



Smarte Antworten aus geschütztem Fachcontent

CrossMediaForum KI-Spezial
05.12.2024





Lazaros Koutsianos
Bachelor of Computer Linguistik (LMU)
Consultant Search & Analytics
lazaros.koutsianos@shi-gmbh.com



Arkadius Nowakowski
Diplom Informatiker (FH)
Bereichsleitung Publisher Products und Web Development
arkadius.nowakowski@shi-gmbh.com
+49 176 11 633 637



Zum Unternehmen

- Standort: Augsburg
- Gegründet 1994
- 25 Mitarbeiter

- 3 Bereiche
 - Publisher Solutions
 - Web Development
 - Search & Analytics





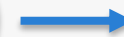
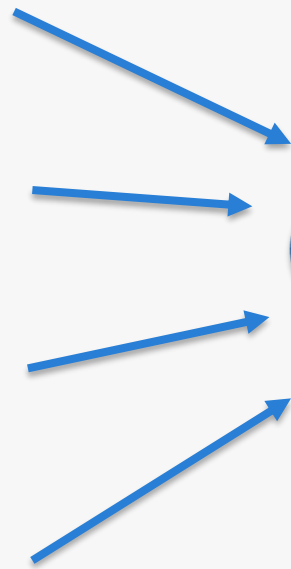
Agenda

- **KI mit Schutz von Fachcontent**
- **Semantische Suche in Apache Solr**
- **Large Language Modells (LLMs)**
- **RAG (Retrieval Augmented Generation)**
- **RAG-Voraussetzung: Chunking und Vektorspeicherung**
- **Methoden zur Verbesserung von RAG**
- **Demo**



KI mit Schutz für Fachcontent

- Wissensbasierter Chatbot
- Q&A Systeme
- Erweiterte Suchfunktionen
- Textzusammenfassung



Fachcontent





« KI mit Schutz für Fachcontent

Anforderungen:

- Maximale Datensicherheit
- Volle Kontrolle über Daten
- Compliance-Anforderungen

Lösung:

- ✓ RAG On-Prem
- ✓ LLM auf eigenem Server
- ✓ Fachcontent weiterhin in eigenem Unternehmensnetzwerk





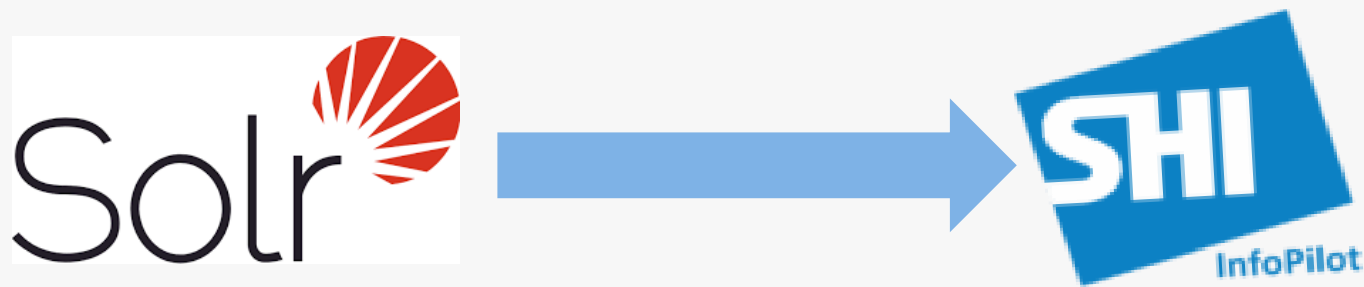
Agenda

- ✓ **KI mit Schutz für Fachcontent**
- **Semantische Suche in Apache Solr**
- **Large Language Modells (LLMs)**
- **RAG (Retrieval Augmented Generation)**
- **RAG-Voraussetzung: Chunking und Vektorspeicherung**
- **Methoden zur Verbesserung von RAG**
- **Demo**



Semantische Suche in Apache Solr

- Im Jahr 2022 mit der Version 9.0 in Apache Solr eingeführt
- Vorteil: SHI InfoPilot Retrieval Plattform basiert seit 10 Jahren auf Apache Solr





Semantische Suche

Konzepte:

- Bedeutung der Suchanfrage
- System versteht Kontext und Intention
- Einsatz von Natural Language Processing (NLP) & maschinelles Lernen (ML)

Ergebnisse:

- ✓ Verbessert die Relevanz
- ✓ Erhöht die Benutzerzufriedenheit
- ✓ Reduzierter Rechercheaufwand

