



Berner Fachhochschule
Haute école spécialisée bernoise
Bern University of Applied Sciences



Umsetzung der Chatbot-Lösung

Technische Sicht

► Institut Public Sector Transformation

Vorteile von Open Source AI

1. Datenschutz
2. Digitale Souveränität
3. Transparentere Preisgestaltung
4. Wettbewerbsfähigkeit

Retrieval-Augmented Generation (RAG)

- ▶ RAG ist eine Technik, mit der Sprachmodelle durch **externe Wissensdatenbanken** erweitert werden, was präzisere Antworten ermöglicht
- ▶ Folgender Prozess wird durchlaufen:
 - 1) Benutzer*in stellt eine Frage
 - 2) Eine Suche wird gestartet, welche die passendsten Textpassagen zu dieser Frage zurückgeben (**Retrieval**)
 - 3) Die Textpassagen werden zusammen mit der Frage in einem Prompt für ein Sprachmodell kombiniert (**Augmentation**)
 - 4) Das Sprachmodell liest den Prompt und generiert eine passende Antwort (**Text Generation**)

Wissensdatenbank

- ▶ Wissensdatenbank = **Sammlung von Dokumenten**
- ▶ Diese werden in **einzelne Textpassagen (Chunks)** aufgeteilt
- ▶ Die Wissensdatenbank für diesen Chatbot umfasste **296 Dokumente**
- ▶ Dies entspricht 3952 Seiten oder 1'289'103 Wörtern
- ▶ Dokumente wurden in Textpassagen mit **jeweils 400 Wörtern** unterteilt (Überlappung: 50 Wörter)
- ▶ Dadurch entstanden **3522 Chunks**



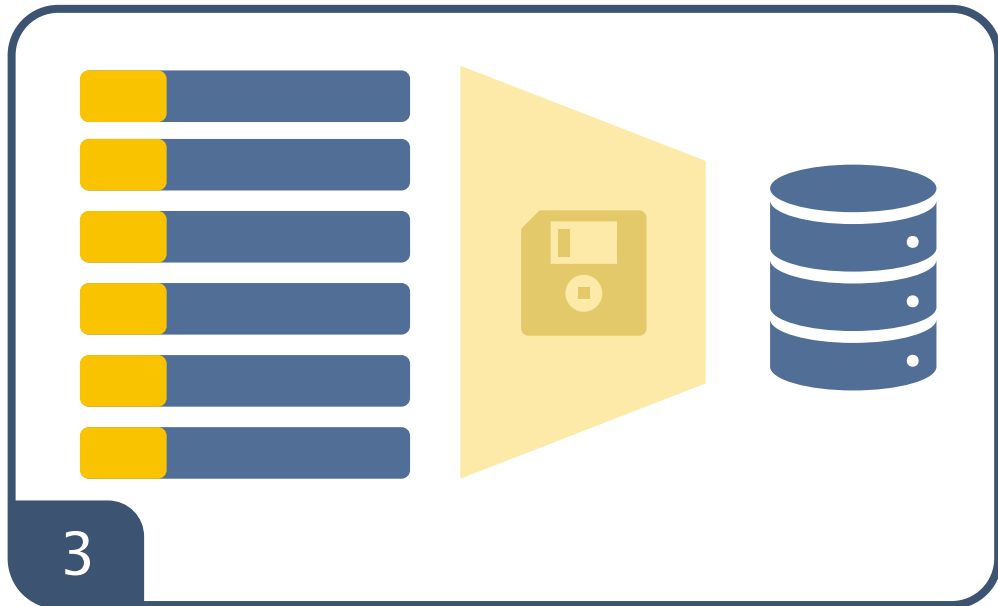
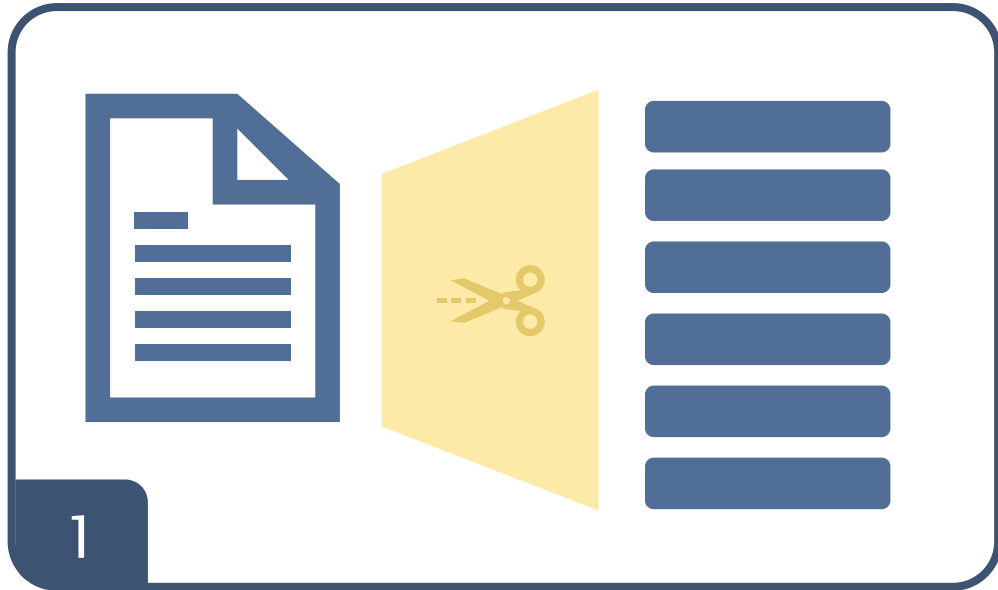
Contextual Retrieval

- ▶ Methode, die von Anthropic vorge stellt wurde, um die einzelnen Chunks durch **kontextuelle Informationen** zu erweitern
- ▶ Jeder Chunk wird **zusammen mit dem gesamten Dokument** an ein Sprachmodell übergeben, welches dem Chunk Zusatzinformationen hinzufügt

```
original_chunk = "Der Umsatz des Unternehmens stieg im Vergleich zum vorherigen Quartal um 3 %."  
  
contextualized_chunk = "Dieser Abschnitt stammt aus einer SEC-Einreichung zur Leistung der ACME Corp im zweiten Quartal 2023; der Umsatz des vorherigen Quartals betrug 314 Millionen US-Dollar. Der Umsatz des Unternehmens stieg im Vergleich zum vorherigen Quartal um 3 %."
```

Contextual Retrieval

- ▶ **Herausforderung:** Das Sprachmodell, welches wir verwendet haben, erlaubt eine maximale Textlänge von 128'000 Tokens
- ▶ Einige Dokumente waren jedoch länger
- ▶ **Lösung:** Adaptierte Version von Contextual Retrieval basierend auf **sechs Chunks** (drei vor dem aktuellen und drei nach dem aktuellen Chunk), sowie einer **Zusammenfassung** des Dokuments und **wichtiger Restriktionen** basierend auf den ersten 10 Chunks.



Hier sind die **ersten paar Seiten** eines Dokuments:

```
<excerpt>  
{{ excerpt }}  
</excerpt>
```

Bitte schreibe eine kurze Zusammenfassung über den Inhalt des Dokuments, auch wenn du **nicht das ganze Dokument gesehen hast**. Du solltest in der Lage sein, anhand den ersten paar Seiten zu erkennen, was der Inhalt des gesamten Dokuments ungefähr ist. Halte dich kurz aber sei präzise. Ausserdem solltest du einen weiteren Text schreiben, welcher die Einschränkungen des Dokuments beinhaltet. Wenn sich ein Dokument beispielsweise **nur auf eine gewisse Personengruppe, einen bestimmten Zeitraum oder eine bestimmte Bedingung oder Grundvoraussetzung bezieht**, dann sollte dies in diesem kurzen Text festgehalten werden. Bitte antworte wie folgt:

Summary: <Die **Zusammenfassung, worum es in dem Dokument geht**>

Restrictions: <Die **Einschränkungen, die auf dieses Dokument zutreffen**>

Antworte nebst diesen beiden Präfixen und den entsprechenden Texten nichts anderes!

