

# DBE AI-powered Quantum Expert

Zwischenstandspräsentation – 11.03.2025

1

Unser Projekt mit KIPU Quantum

2

Retrieval Augmented Generation

3

Ausblick des Projekts



01

# Unser Projekt mit KIPU Quantum

## Projektziel:

Entwicklung eines funktionierenden Chatbots, der mithilfe des RAG-Ansatzes (Retrieval-Augmented Generation) auf eine Datenbank mit Quantenalgorithmen zugreift, um auf Nutzeranfragen passende Algorithmen bereitzustellen.

## Erwartetes Ergebnis:

Ein Chatbot, der den RAG-Ansatz nutzt, um in Echtzeit intelligente und kontextbezogene Antworten auf die PlanQK-Wissensartefakte bereitstellt.

## Kontext:

In einer der größten Quanten-Communities von KIPU Quantum teilen Experten Algorithmen und Anwendungsfälle. Entwickler bieten Dienste auf einem Marktplatz an, während Nutzer Anwendungen abonnieren. Die Herausforderung ist die Bereitstellung intelligenter, kontextsensitiver Antworten auf die wachsende Wissensbasis in Echtzeit.

## Methoden:

Agiles Projektmanagement, regelmäßiger Austausch mit dem Partnerunternehmen sowie interne Recherche und Umsetzung zur Realisierung der RAG-Architektur mit Einsatz eines LLMs.

## Projektleiter & Team:

Projektleiter –  
Prof. Dr. Christian Decker, Prof. Dr. Uwe Breitenbücher

Team –  
Sinem Kayatas, Michelle Fahrner, Liv Rottler, Glen Braun, Philip de Bakker, Lars Kunik, Dominik Erman, Katharina Willig





**Unternehmensname:**  
KIPU Quantum GmbH

**Geschäftsbereich:**  
Quantencomputing

**Gründungsjahr:**  
2021

**Standorte:**  
Karlsruhe und Berlin

**Mission:** KIPU Quantum entwickelt Quantencomputing-Lösungen, die reale Geschäftsprobleme lösen. Der Fokus liegt auf spezifischen Quantenalgorithmen, die bereits heute in Optimierung, KI und Simulation einen „Quantum Advantage“ bieten.

**Starkes Team:** Über 30+ PhDs in Quantenphysik, Mathematik, Informatik, KI und Chemie

**Zielgruppen:** Entwickler, Unternehmen, Forscher & Quantenhardware-Anbieter

**Branchen:** Pharmazie, Chemie, Logistik, Finanzen

Projektrelevantes Produkt:

- PlanQK Plattform: Eine Cloud-basierte Plattform, die die besten Quantenalgorithmen mit einer modernen Infrastruktur kombiniert und eine einfache Implementierung für Unternehmensanwender ermöglicht.

**Unsere Ansprechpartner:**

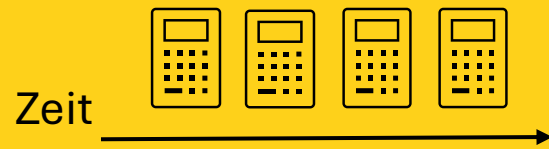


**Michael Falkenthal**  
Head of Plattform

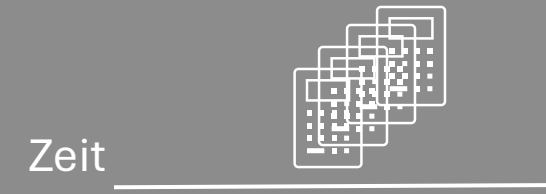


**Christoph Krieger**  
Software Engineer

**Quantencomputing** ist eine neue Art von Computertechnik, die erlaubt Berechnungen viel schneller und effizienter durchzuführen als klassische Computer.