Indexing



Text zu Vektoren

Searching



Semantische Suche



Agenda

- ✓ **KI mit Schutz für Fachcontent**
- ✓ **Semantische Suche in Apache Solr**
- **Large Language Modells (LLMs)**
 - **RAG (Retrieval Augmented Generation)**
 - **RAG-Voraussetzung: Chunking und Vektorspeicherung**
 - **Methoden zur Verbesserung von RAG**
 - **Demo**



LLM (Large Language Model)

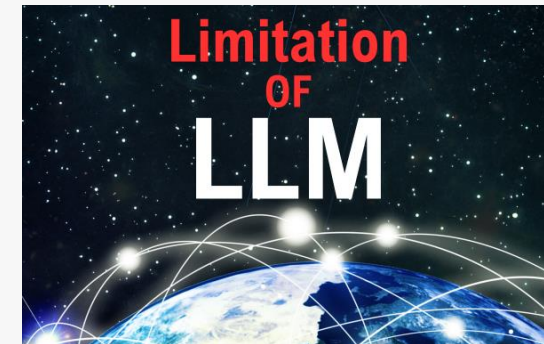
- Große Modelle des maschinellen Lernens
- Sie können viele Aufgaben erfüllen:
 - ✓ Beantworten von Fragen
 - ✓ Zusammenfassen von Dokumenten
 - ✓ Übersetzen von Sprachen
 - ✓ Generieren von Texten
- Mit genügend Beispielen trainiert, um menschliche Sprache erkennen





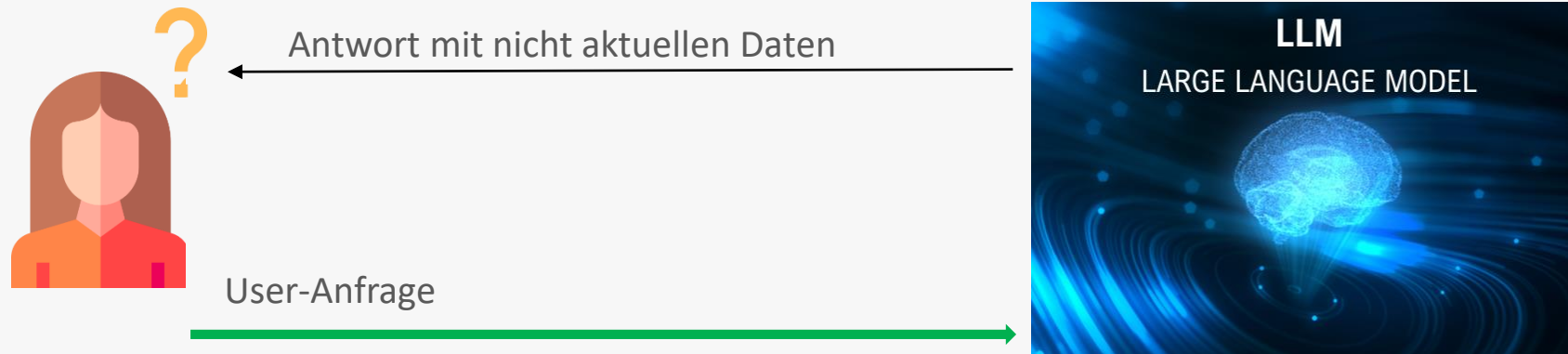
LLM-Einschränkungen

- Nicht auf dem neuesten Stand der Informationen
- Halluzinationen
- Keine bereichsspezifischen Informationen
- Das Neu-Training mit Informationen (Fine-Tuning) erfordert Ressourcen





Ohne RAG





Agenda

- ✓ **KI mit Schutz für Fachcontent**
- ✓ **Semantische Suche in Apache Solr**
- ✓ **Large Language Modells (LLMs)**
- **RAG (Retrieval Augmented Generation)**
 - **RAG-Voraussetzung: Chunking und Vektorspeicherung**
 - **Methoden zur Verbesserung von RAG**
 - **Demo**



RAG

- Kombiniert Abruf & Generierungsprozesse: Verbessert die Fähigkeiten großer Sprachmodelle

Funktionsweise:

- Abrufphase: Relevante Daten oder Dokumente werden abgerufen
- Generierungsphase: Das Sprachmodell nutzt die abgerufenen Daten & sein internes Wissen

- ✓ Ziel: Antworten basieren auf spezifischen Daten





Warum RAG?