Im Folgenden ein längerer Ausschnitt aus einem Dokument:

{{ excerpt }}

Zusätzlich wissen wir, dass das gesamte Dokument folgenden Inhalt hat:

{{ summary }}

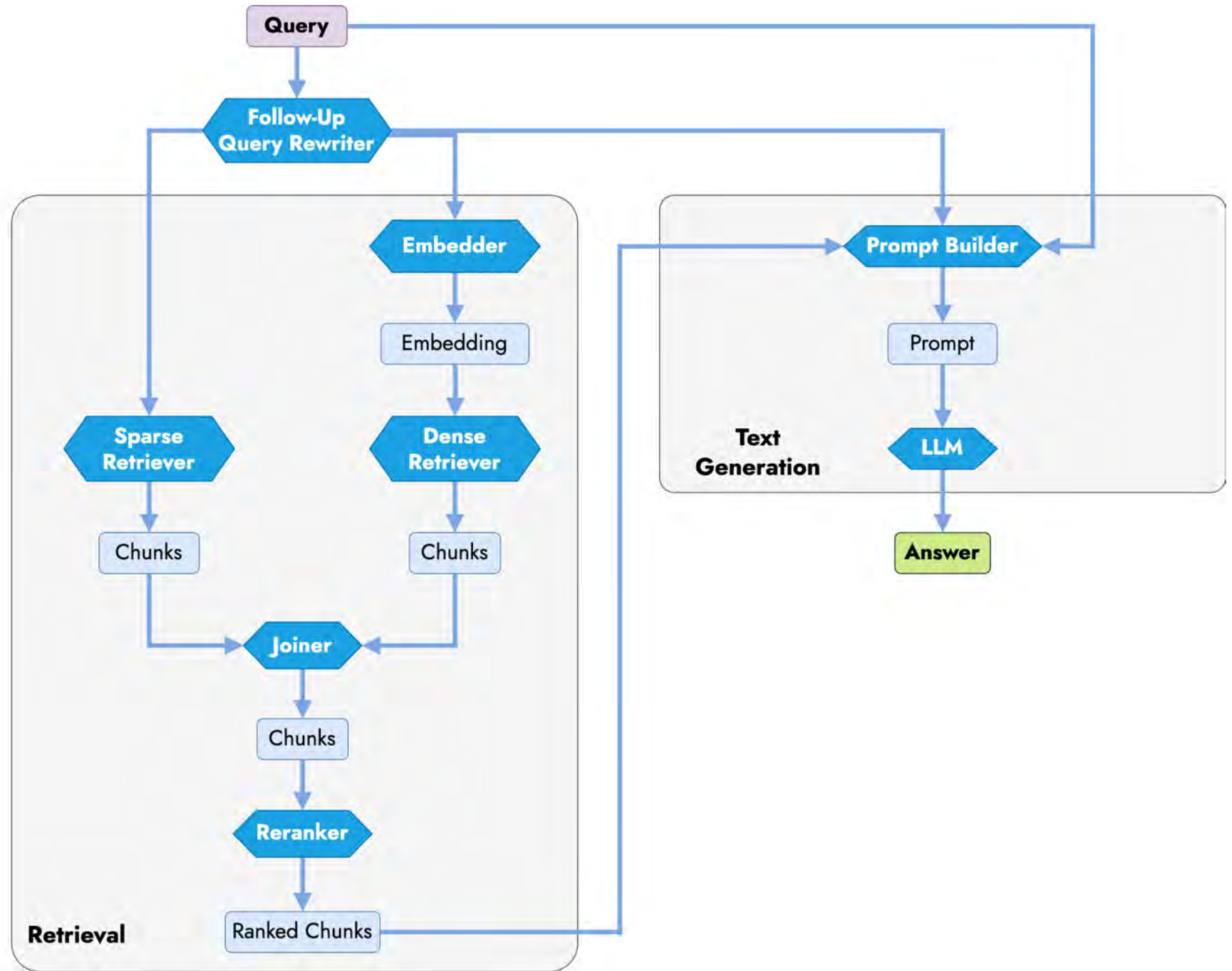
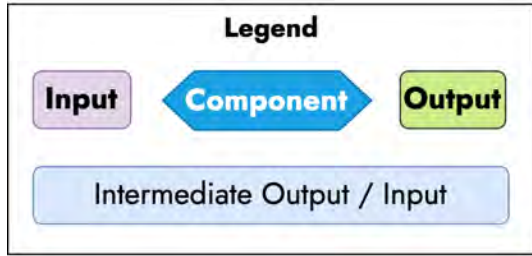
Wir wissen auch, dass folgende Einschränkungen berücksichtigt werden müssen:

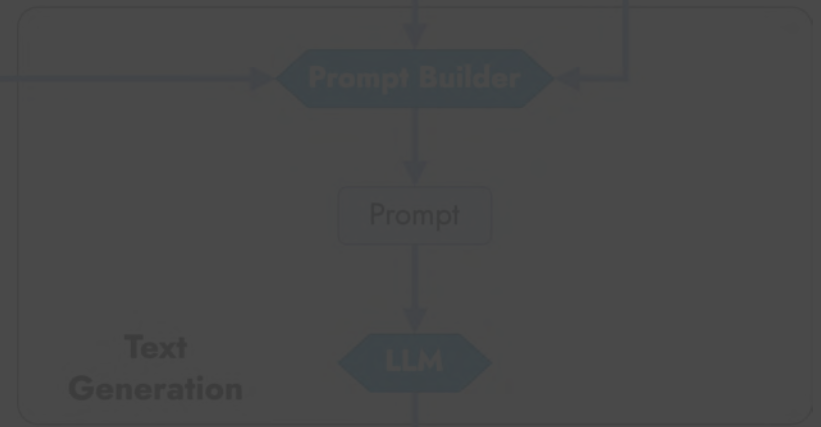
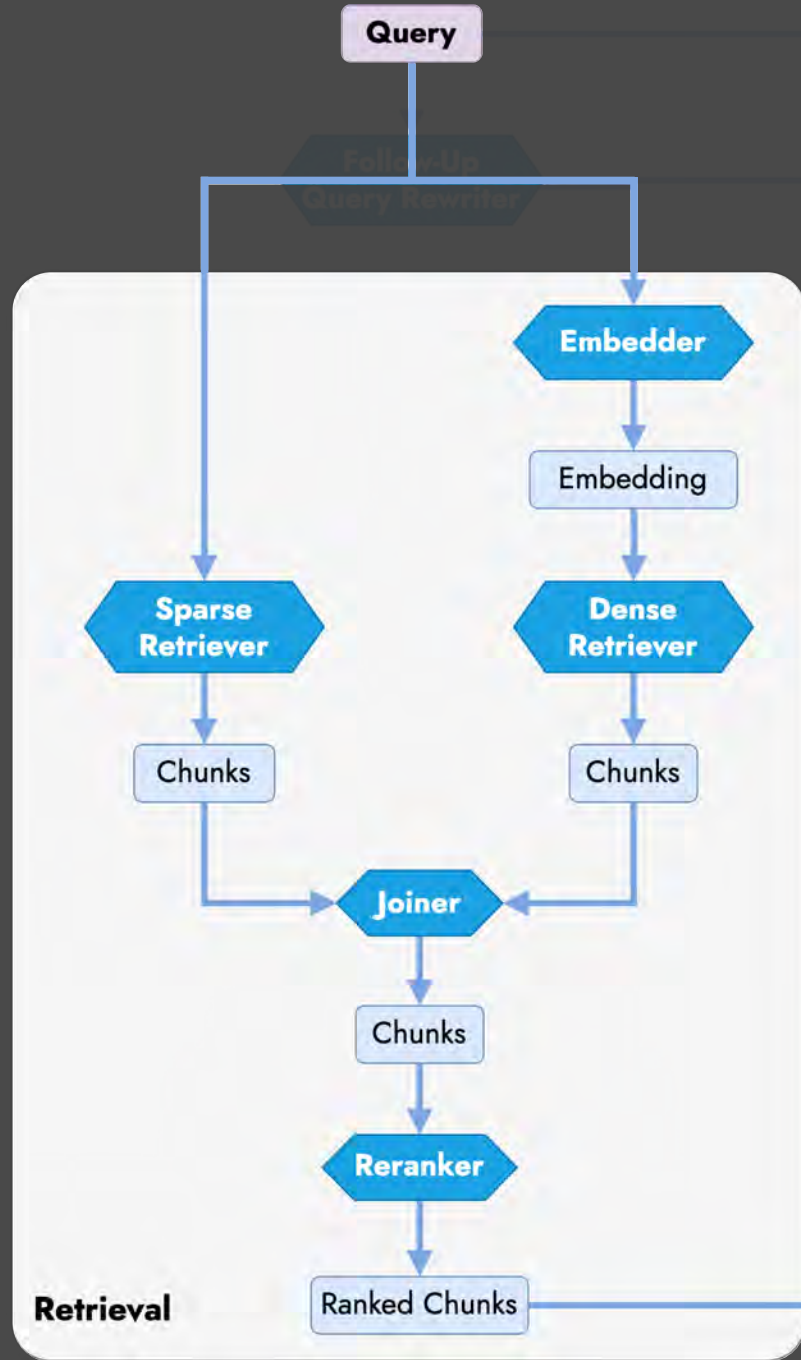
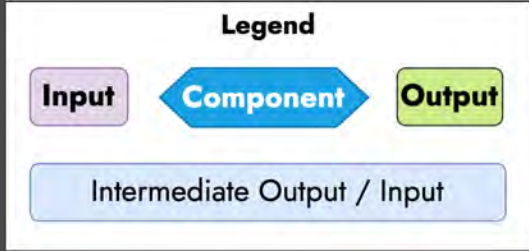
{{ restrictions }}