Klassische Computer



Quantencomputer

**Wie kann man die PlanQK Plattform nutzbar und das Thema Quantencomputing mit all seinen Vorteilen verständlich machen?**



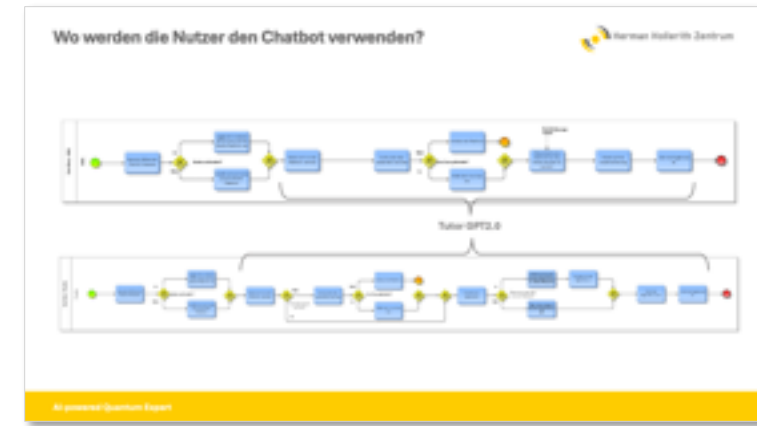
**Mit einem Chatbot auf der PlanQK Plattform der bestens mit der Arbeit von KIPU und der PlanQK Plattform vertraut ist!**

## Wer sind die Nutzer?



The screenshot shows a document titled "Persone" with two columns. The left column is headed "Business" and the right column is headed "Physiker". Both columns contain a "Ziel:" (Goal) section and a "Werkzeuge/Tools/Methoden:" (Tools/Methods) section. The "Business" column lists goals related to user experience and system optimization, while the "Physiker" column lists goals related to quantum computing and system performance. The "Werkzeuge/Tools/Methoden:" section in both columns lists various tools and methods used for system analysis and optimization.

## Interaktion mit dem Chatbot





## Business

### Ziel:

Die betriebswirtschaftlichen Nutzenden möchte die Plattform sicher bedienen, ihre Funktionen verstehen und sie gezielt für seine Anforderungen nutzen. Sie wollen relevante Anwendungsfälle finden, bewerten, berechnete Ergebnisse wirtschaftlich interpretieren und erstellte Modelle langfristig für Analysen und Optimierungen verwalten.

### Unterstützung durch Chatbot:

- Der Chatbot kann Fachbegriffe und technische Konzepte in einfachen und verständlichen Worten erläutern, sodass auch Nutzer ohne tiefgehendes Vorwissen sich schnell zurechtfinden.
- Empfehlung relevanter Dokumentationen basierend auf der Problemstellung des Nutzers.
- Unterstützung als Navigationshelfer innerhalb der Plattform.
- Analyse der vom Nutzer beschriebenen Herausforderungen und Empfehlung passender Anwendungsfälle.
- Gegenüberstellung verschiedener Use Cases mit Erläuterung ihrer Stärken und Vorteile zur besseren Entscheidungsfindung.
- Unterstützung bei der Interpretation der berechneten Ergebnisse im wirtschaftlichen Kontext zur fundierten Entscheidungsfindung.
- Hilfestellung bei der Speicherung und Verwaltung erstellter Modelle zur flexiblen Wiederverwendung und Optimierung.

