# RAG ‘n Taxes

Knowledge-Assistant für die Steuererklärung

Adrian Höhn

Peter Fust

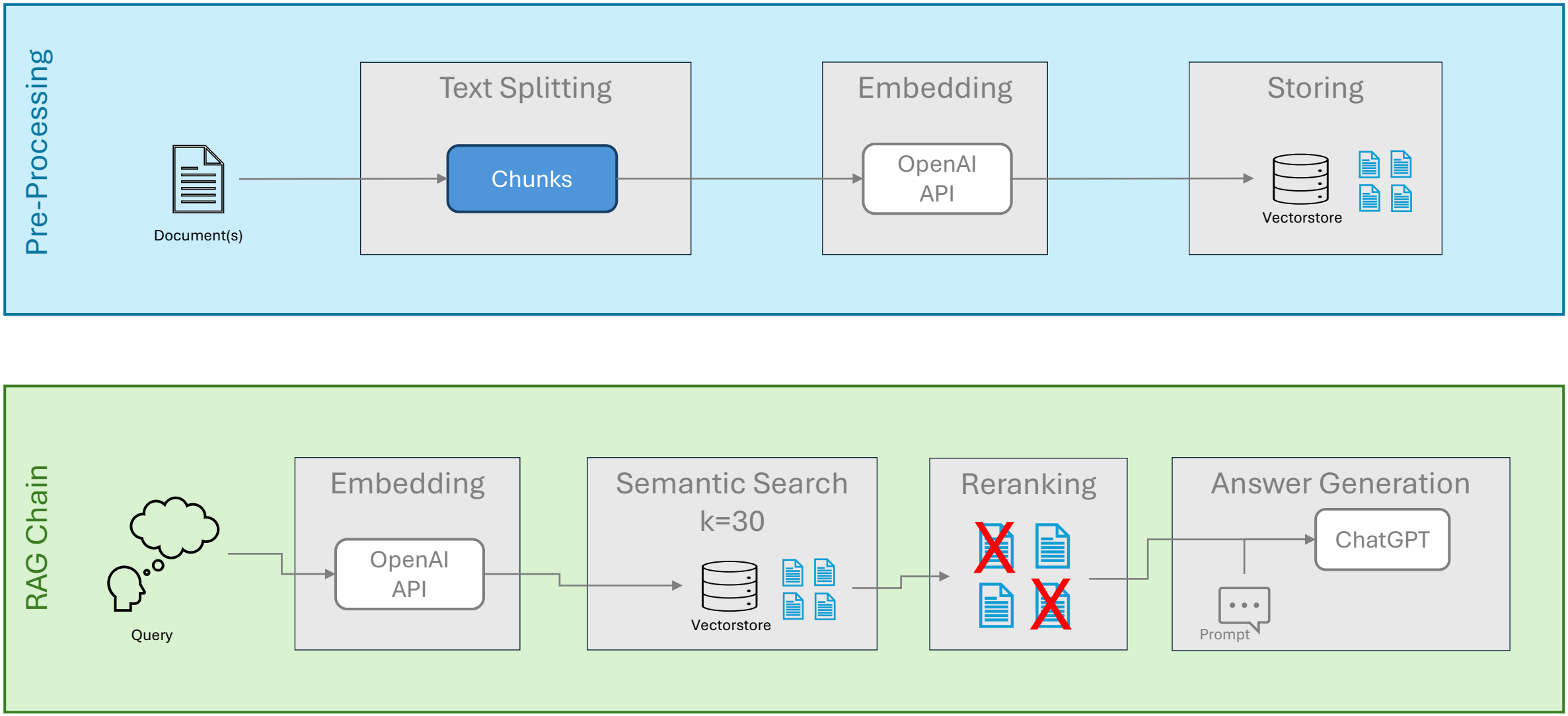
Juni 2024

# Ausgangslage & Ziel

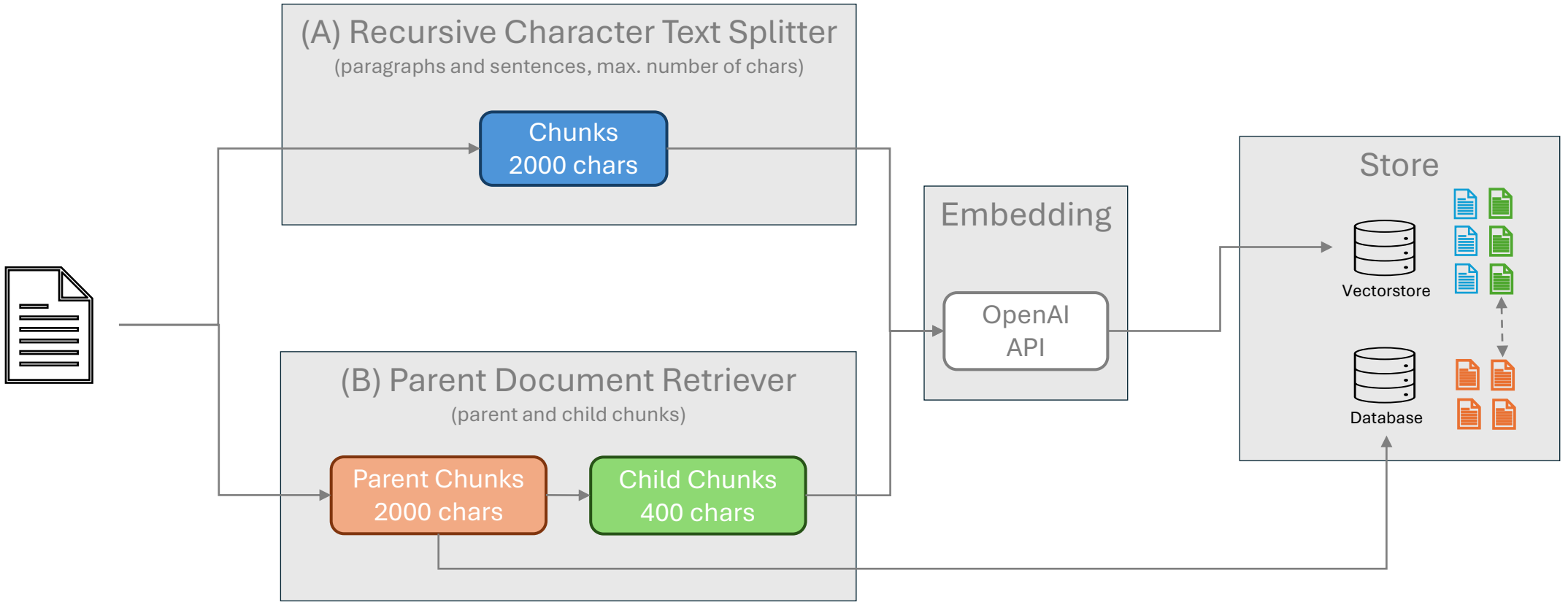
- RAG-basierter Knowledge-Assistant als Hilfe zum Ausfüllen der Steuererklärung
- Langchain als Framework kennenlernen