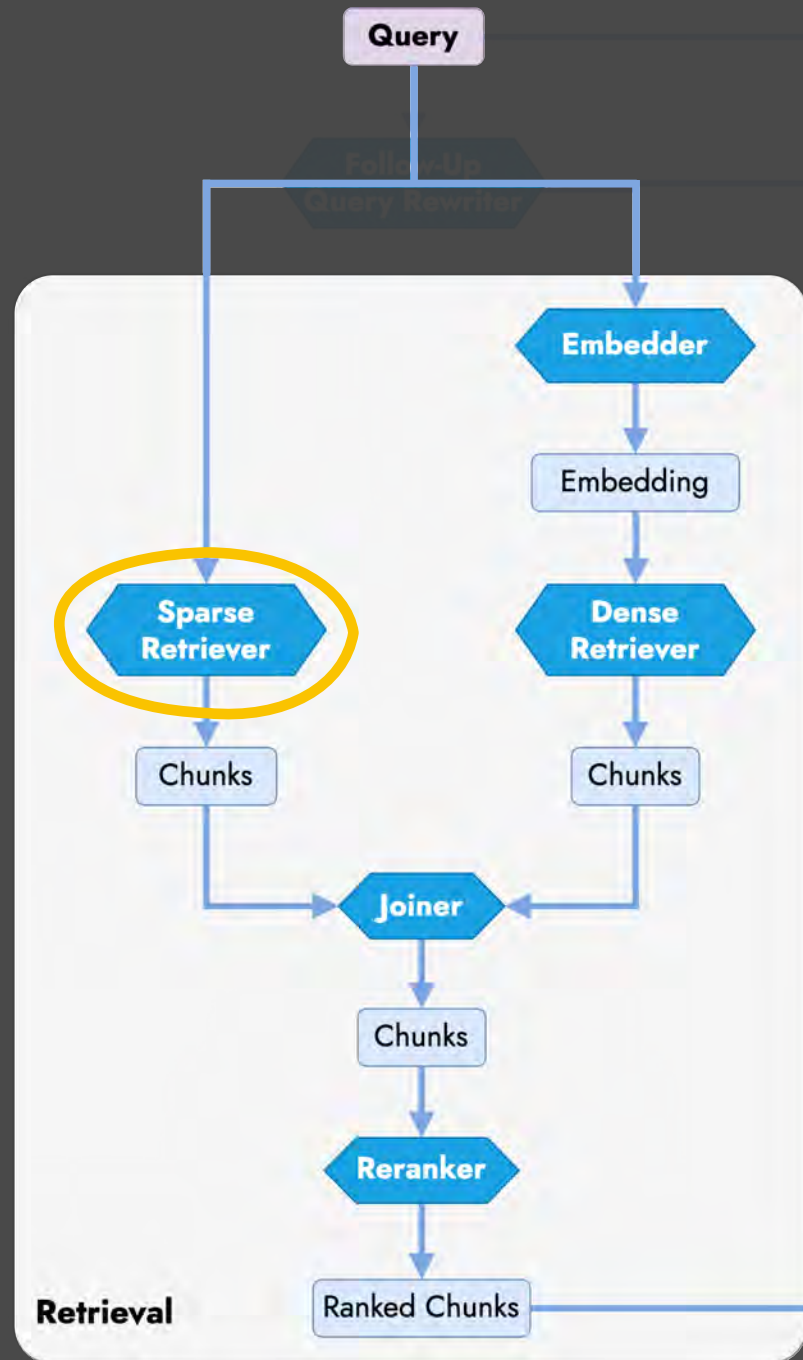
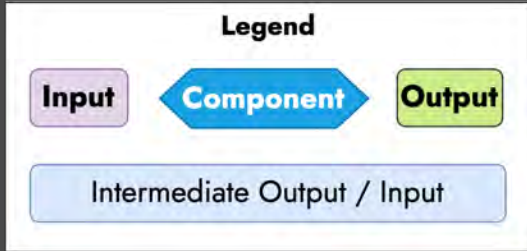
Bitte schreibe basierend auf diesen Informationen einen kurzen Text, welcher als Kontext dient, um diesen Textausschnitt **innerhalb des Gesamtdokuments zu situieren**. Der Kontext sollte genug Informationen enthalten, dass beim Lesen des Ausschnitts klar wird, **auf was sich der Text bezieht und wie dieser verstanden werden soll**. Antworte nur mit dem Kontext und sonst nichts.

Retrieval

- ▶ **Wichtigster Part** des Chatbots: Ohne die richtigen Textpassagen, keine korrekte Antwort
- ▶ Für das Auffinden von passenden Textstellen wird i.d.R. eine der folgenden Methoden verwendet:
 - ❑ **Sparse Retrieval:** Stichwortsuche
 - ❑ **Dense Retrieval:** Vektor-Embeddings (Semantische Ähnlichkeit)
 - ❑ **Hybrid Retrieval:** Kombination der oben genannten Methoden