## Physiker

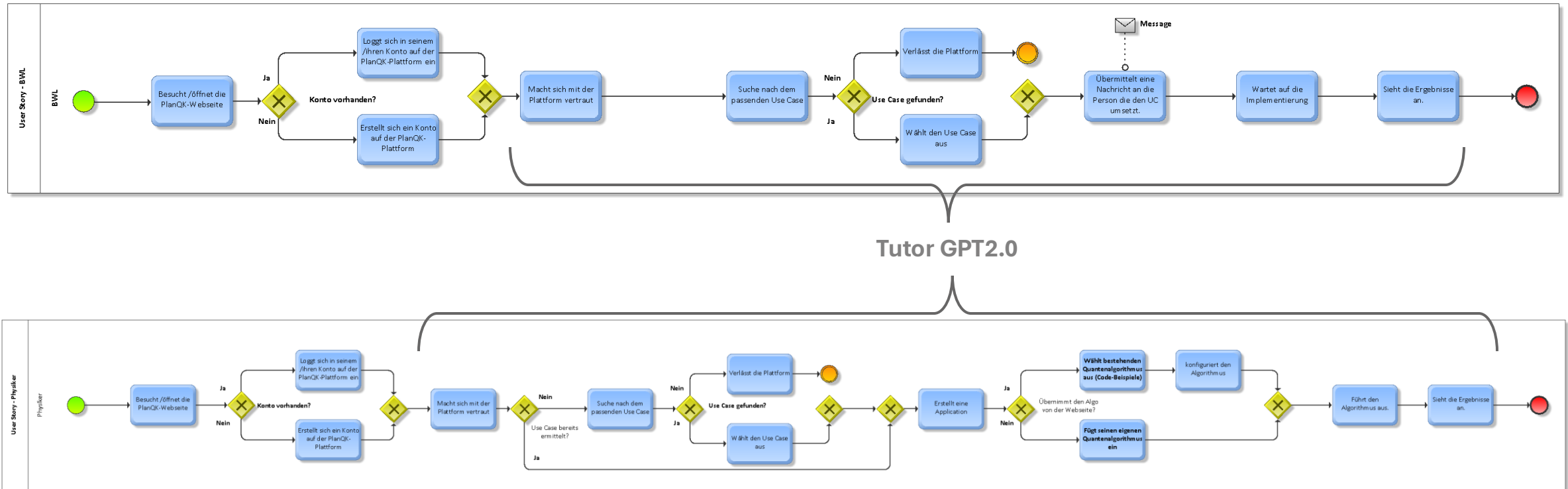
### Ziel:

Die Physiker\*innen kennen Quantenalgorithmen, sind aber neu auf der Plattform und möchte sich schnell zurechtfinden. Besonders interessiert Sie die Anwendung der Algorithmen, um neue Ansätze zu entdecken und Ihr Unternehmen strategisch weiterzuentwickeln. Zudem wollen Sie Modelle flexibel mit verschiedenen Daten und Use Cases reproduzieren können.

### Unterstützung durch Chatbot:

- Der Chatbot soll den Nutzer durch die Vorbereitungen für die Ausführung eines Modells leiten mit Hilfe relevanter Dokumentation.
- Der Chatbot soll den Nutzer durch die Ausführung des Modells leiten, bis es zu einem Ergebnis kommt mit Hilfe relevanter Dokumentation
- Der Chatbot hilft dem Nutzer bei der Erstellung eines eigenen Algorithmus bzw. Dem Hochladen seines eigenen Algorithmus
- Empfehlung relevanter Dokumentationen basierend auf der Problemstellung des Nutzers.
- Unterstützung als Navigationshelfer innerhalb der Plattform.
- Analyse des vom Nutzer beschriebenen Algorithmus und Identifizierung möglicher vorhandener Anwendungsfälle.
  - Gegenüberstellung verschiedener Use Cases mit Erläuterung ihrer Stärken und Vorteile zur besseren Entscheidungsfindung.
- Hilfestellung bei der Speicherung und Verwaltung erstellter Modelle zur flexiblen Wiederverwendung und Optimierung.

# Wo werden die Nutzer den Chatbot verwenden?





## Ausgangssituation:

Als Tech Lead bei Lufthansa Consulting stehe ich vor der Herausforderung, die Flugsteuerung effizienter zu gestalten. Die aktuellen Ressourcen reichen nicht mehr aus, um den steigenden Anforderungen gerecht zu werden.

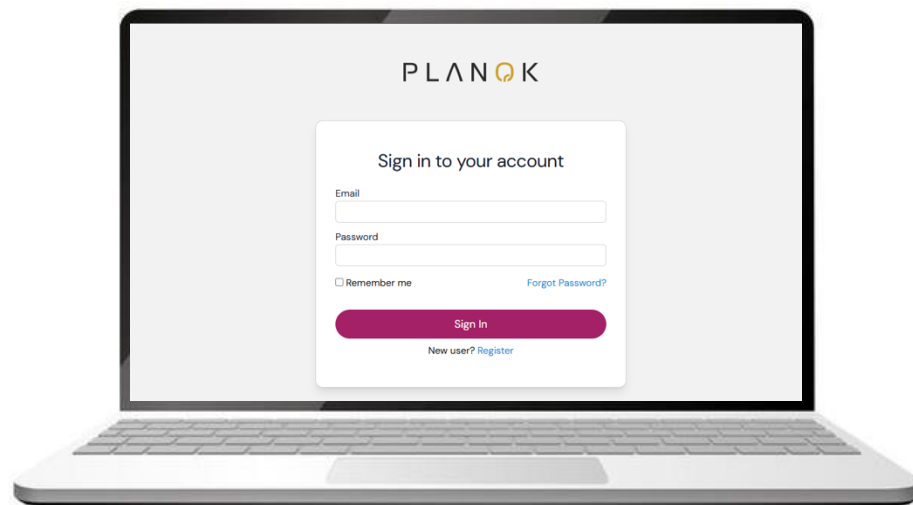
## Herausforderung:

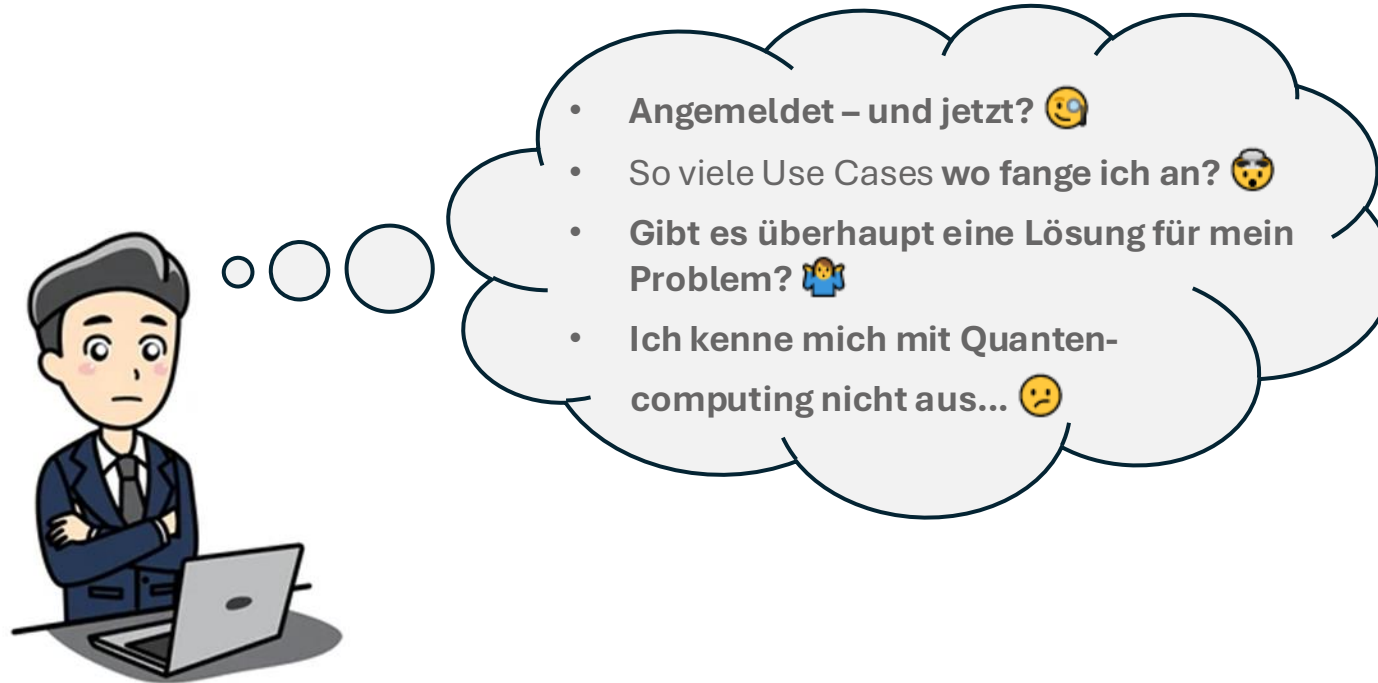
Die Flugsteuerung ist ein hochkomplexes System, das zahlreiche dynamische Faktoren berücksichtigen muss:

- **Flugrouten:** Ständig wechselnde Flugpläne und Streckenführungen
- **Wetterbedingungen:** Unvorhersehbare Einflüsse auf die Routenwahl und Sicherheit
- **Verkehrsdichte:** Steigende Anzahl an Flugbewegungen erfordert optimale Koordination

# AI Chatbot auf der PlanQK Plattform

Bei seiner Suche stößt er auf KIPU Quantum und die PlanQK-Plattform. Neugierig meldet er sich an, ist aber unsicher – *Quantencomputing* ist für ihn völliges Neuland.