- Grundlage ist die Wegleitung zur Steuererklärung Kanton SG
- PDF-Dokument mit 52 Seiten, inkl. Tabellen

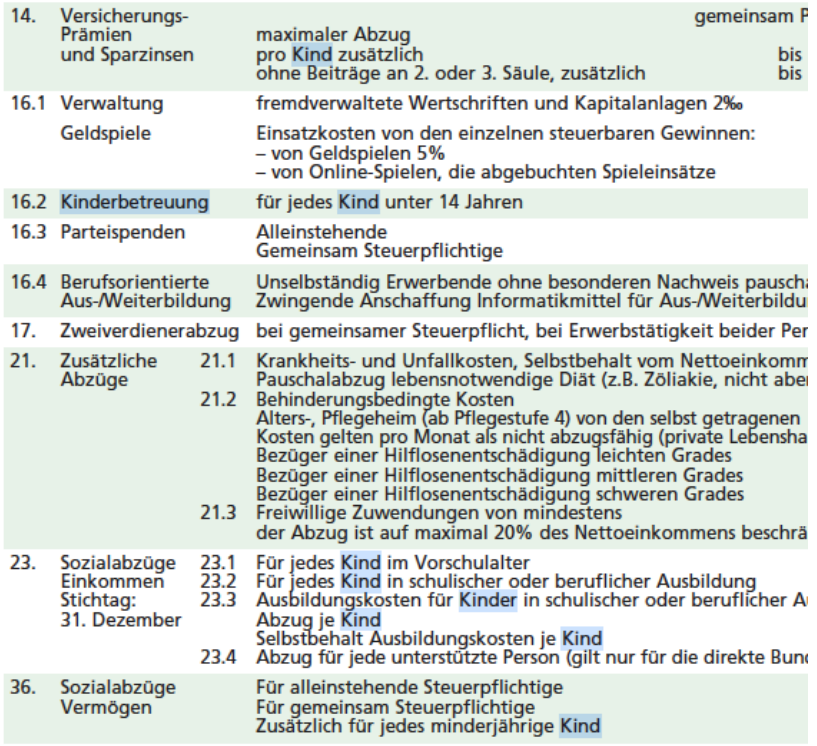
# Big Picture


Pre-Processing
Document(s)
Text Splitting
Chunks
Embedding
OpenAI API
Storing
Vectorstore
RAG Chain
Query
Embedding
OpenAI API
Semantic Search k=30
Vectorstore
Reranking
Answer Generation
ChatGPT
Prompt