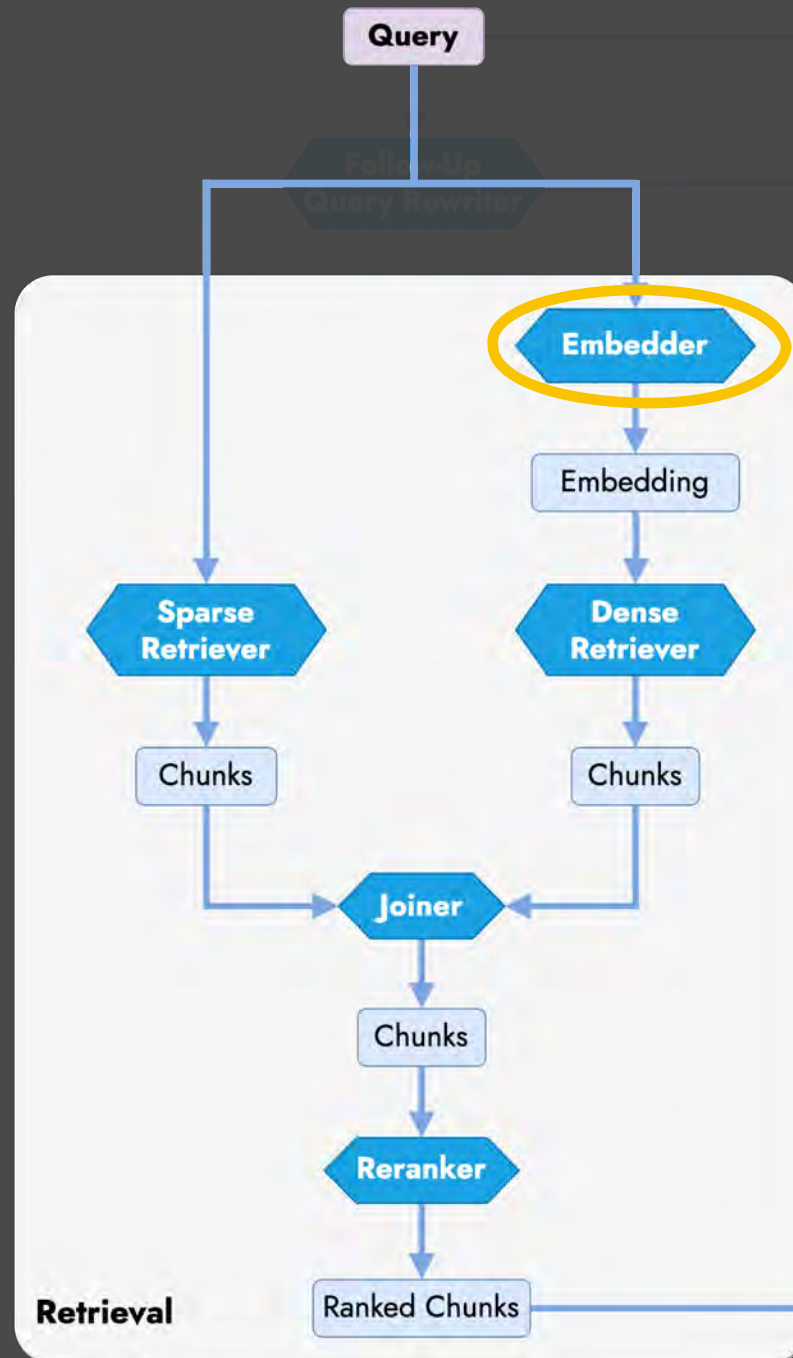






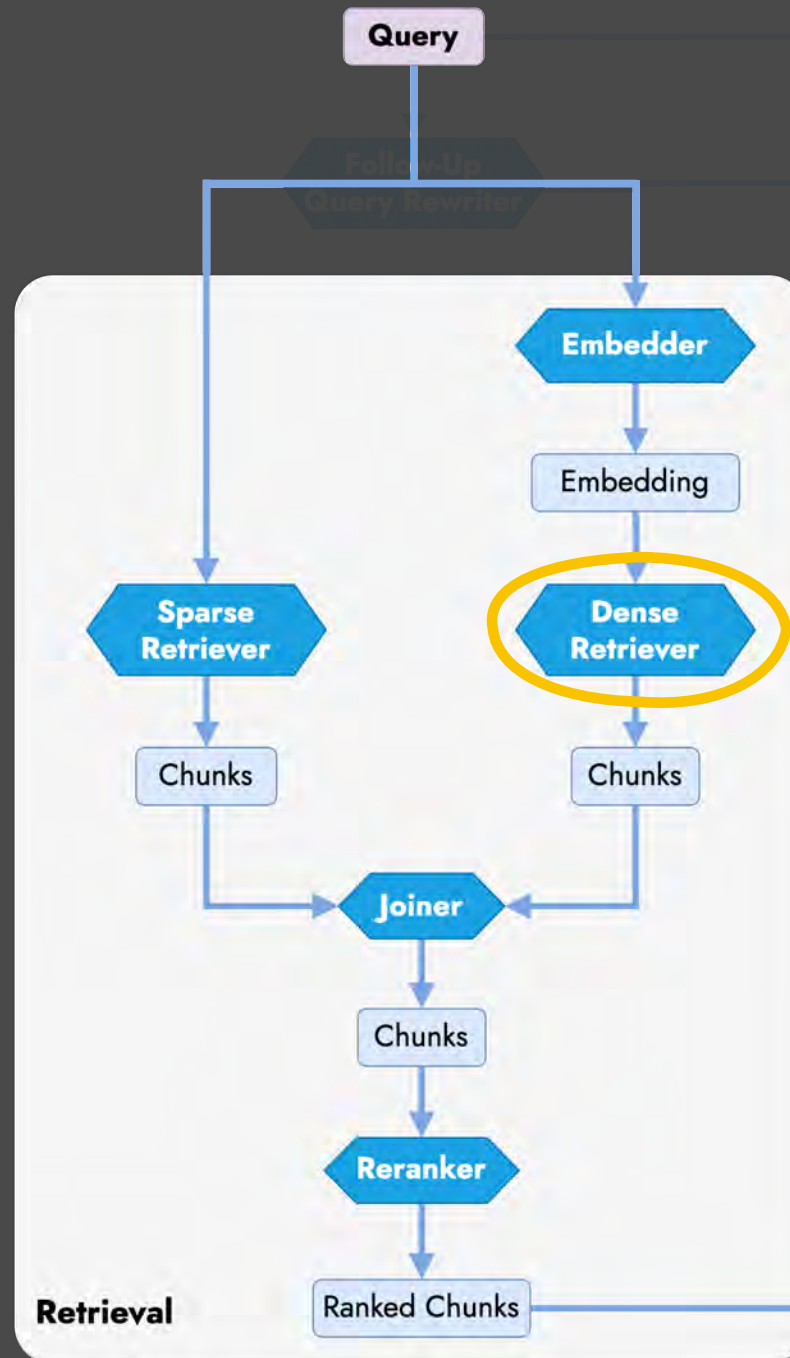
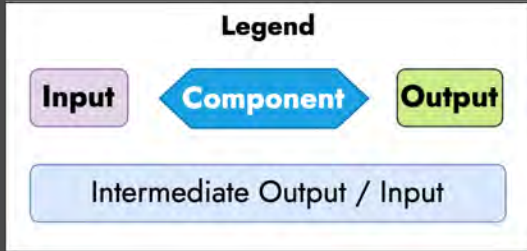
Best Matching 25 (BM25)

- ❑ In vielen Suchmaschinen eingesetzt
- ❑ Basiert auf Stichwörtern
- ❑ Gibt die k ähnlichsten Chunks aus