- ✓ **RAG ermöglicht den Zugriff auf eine stets aktuelle Wissensdatenbank**
- Aktualität und Genauigkeit
- Fachspezifische Antworten
- Kostengünstige Implementierung
- Reduktion von Halluzinationen





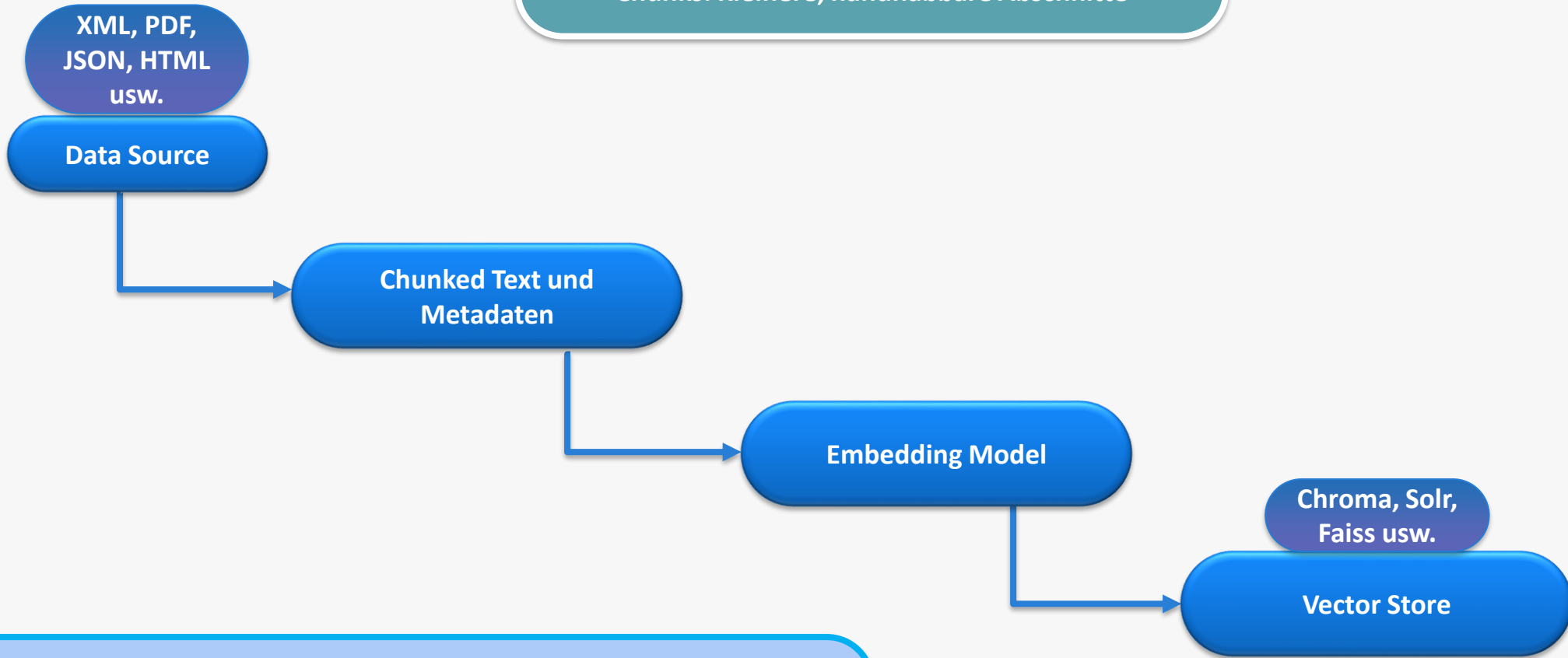
Agenda

- ✓ **KI mit Schutz für Fachcontent**
- ✓ **Semantische Suche in Apache Solr**
- ✓ **Large Language Modells (LLMs)**
- ✓ **RAG (Retrieval Augmented Generation)**
- **RAG-Voraussetzung: Chunking und Vektorspeicherung**
- **Methoden zur Verbesserung von RAG**
- **Demo**



