



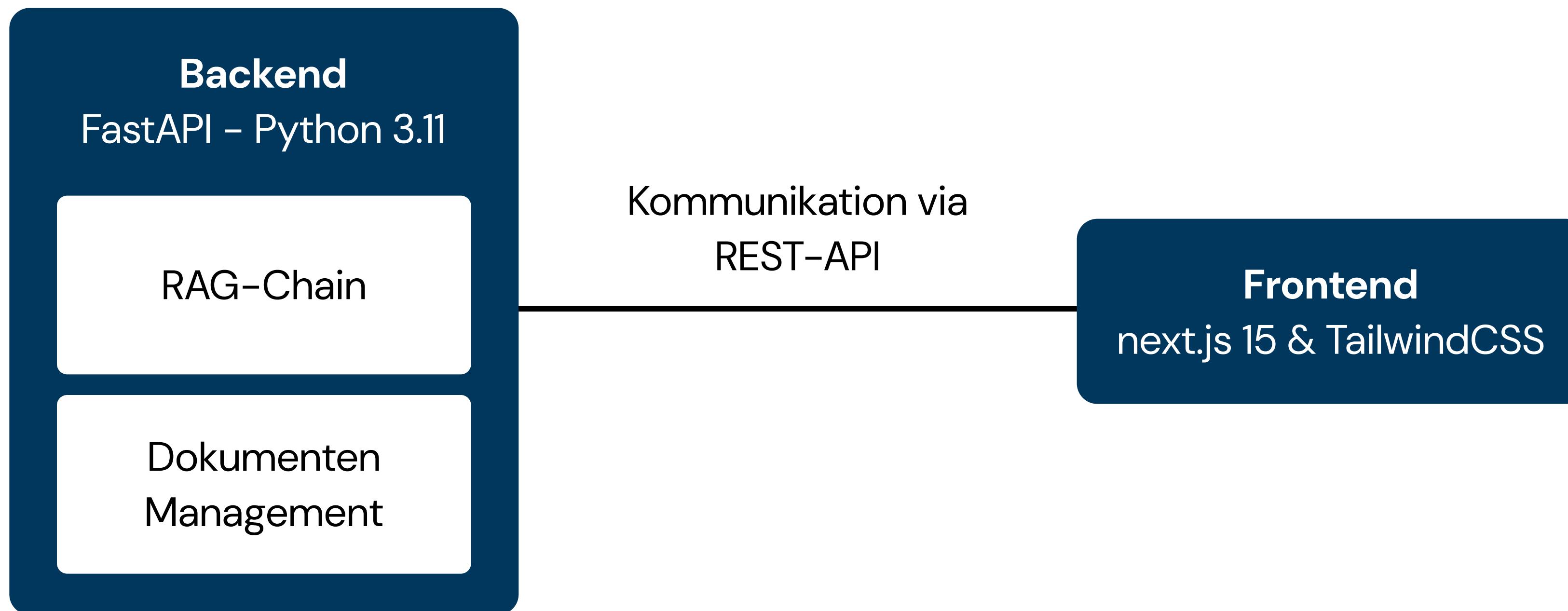
Embedding Model
bge-m3

Vector DB
Qdrant (via Docker)

LLM Model
Llama 4 Scout

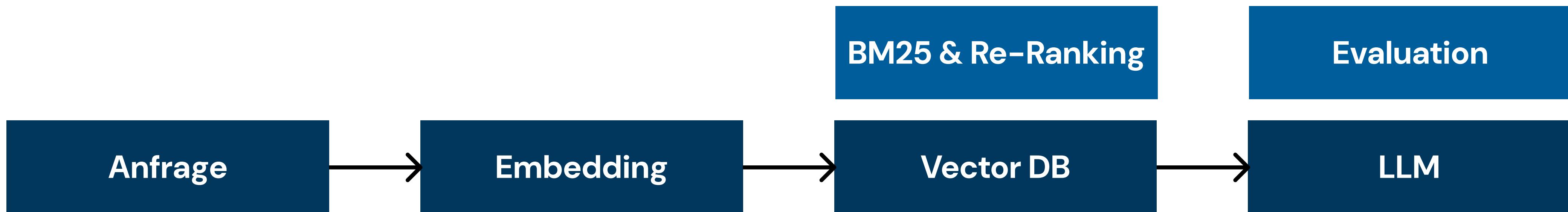
TECHNISCHE UMSETZUNG

Frontend und Backend können getrennt umgesetzt werden.



TECHNISCHE UMSETZUNG

RAG hat verschiedene Optimierungspunkte



Embedding Model

bge-m3

Vector DB

Qdrant (via Docker)

LLM Model

Llama 4 Scout

VORTEILE VON RAG

RAG kann kostengünstigere, aktuellere und relevantere Antworten geben als ein LLMS

①

Zugriff auf interne Daten

RAG-Chains können auf semantisch relevante interne Daten zugreifen.

②

Aktualisierung ohne Re-Trainieren

RAG-Chain können das Wissen eines Chatbots aktualisieren ohne das LLM anpassen zu müssen.

③

Kosteneffizienz

RAG-Chains sind meist kosteneffizienter als alternative Optionen.