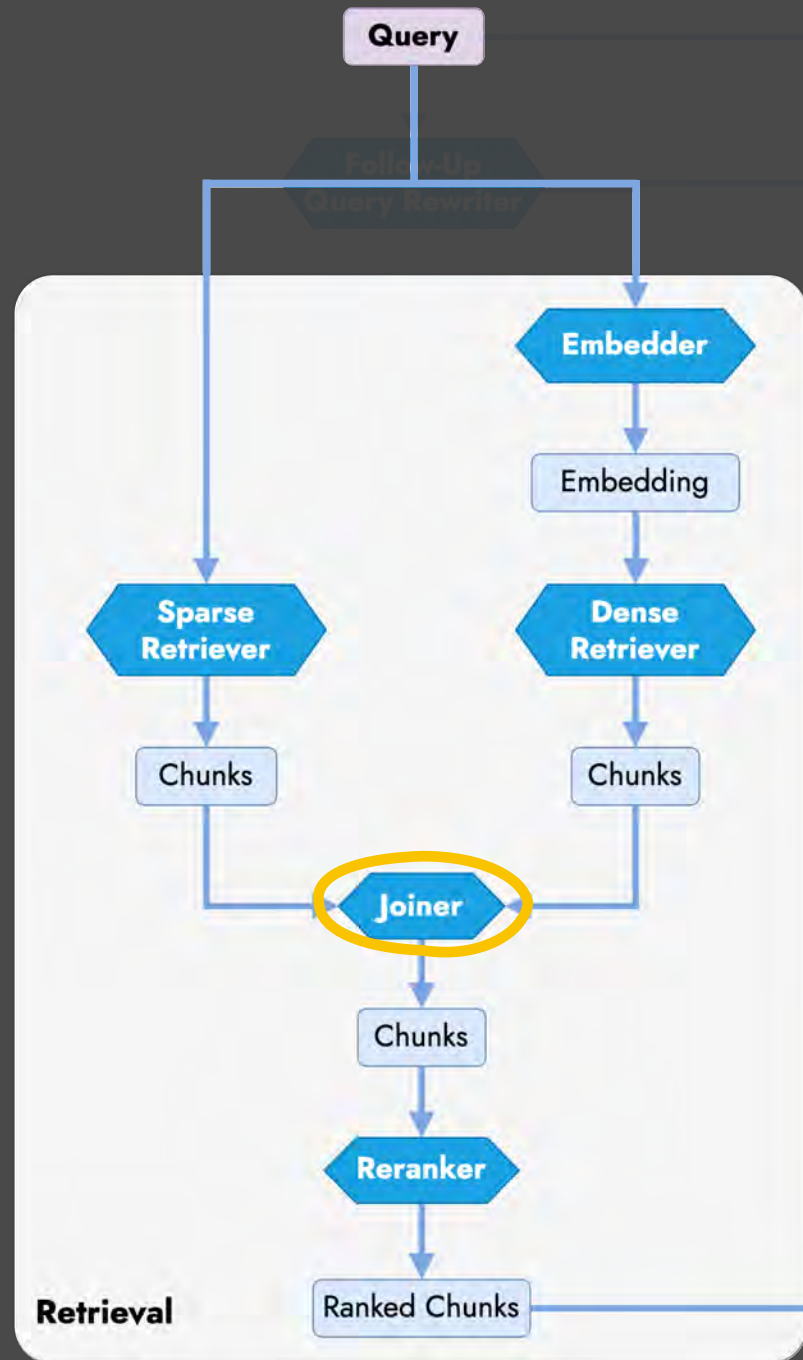
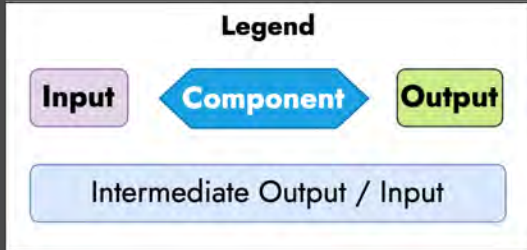
Sentence Transformer

- ❑ Sprachmodell, wandelt Text in numerischen Vektor um
- ❑ Texte mit ähnlichem Inhalt sind nahe beieinander in Vektorraum



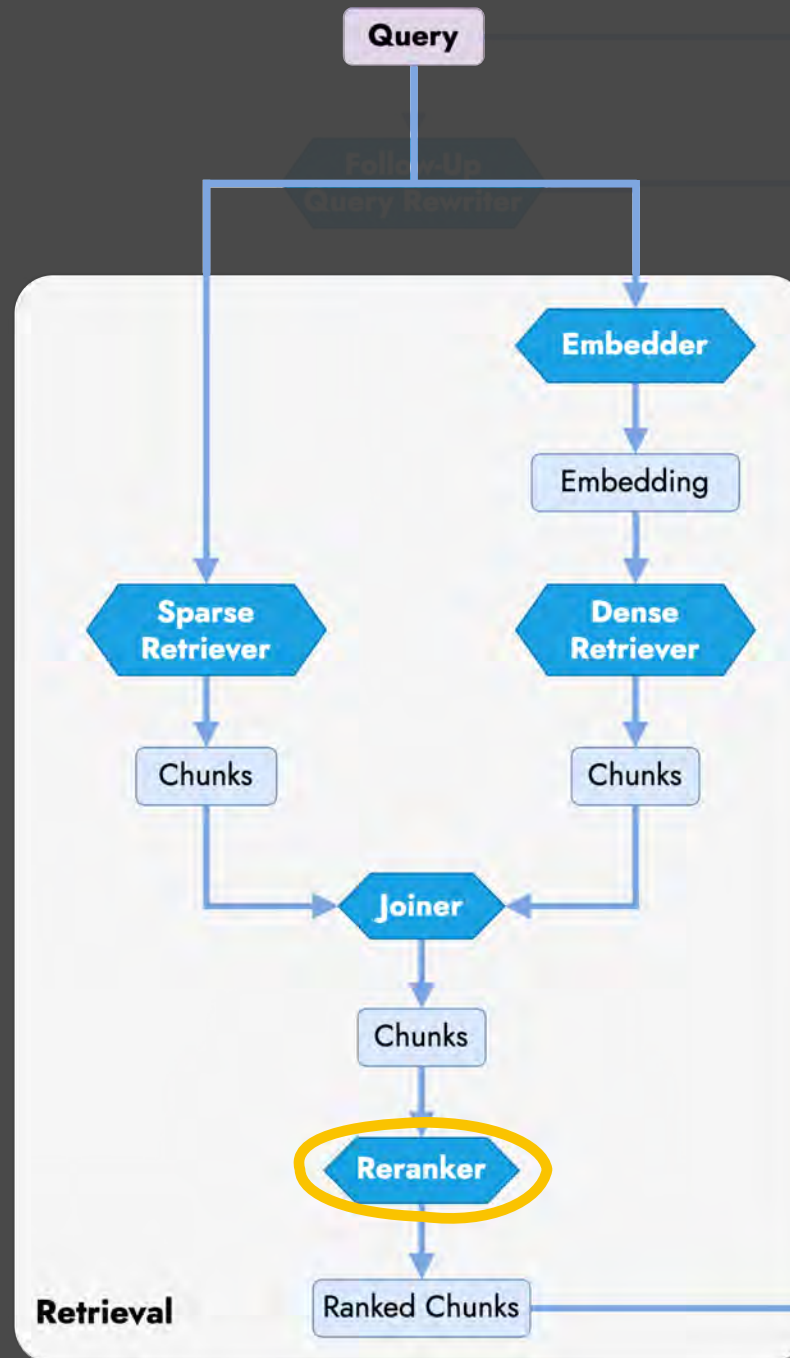
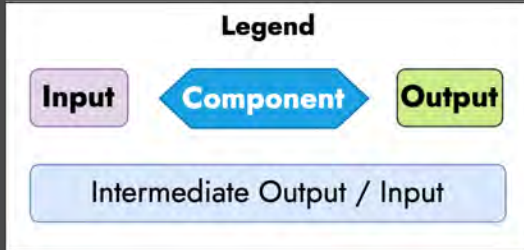
Cosine Similarity

- ❑ Berechnet den Kosinus des Winkels zwischen zwei Embedding-Vektoren
- ❑ Dieses Mass zeigt auf, wie ähnlich sich die Embeddings sind
- ❑ Gibt die k ähnlichsten Chunks aus