# Text Splitting



# Text Splitting - Erkenntnisse



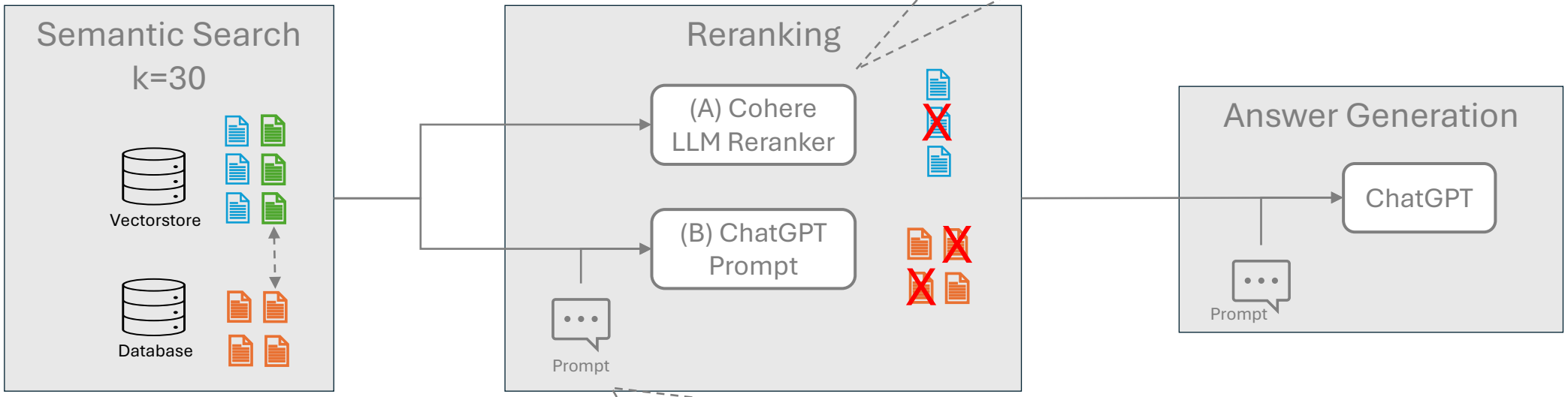
| | | | |
|---|---|---|---|
| 14. | Versicherungs-Prämien und Sparzinsen | | gemeinsam P<br>maximaler Abzug<br>pro Kind zusätzlich bis<br>ohne Beiträge an 2. oder 3. Säule, zusätzlich bis |
| 16.1 | Verwaltung | | fremdverwaltete Wertschriften und Kapitalanlagen 2‰ |
| | Geldspiele | | Einsatzkosten von den einzelnen steuerbaren Gewinnen:<br>– von Geldspielen 5%<br>– von Online-Spielen, die abgebuchten Spieleinsätze |
| 16.2 | Kinderbetreuung | | für jedes Kind unter 14 Jahren |
| 16.3 | Parteispenden | | Alleinstehende<br>Gemeinsam Steuerpflichtige |
| 16.4 | Berufsorientierte Aus-/Weiterbildung | | Unselbständig Erwerbende ohne besonderen Nachweis pausch<br>Zwingende Anschaffung Informatikmittel für Aus-/Weiterbildu |
| 17. | Zweiverdienerabzug | | bei gemeinsamer Steuerpflicht, bei Erwerbstätigkeit beider Per |
| 21. | Zusätzliche Abzüge | 21.1 | Krankheits- und Unfallkosten, Selbstbehalt vom Nettoeinkomm<br>Pauschalabzug lebensnotwendige Diät (z.B. Zöliakie, nicht abe |
| | | 21.2 | Behinderungsbedingte Kosten<br>Alters-, Pflegeheim (ab Pflegestufe 4) von den selbst getragenen<br>Kosten gelten pro Monat als nicht abzugsfähig (private Lebensha<br>Bezüger einer Hilflosenentschädigung leichten Grades<br>Bezüger einer Hilflosenentschädigung mittleren Grades<br>Bezüger einer Hilflosenentschädigung schweren Grades |
| | | 21.3 | Freiwillige Zuwendungen von mindestens<br>der Abzug ist auf maximal 20% des Nettoeinkommens beschrä |
| 23. | Sozialabzüge<br>Einkommen<br>Stichtag:<br>31. Dezember | 23.1<br>23.2<br>23.3 | Für jedes Kind im Vorschulalter<br>Für jedes Kind in schulischer oder beruflicher Ausbildung<br>Ausbildungskosten für Kinder in schulischer oder beruflicher A<br>Abzug je Kind<br>Selbstbehalt Ausbildungskosten je Kind |
| | | 23.4 | Abzug für jede unterstützte Person (gilt nur für die direkte Bun |
| 36. | Sozialabzüge Vermögen | | Für alleinstehende Steuerpflichtige<br>Für gemeinsam Steuerpflichtige<br>Zusätzlich für jedes minderjährige Kind |