RAG Voraussetzung: Chunking und Vektorspeicherung

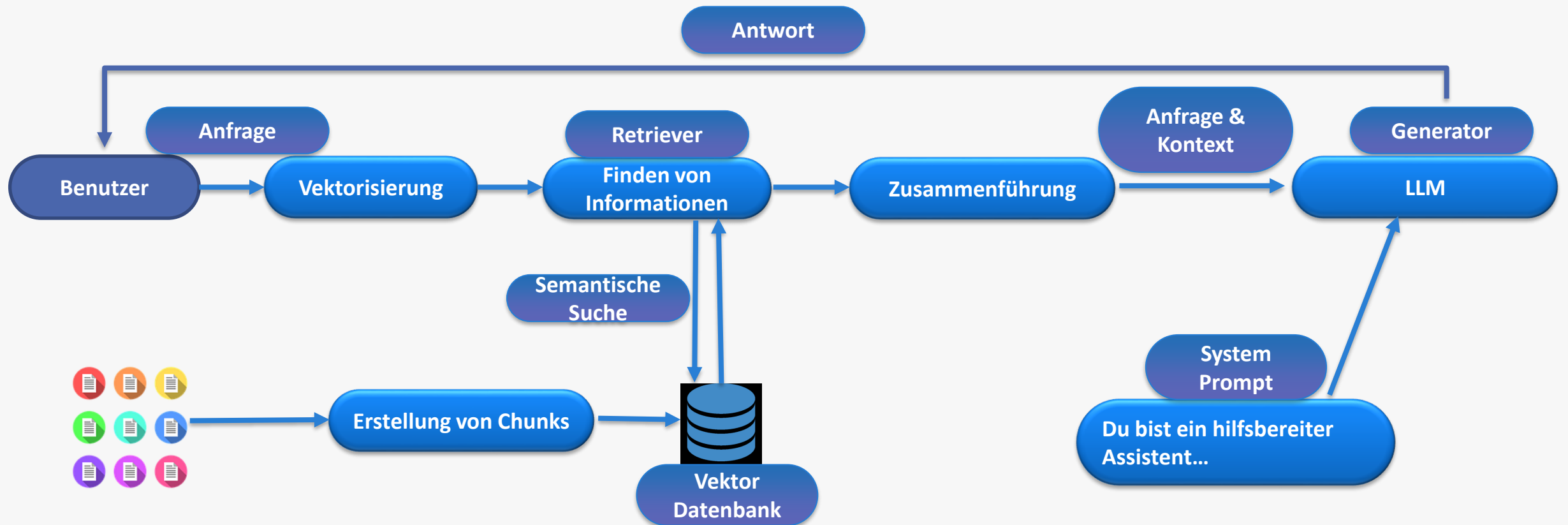
Chunks: Kleinere, handhabbare Abschnitte



Blogbeitrag: [Effiziente Datenaufbereitung und Basiskomponenten für RAG](#)



RAG Architektur





Agenda

- ✓ **KI mit Schutz für Fachcontent**
- ✓ **Semantische Suche in Apache Solr**
- ✓ **Large Language Modells (LLMs)**
- ✓ **RAG (Retrieval Augmented Generation)**
- ✓ **RAG-Voraussetzung: Chunking und Vektorspeicherung**
- **Methoden zur Verbesserung von RAG**
 - **Chunking Strategien**
 - **Hybride Suche**
 - **Contextual Retrieval**
 - **Demo**



Chunking Strategien

Strategie	Vorteil	Nachteil
→ Chunking mit fester Größe	Einfach zu implementieren	Zerschneidet Sätze oder Wörter
→ Token-basiertes Chunking	Präzise Kontrolle über die Größe der Chunks	Kontext kann verloren gehen
→ Rekursiv-basiertes Chunking	Bewahrt die logische Struktur des Textes	Kann langsamer sein
→ Dokumentenbasiertes Chunking	Folgt der Struktur des Dokuments	Chunks können variieren
→ Semantisches Chunking	Inhaltlich kohärent	Rechenintensiv und komplex

→ Blogbeitrag: [Chunking für RAG: Effektive Methoden im Überblick](#)



Agenda

- ✓ **KI mit Schutz für Fachcontent**
- ✓ **Semantische Suche in Apache Solr**
- ✓ **Large Language Modells (LLMs)**
- ✓ **RAG (Retrieval Augmented Generation)**
- ✓ **RAG-Voraussetzung: Chunking und Vektorspeicherung**
- **Methoden zur Verbesserung von RAG**
 - ✓ **Chunking Strategien**
 - **Hybride Suche**
 - **Contextual Retrieval**
- **Demo**