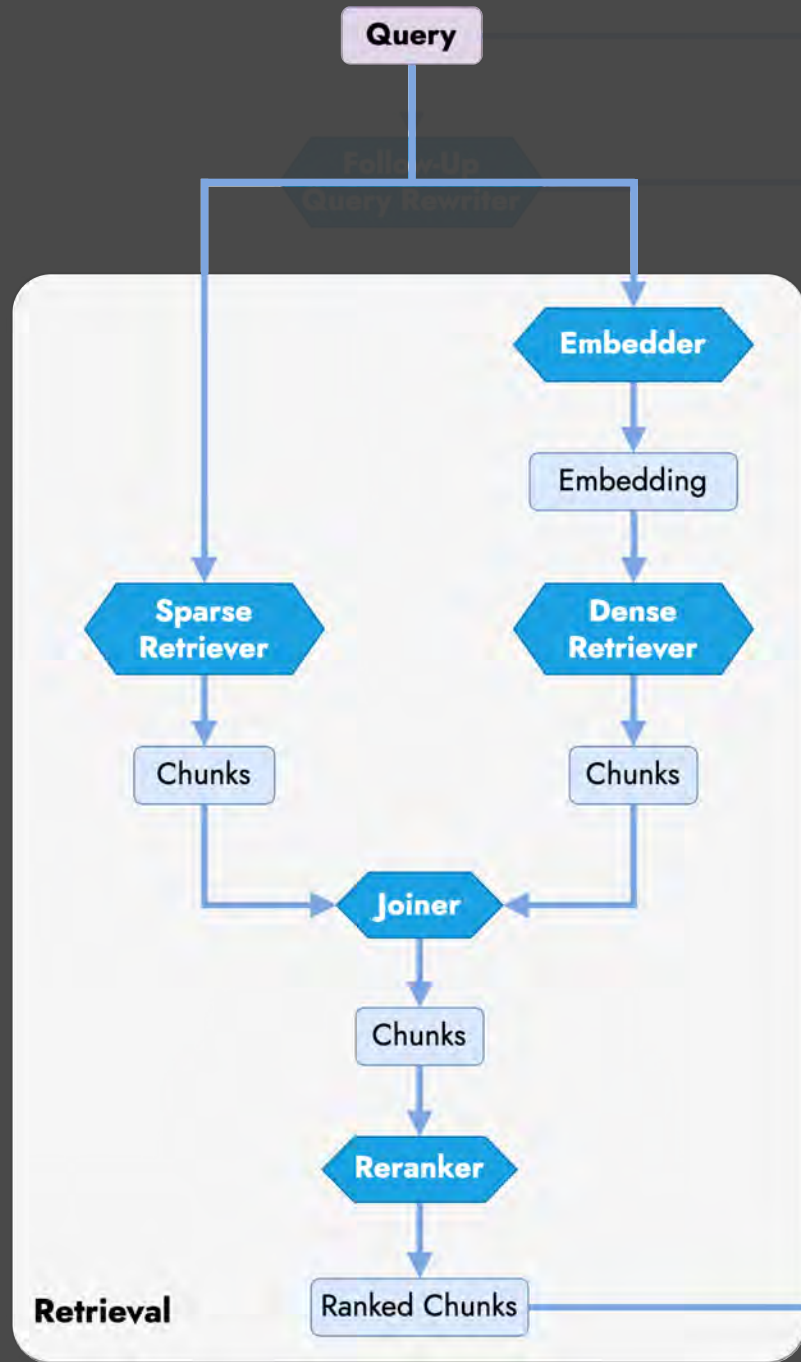
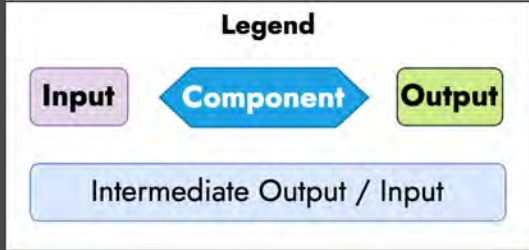

Zusammenfügen

- ❑ Kombiniert die gefundenen Chunks
- ❑ Gibt eine Liste mit gefundenen Chunks aus
- ❑ Es kann sein, dass ein Chunk von beiden Methoden gefunden wurde (Duplikate)




Reranking

- Anderes Sprachmodell, welches Embeddings erzeugt
- Mächtiger aber auch langsamer als das initiale Modell
- Bewertet die Relevanz der kombinierten Suchresultate neu
- Gibt die k relevantesten Chunks aus

Snowflake/snowflake-arctic-embed-l-v2.0

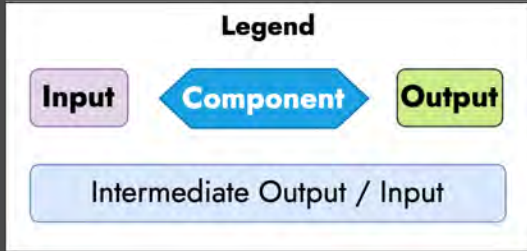



BAAI/bge-reranker-v2-m3

Text Generation

- ▶ Um die Antwort auf die Frage zu generieren, wird ein **Large Language Model (LLM)** eingesetzt
- ▶ Dieses muss intelligent genug sein, die gefundenen Textpassagen zu **interpretieren** und diese **in Bezug zur Frage** zu stellen
- ▶ Wenn die gefundenen Informationen irrelevant sind, dann muss das LLM dies **merken und keine Antwort liefern**





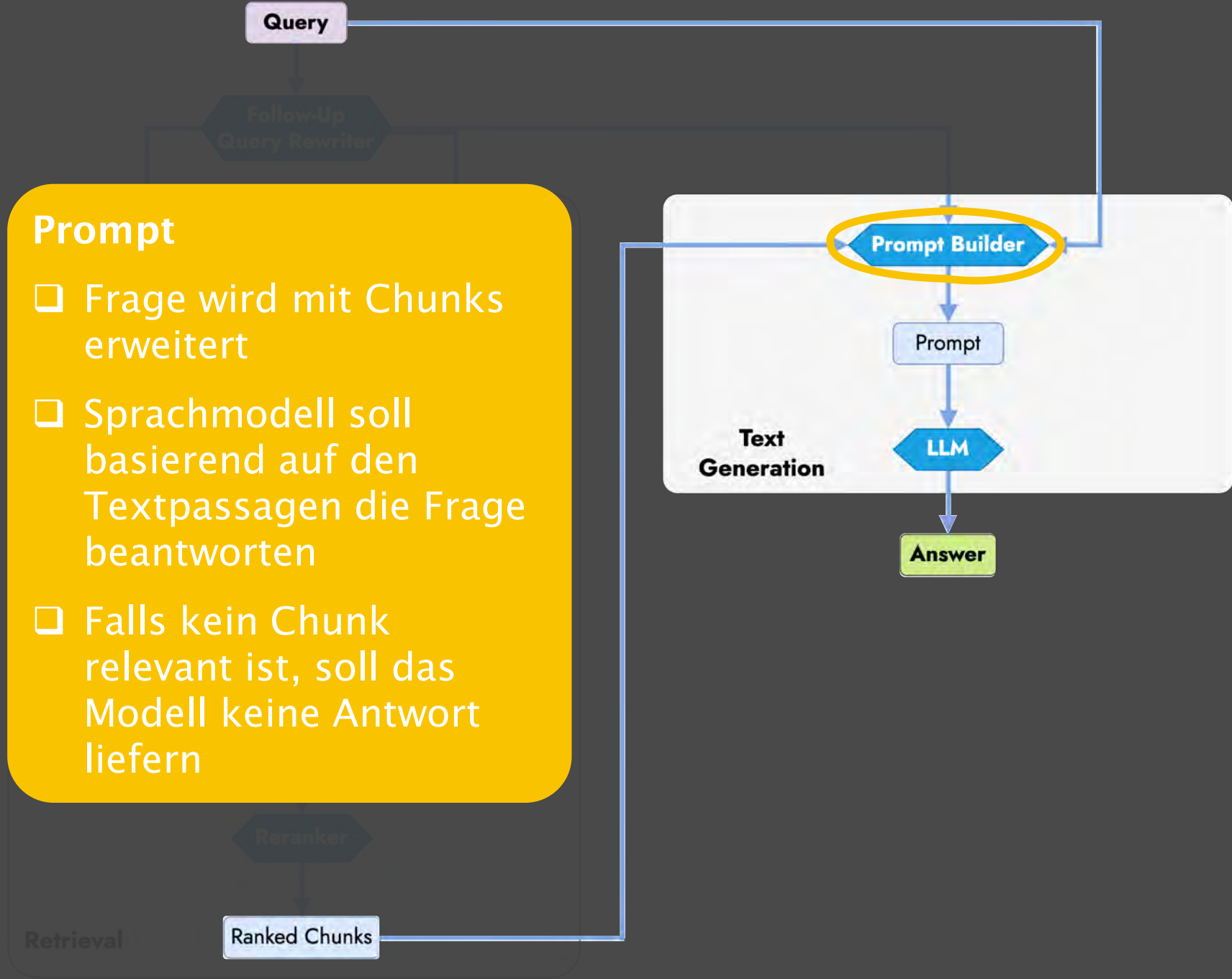
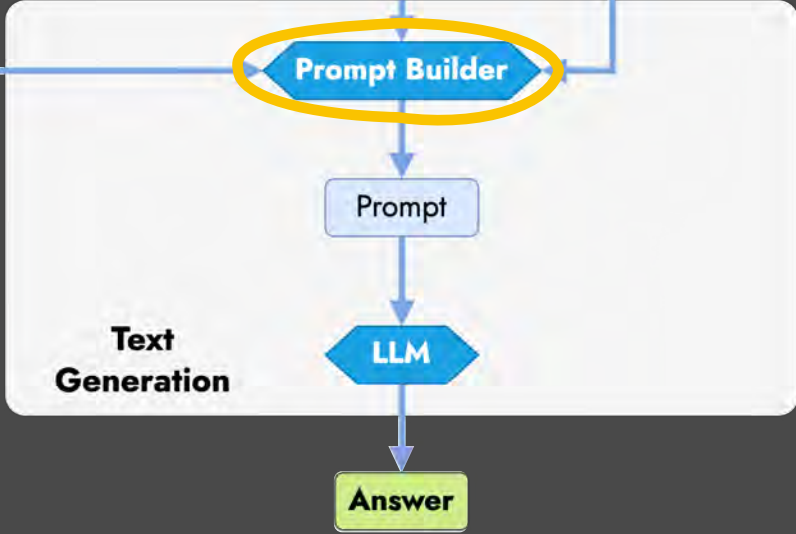
Prompt