- Similarity Threshold schwierig zu bestimmen
  - 30 Chunks zwischen 0.32 und 0.35
- Parent Document Retriever bzw. kleinere Chunks finden semantisch feinere Informationen
  - Einzelne Wörter im Inhaltverzeichnis werden gefunden
  - Bester similarity score ist 0.2882 vs. 0.3266 bei Text Splitting
  - Ob Miteinbezug von mehr Kontext relevant ist hängt von Dokumentenstruktur ab

- Die Qualität der finalen Antwort auf das Query hat es aber nicht massgeblich beeinflusst
- PDR tendenziell etwas genauer, da weniger Chunks ans LLM gelangen

# Reranking

```
index=2 relevance_score=0.6888872 - ../raw_data/Wegleitung_NP_2023.pdf__12
index=5 relevance_score=0.32863715 - ../raw_data/Wegleitung_NP_2023.pdf__13
index=3 relevance_score=0.123894654 - ../raw_data/Wegleitung_NP_2023.pdf__30
index=9 relevance_score=0.0894546 - ../raw_data/Wegleitung_NP_2023.pdf__27
index=0 relevance_score=0.080069266 - ../raw_data/Wegleitung_NP_2023.pdf__30
index=18 relevance_score=0.022414217 - ../raw_data/Wegleitung_NP_2023.pdf__19
```




```
Bitte bewerte die Relevanz jedes der oben aufgeführten Dokumente im Hinblick auf die gestellte Frage.
Berücksichtige dabei, wie gut jedes Dokument die Frage beantwortet, ob es relevante Informationen
enthält und ob diese Informationen korrekt und vertrauenswürdig erscheinen.
Die einzelnen Dokumente sind mit ===START=== und ===ENDE=== markiert. Der Name des Dokuments steht
gleich nach dem Marker ===START=== und ist mit ==NAME_START und ==NAME_END== markiert. Danach folgt
der Inhalt des Dokuments.
Gib als Antwort ein JSON Array zurück mit einer Liste von JSON Objekten. Jedes Objekt enthält die
folgenden Felder:
- name: Name des Dokuments
- relevant: true, wenn das Dokument relevant ist, false, wenn nicht
```

# Reranking - Erkenntnisse

- Cohere
  - Threshold?
  - x Dokumente?

document=None index=2 relevance_score=0.6888872 - ../raw_data/Wegleitung_NP_2023.pdf__12
document=None index=5 relevance_score=0.32863715 - ../raw_data/Wegleitung_NP_2023.pdf__13
document=None index=3 relevance_score=0.123894654 - ../raw_data/Wegleitung_NP_2023.pdf__30
document=None index=9 relevance_score=0.0894546 - ../raw_data/Wegleitung_NP_2023.pdf__27
document=None index=0 relevance_score=0.080069266 - ../raw_data/Wegleitung_NP_2023.pdf__30

- ChatGPT
  - Kein Ranking
  - Deklariert mehr Dokumente relevant als Cohere (mit 0.5)

- Kein klares Bild
  - Kleiner relevanter Abschnitt in nur einem Dokument („Heirat“)
    - Cohere: 0.0038 => nicht relevant
    - ChatGPT 3 relevante Dokumente
  - Viele relevante Dokumente („Arbeitskosten“)
    - Cohere: 12 relevant Dokumente (0.99 – 0.62)
    - ChatGPT: 9 relevante Dokumente

# Summary

- **Qualitätskontrolle der Chain sehr aufwendig**
  - Manuell, Spezialisten Know-How nötig
  - Viele Tuning-Möglichkeiten

- Generischer Chain-Ansatz wohl eher ungeeignet
  - Dokumenten-Art, Tabellen, etc.
  - Verschiedene Chain-Architekturen möglich

- Langchain Framework bedingt etwas Einarbeitung, bietet aber eine sehr einfache Anwendung von vielen Standard RAG Tasks:
  - Document Loading (PDF, Markdown, HTML, TXT, etc.)
  - Text Splitting
  - VectorDB Anbindung
  - Semantic Search
  - LLM calls