Hybride Suche

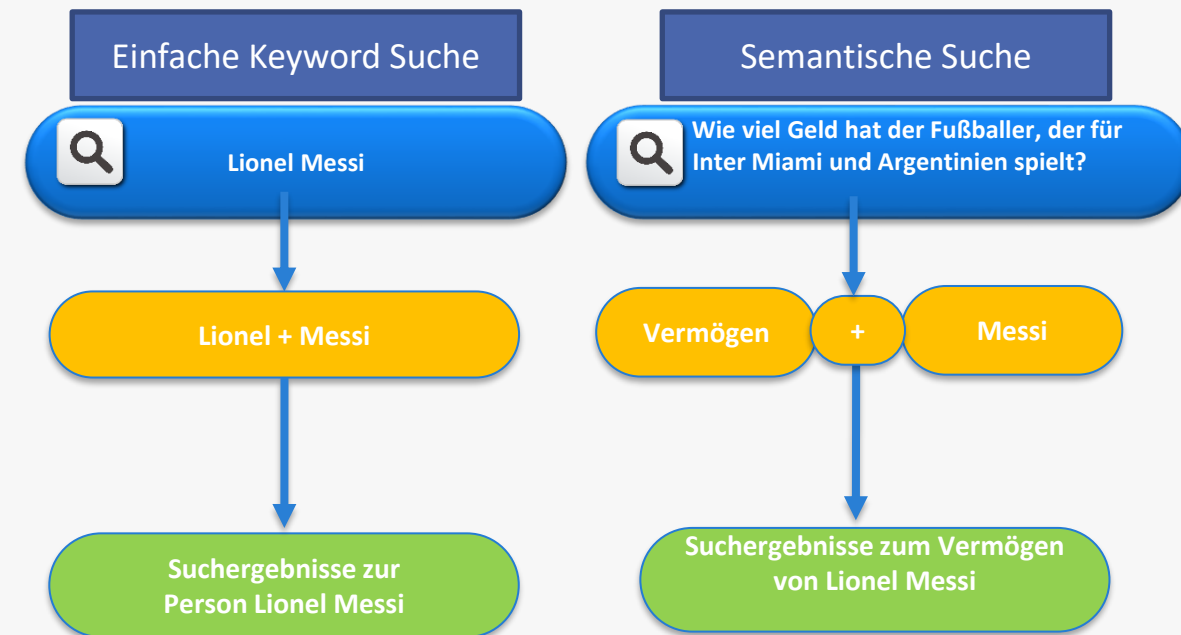
Semantische Suche:

- Besser auf Absätzen - Chunks
- Keine exakten Schlüsselwörter
- Nutzer erwarten lexikalische Übereinstimmungen

➤ Warum erscheint das Dokument mit den Schlüsselwörtern nicht?

Hybride Suche:

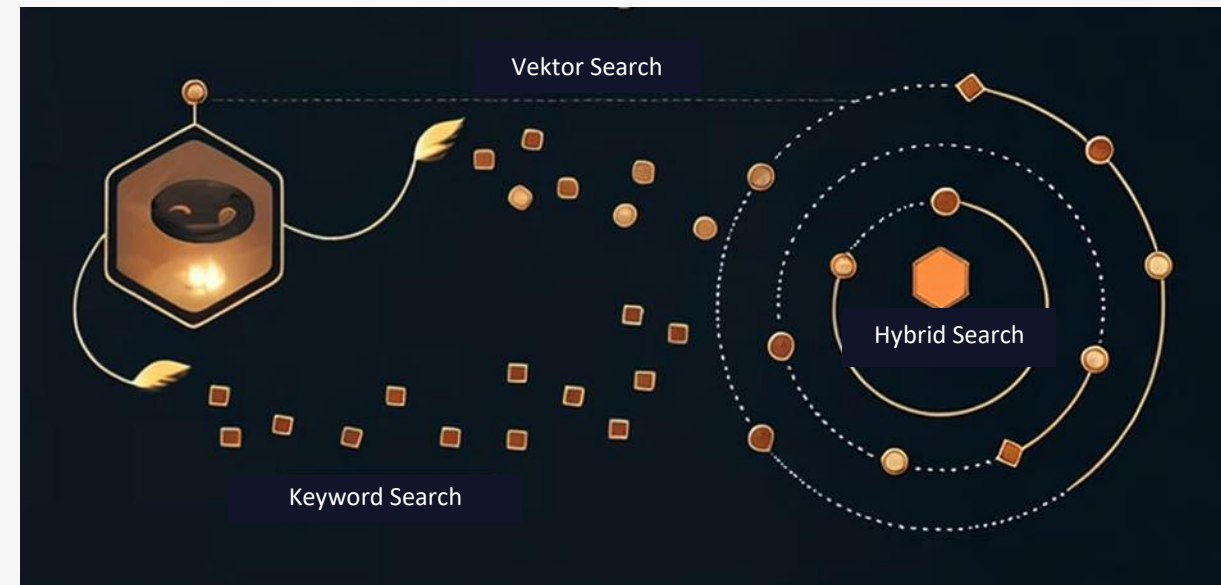
- ✓ Lexikalische und semantische Suche kombinieren