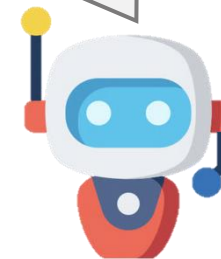






Willkommen auf der PlanQK Plattform!

Ich bin dein persönlicher Assistent und unterstütze dich bei der Suche und Implementierung von Use Cases. Wobei kann ich dich unterstützen? ...





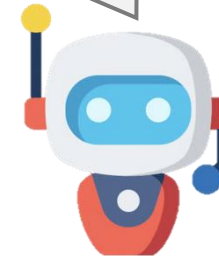
# AI Chatbot auf der PlanQK Plattform



Ich bin Tech Lead bei Lufthansa Consulting. Wir suchen nach einer Möglichkeit, die Herausforderungen bei der Bereitstellung ausreichender Luftraumkapazität zu lösen, um eine Luftraumüberlastung zu vermeiden.

Willkommen auf der PlanQK Plattform!

Ich bin dein persönlicher Assistent und unterstütze dich bei der Suche und Implementierung von Use Cases. Wobei kann ich dich unterstützen?

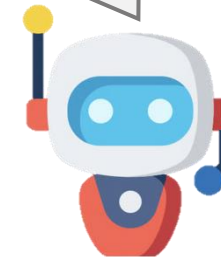


# AI Chatbot auf der PlanQK Plattform



Ich bin Tech Lead bei Lufthansa Consulting. Wir suchen nach einer Möglichkeit, die Herausforderungen bei der Bereitstellung ausreichender Luftraumkapazität zu lösen, um eine Luftraumüberlastung zu vermeiden.

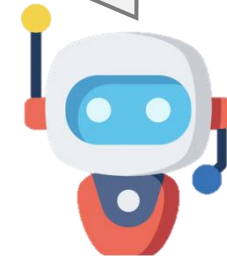
Das Problem des Air Traffic Flow Management (ATFM) wird im ATFM Use Case auf PlanQK beschrieben. Möchten Sie mehr über die wichtigsten Konzepte erfahren?





Ja, gerne!

Das Problem des Air Traffic Flow Management (ATFM) wird im ATFM Use Case auf PlanQK beschrieben. Möchten Sie mehr über die wichtigsten Konzepte erfahren?



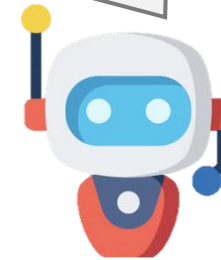


Ja, gerne!

Hier die wichtigsten Konzepte sind:

- Optimierung der Luftraumnutzung zur Vermeidung von Überlastung und Verspätungen
- Predictive Analytics mit Quantum Computing für präzisere Flugroutenoptimierung
- Dynamische Anpassungen basierend auf Echtzeit-Verkehrsdaten

Mit KIPU's Quantum-Optimierungsalgorithmen (DCQO) können größere Probleme gelöst werden als mit klassischen Methoden. Der Service ist auf unserem Marketplace verfügbar: [[Link zum ATFM Use Case](#)]. Möchten Sie mehr Informationen?



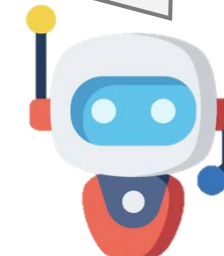


Wie rufe ich den Service mit Python auf?

Sie können den ATFM-Optimierungsservice in Ihre bestehenden Workflows integrieren, indem Sie den folgenden Python-Code verwenden:

```
python Kopieren Bearbeiten  
  
import requests  
  
url = "https://api.planqk.de/kipu/atfm"  
headers = {"Authorization": "Bearer YOUR_ACCESS_TOKEN"}  
data = {"airspace_data": "your_input_data"}  
  
response = requests.post(url, json=data, headers=headers)  
print(response.json())
```

Um Ihren Zugangstoken zu erhalten, folgen Sie bitte dieser [\[Authentifizierungsanleitung\]](#). Lassen Sie mich wissen, wenn Sie weitere Unterstützung benötigen!