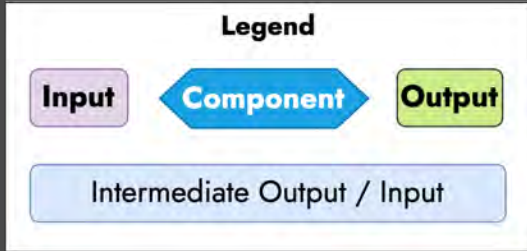
- ❑ Frage wird mit Chunks erweitert
- ❑ Sprachmodell soll basierend auf den Textpassagen die Frage beantworten
- ❑ Falls kein Chunk relevant ist, soll das Modell keine Antwort liefern



Query

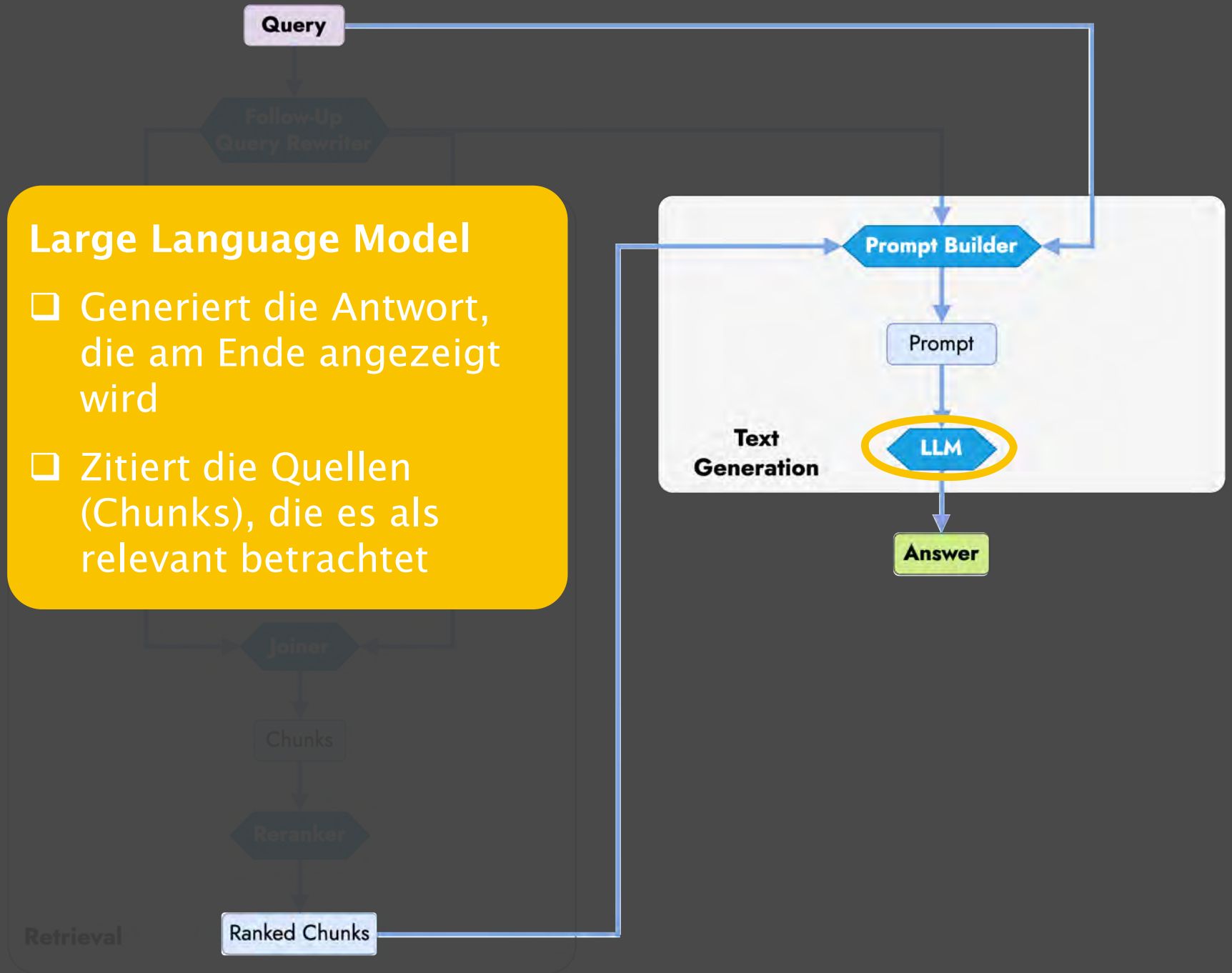
Follow-Up Query Rewriter

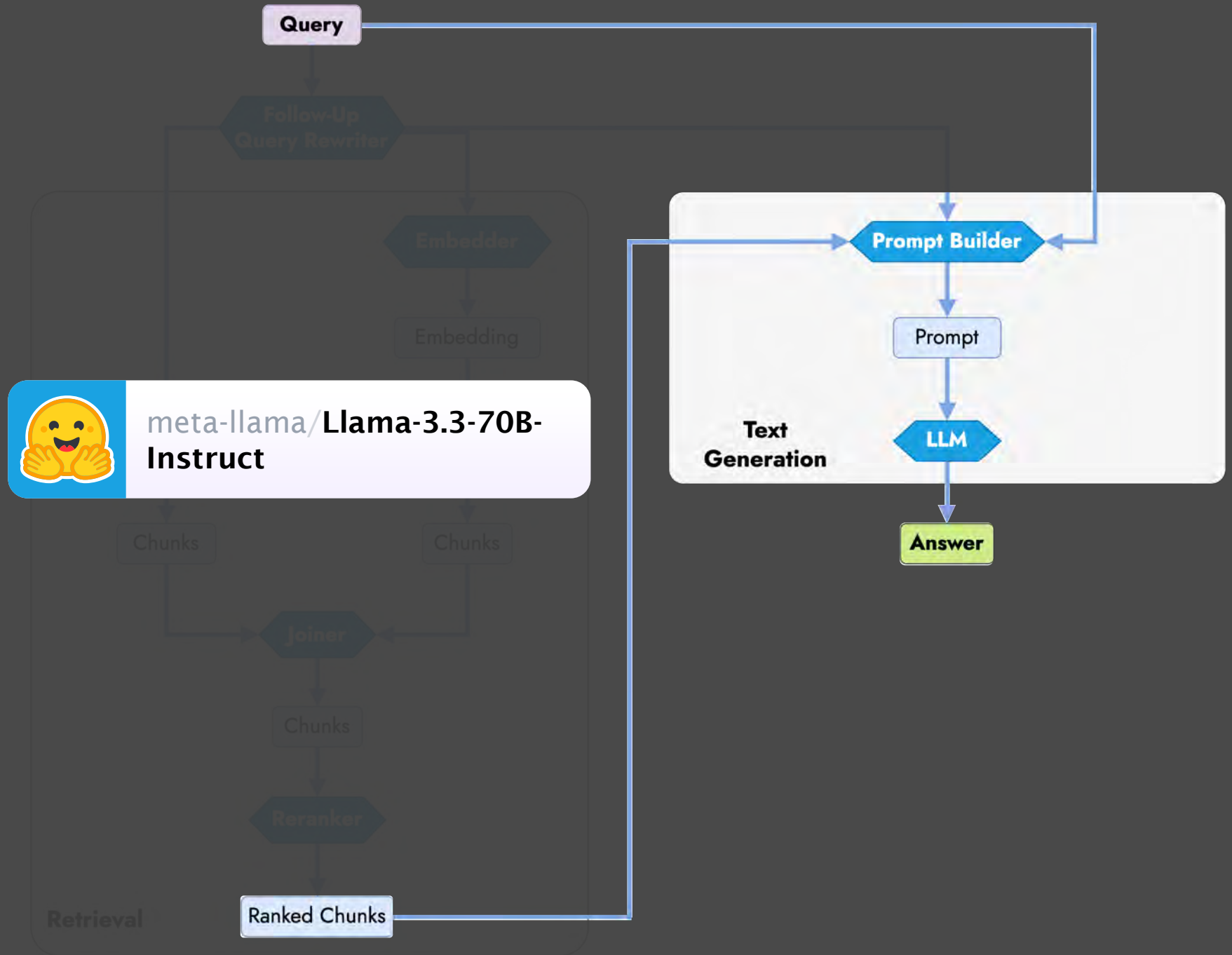
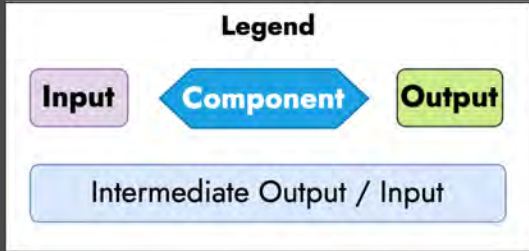




Large Language Model

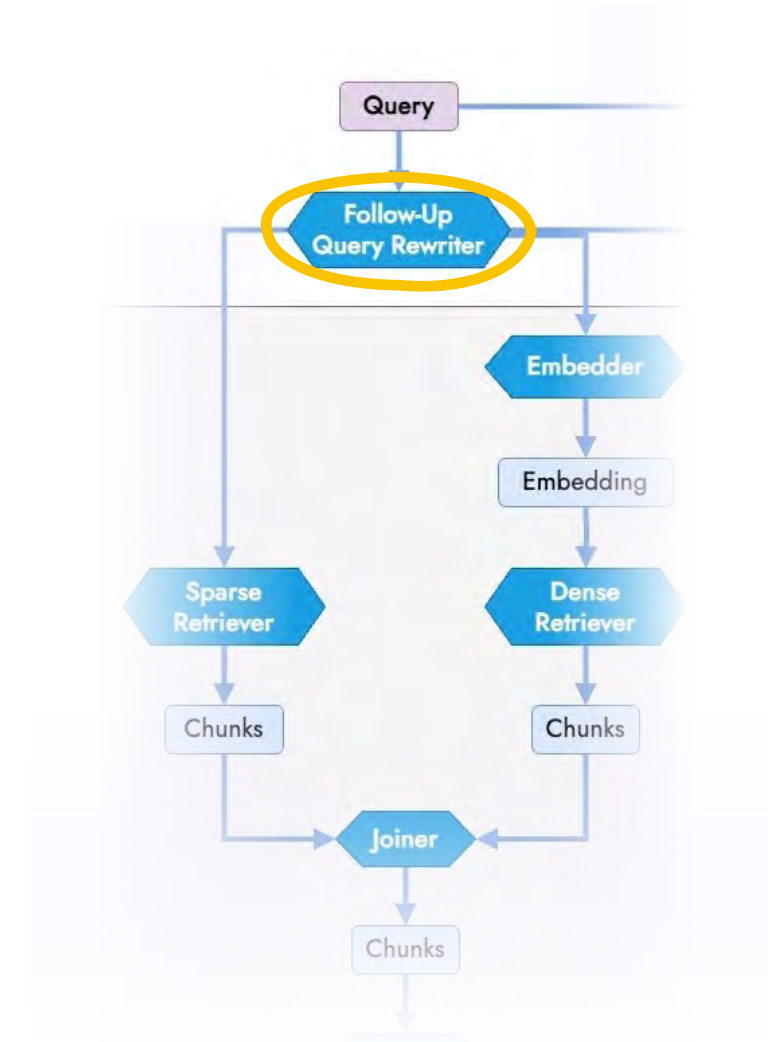
- ❑ Generiert die Antwort, die am Ende angezeigt wird
- ❑ Zitiert die Quellen (Chunks), die es als relevant betrachtet





Follow-Up Questions

- ▶ **Challenge:** Manchmal stellen Benutzer*innen Folgefragen. Unser Chatbot behandelt jedoch jede Nachricht als neue Anfrage.
- ▶ **Lösung:** Follow-Up Query Rewriter
 - ▶ **LLM entscheidet**, ob es sich um eine neue Frage oder eine Folgefrage handelt
 - ▶ Falls es eine Folgefrage ist, wird diese anhand der Chat-Historie so **umformuliert**, dass sie alle relevanten Details enthält
 - ▶ Stellt sicher, dass Folgefragen korrekt interpretiert werden





Was ist der Wohnsitz einer obdachlosen Person?



Der Wohnsitz einer obdachlosen Person bleibt in der Gemeinde registriert, in der sie früher einmal Wohnsitz hatte [1]. Wenn eine obdachlose Person jedoch während mehr als drei Monaten (...)

Quellen



Was, wenn die Person nie einen hatte?