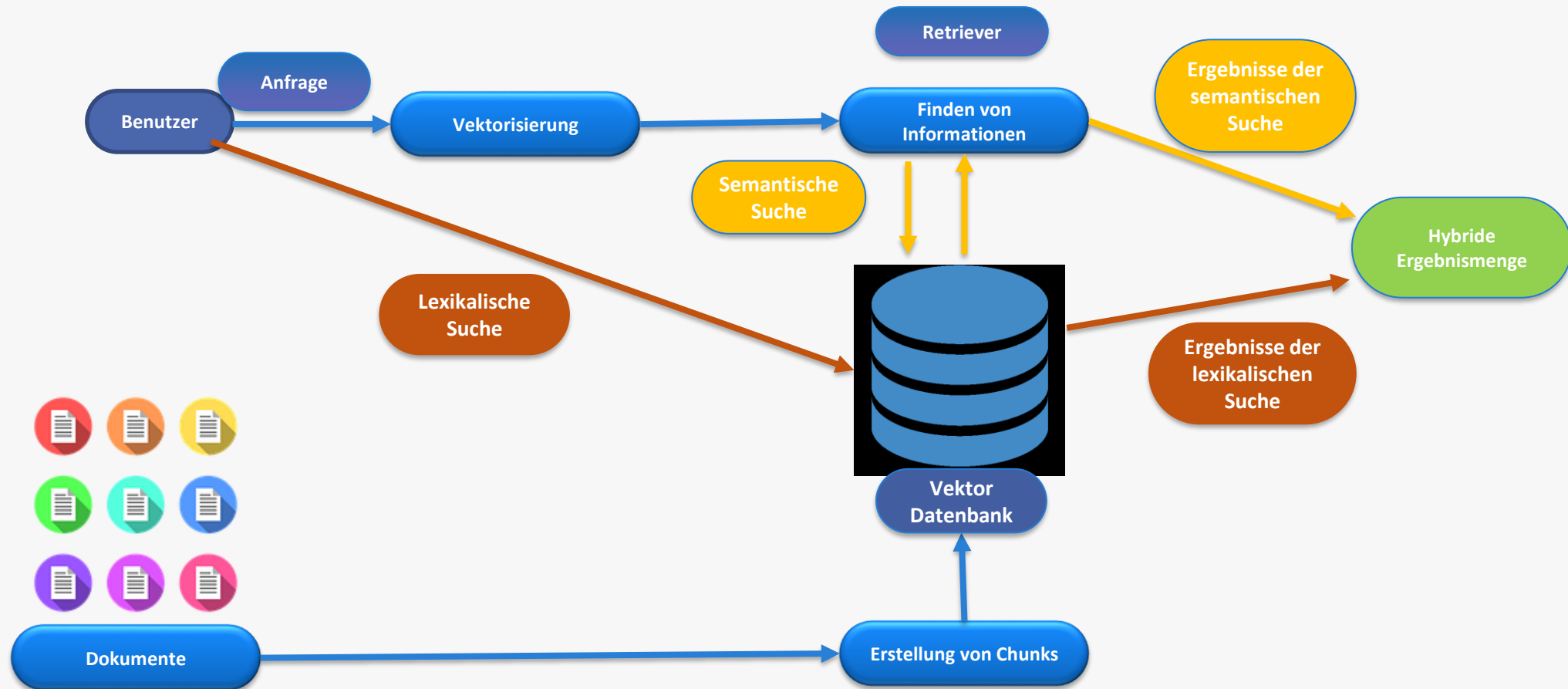
Hybride Suche

- Was bedeutet es, diese beiden Welten zu kombinieren?
- Abrufphase
 - Zwei Gruppe von Kandidaten
 - Die Hybride Ergebnismenge wird erstellt
- Rankingphase
 - Jedes Dokument der hybriden Ergebnismenge erhält einen Score





Hybride Suche





Agenda

- ✓ **KI mit Schutz für Fachcontent**
- ✓ **Semantische Suche in Apache Solr**
- ✓ **Large Language Modells (LLMs)**
- ✓ **RAG (Retrieval Augmented Generation)**
- ✓ **RAG-Voraussetzung: Chunking und Vektorspeicherung**
- ✓ **Methoden zur Verbesserung von RAG**
 - ✓ **Chunking Strategien**
 - ✓ **Hybride Suche**
 - **Contextual Retrieval**
- **Demo**

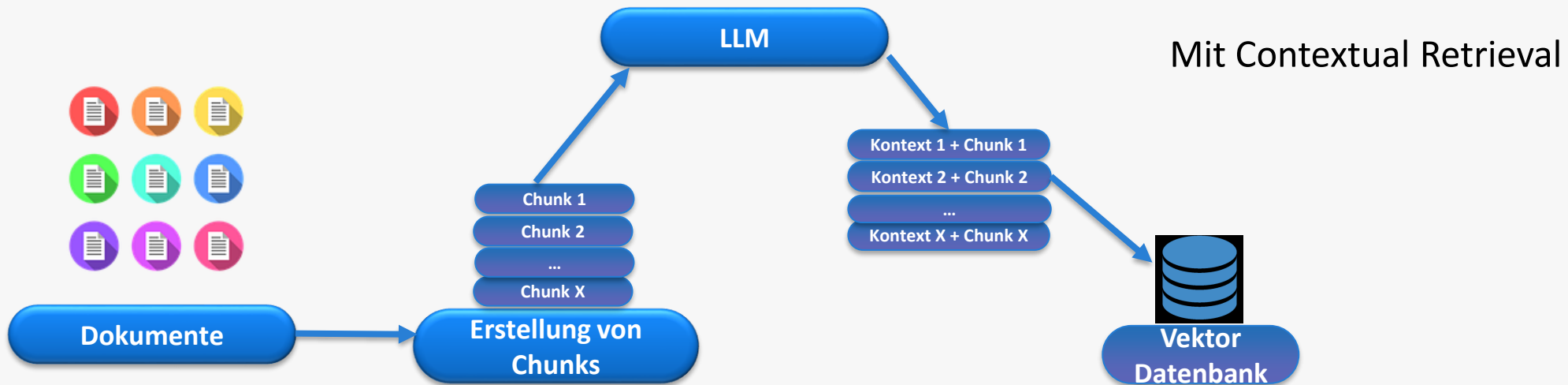
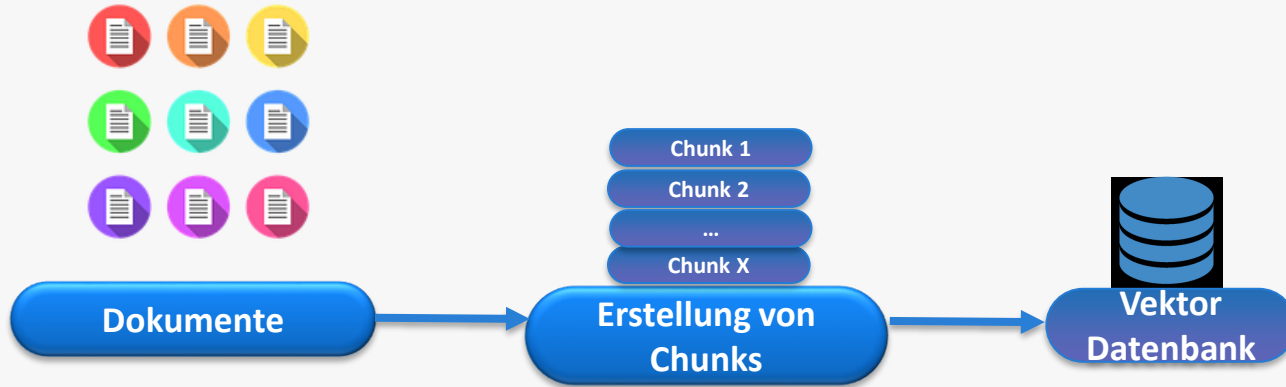


Contextual Retrieval

- RAG: Gut für viele Anwendungen
 - Problem?
 - Einzelne Chunks enthalten oft nicht genügend Kontext
- Contextual Retrieval:
 - Fügt jedem Chunk vor der Vektorisierung zusätzlichen erklärenden Kontext hinzu



Contextual Retrieval





Contextual Retrieval

➤ Frage:

In welchem Normkapitel finde ich die Anforderungen an interne Audits?

➤ Chunk:

Durchführung des Audits. Hierbei werden die Anforderungen an das „interne Audit“ gemäß den Vorgaben des **Normkapitels 9.2 der ISO 9001** zugrunde gelegt.