Wenn eine Person nie einen Wohnsitz hatte, kann der Wohnsitz nach Art. 24 ZGB [1] bestimmt werden. Demnach bleibt der einmal begründete Wohnsitz einer Person bestehen, bis zum Erwerb eines neuen Wohnsitzes. Wenn jedoch ein früher begründeter Wohnsitz nicht nachweisbar ist oder ein im Ausland begründeter Wohnsitz aufgegeben und in der Schweiz kein neuer begründet worden ist, so gilt der Aufenthaltsort als Wohnsitz. Im Falle (...)



Was ist der Wohnsitz einer obdachlosen Person?

Ursprüngliche Frage

Was, wenn die Person nie einen hatte?

Neuformulierung aufgrund Follow-Up

Was ist der Wohnsitz einer Person, die obdachlos ist und noch nie einen Wohnsitz hatte?



Was, wenn die Person nie einen hatte?



Wenn eine Person nie einen Wohnsitz hatte, kann der Wohnsitz nach Art. 24 ZGB [1] bestimmt werden. Demnach bleibt der einmal begründete Wohnsitz einer Person bestehen, bis zum Erwerb eines neuen Wohnsitzes. Wenn jedoch ein früher begründeter Wohnsitz nicht nachweisbar ist oder ein im Ausland begründeter Wohnsitz aufgegeben und in der Schweiz kein neuer begründet worden ist, so gilt der Aufenthaltsort als Wohnsitz. Im Falle (...)

Open-Source Software

- ▶ Benutzeroberfläche: Streamlit
- ▶ Backend (RAG): Haystack
- ▶ Chunking: NLTK
- ▶ Embeddings: Sentence Transformers
- ▶ Reranking: FlagEmbedding
- ▶ Evaluation: DeepEval (via Haystack)
- ▶ Deployment: Docker Engine



Ende der Präsentation



Falls Sie noch Fragen haben, beantworte ich Ihnen diese gerne!

Frage stellen