➤ Kontext Information:

In diesem Textabschnitt **geht es um die Auditplanung und -durchführung nach ISO 9001:** Hier wird empfohlen, **bei der Umsetzung des in Kapitel 9.1.3 beschriebenen Audits die Anforderungen an das interne Audit aus Kapitel 9.2 der ISO 9001 zu berücksichtigen.**



Vorteile von Contextual Retrieval

- Reduzierung von Abruffehlern
- Verbesserte Relevanz der Suchergebnisse
- Einfache Implementierung



Agenda

- ✓ **KI mit Schutz für Fachcontent**
- ✓ **Semantische Suche in Apache Solr**
- ✓ **Large Language Modells (LLMs)**
- ✓ **RAG (Retrieval Augmented Generation)**
- ✓ **RAG-Voraussetzung: Chunking und Vektorspeicherung**
- ✓ **Methoden zur Verbesserung von RAG**
 - ✓ **Chunking Strategien**
 - ✓ **Hybride Suche**
 - ✓ **Contextual Retrieval**
- **Demo**



Demo



AI Search: Nutzung des Modells ohne RAG mit Trainingsdaten bis 2023

Geben Sie Ihre Anfrage ein:

Wer hat die Fußball-Europameisterschaft 2024 gewonnen und wer war der beste Spieler?

Absenden

Antwort des Modells:

Es tut mir leid, aber ich kann diese Frage nicht beantworten, da die Fußball-Europameisterschaft 2024 noch nicht stattgefunden hat. Die EM findet in Deutschland statt und wird vom 14. Juni bis zum 14. Juli 2024 ausgetragen. Daher ist es momentan nicht möglich, den Gewinner oder den besten Spieler zu kennen.



Demo: Mit RAG

AI Search: Nutzung des Modells mit RAG und Nachrichten-Daten von Juli 2024

Geben Sie Ihre Anfrage ein:

Wer hat die Fußball-Europameisterschaft 2024 gewonnen und wer war der beste Spieler?

Absenden

Ergebnisse der semantischen Suche:

ID: 10 Metadata: {'id': 10} Text: Die spanische Nationalmannschaft hat am 14 Juli 2024 zum vierten Mal den Europameisterschaft-Titel gewonnen. Im Finale der Europameisterschaft 2024 setzten sich die Iberer im Berliner Olympiastadion mit 2:1 (0:0) gegen England durch. Nico Williams und Mikel Oyarzabal erzielten die Tore für Spanien. Cole Palmer gelang nur der zwischenzeitliche Ausgleich. Die Three Lions müssen weiter auf den ersten Titel seit der WM 1966 warten.

ID: 12 Metadata: {'id': 12} Text: Rodri ist als bester Spieler der Europameisterschaft ausgezeichnet worden. Der Spanier erhielt die Trophäe nach dem 2:1 im Finale gegen England, in dem er zur Pause verletzt ausgewechselt werden musste. Sein Teamkollege Lamine Yamal wurde zum besten Nachwuchsspieler gewählt. Über die Awards entschied eine von der UEFA zusammengestellte Jury. Sie bestand aus zwölf Mitgliedern. Dazu gehören unter anderem Fabio Capello (Italien) und Rafael Benitez (Spanien).

Antwort des Modells:

Vielen Dank für Ihre Frage!

Die Fußball-Europameisterschaft 2024 wurde von der spanischen Nationalmannschaft gewonnen. Im Finale, welches am 14. Juli 2024 im Berliner Olympiastadion stattfand, setzten sich die Iberer mit 2:1 (0:0) gegen England durch. Die Tore für Spanien erzielten Nico Williams und Mikel Oyarzabal, während Cole Palmer nur der zwischenzeitliche Ausgleich gelang.

Der beste Spieler der Europameisterschaft ist Rodri geworden. Diese Auszeichnung erhielt er nach dem Finale gegen England, obwohl er in der ersten Halbzeit verletzt ausgewechselt werden musste. Die Entscheidung über die Awards wurde von einer Jury getroffen, die aus zwölf Mitgliedern bestand, darunter Fabio Capello (Italien) und Rafael Benitez (Spanien).



Danke für die Aufmerksamkeit



Fragen?

KONTAKT

SHI

SHI GmbH

Konrad-Adenauer-Allee 15
D - 86150 Augsburg

info@shi-gmbh.com
0821-74 82 633 - 0
[@SHIEngineers](#)