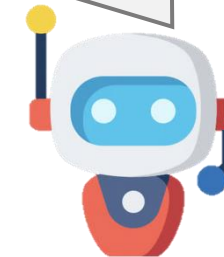





Kann ich die Parameter für verschiedene Szenarien anpassen?

Ja! Die API ermöglicht es, verschiedene Parameter anzupassen, wie zum Beispiel:

- Verkehrsdichte (z. B. hoch/mittel/niedrig)
- Wetterbedingungen (z. B. Windgeschwindigkeit, Turbulenzen)
- Flugumleitungsbeschränkungen (z. B. Sperrgebiete, Höhenbegrenzungen) Möchten Sie einen Link zur detaillierten API-Dokumentation?



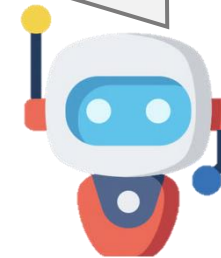




Super, Danke! Erstmal nicht, das gebe ich an meine Kollegen weiter!

Ja! Die API ermöglicht es, verschiedene Parameter anzupassen, wie zum Beispiel:

- Verkehrsdichte (z. B. hoch/mittel/niedrig)
- Wetterbedingungen (z. B. Windgeschwindigkeit, Turbulenzen)
- Flugumleitungsbeschränkungen (z. B. Sperrgebiete, Höhenbegrenzungen) Möchten Sie einen Link zur detaillierten API-Dokumentation?



02

# Retrieval Augmented Generation

# Retrieval Augmented Generation

RAG (Retrieval Augmented Generation) kombiniert KI-Textgenerierung mit einer gezielten Suche in externen Wissensquellen, um genauere, aktuellere und transparentere Antworten zu liefern.

## Limitierungen klassischer LLMs

- ✗ Halluzinationen & Wahrheitsgehalt
- ✗ Begrenztes Wissen
- ✗ Technische Einschränkungen
- ✗ Fehlendes Gedächtnis
- ✗ Stochastisches Verhalten
- ✗ Ressourcenintensität

## Kernbausteine von RAG

### *Indexing* – Vom Rohtext zum KI-Format:

- Datenaufbereitung und -transformation
- Chunking (Textsegmentierung)
- Encoding & Vektorisierung

### *Retrieval* – Relevante Informationen finden

- Kodierung der Nutzeranfrage
- Dokumenten-Retrieval

### *Generation* – KI erstellt die Antwort

- Gefundene Dokumente werden gewichtet und als Kontext ins Modell eingespeist, das daraus die Antwort generiert.

## Vorteile von RAG

- ✓ Verbesserte Textqualität durch Kontext
- ✓ Reduktion von Halluzinationen
- ✓ Erhöhte Transparenz & Quellenangaben
- ✓ Integration und spezialisierter oder privater Daten
- ✓ Weniger aufwendiges Fine-Tuning

## Typische Probleme beim Lernen

### Zu viel Material, zu wenig Überblick

Skripte, Bücher, Vorlesungsfolien, Online-Artikel  
– wo findet man die richtige Info?



### Verstreutes Wissen

Mühsames Zusammensuchen von Infos aus  
verschiedenen Quellen



### Standard-Suchmaschinen sind ungenau

Google liefert tausende Treffer, aber oft nicht  
konkret das, was man braucht



### Fragen bleiben offen

Oft beantworteten Lektüren / Vorlesungen keine  
spezifischen Fragen



## Wie ein RAG-System hier unterstützen kann

### Fragen statt Suchen

Beim Stellen einer konkreten Frage, erhält man  
eine Antwort aus relevanten Quellen



### Kombiniert Wissen aus verschiedenen Materialien

Skripte, Bücher und Notizen werden zu einer  
kohärenten Antwort verbunden



### Spart Zeit

Kein stundenlanges Durchforsten von PDFs  
und Webseiten

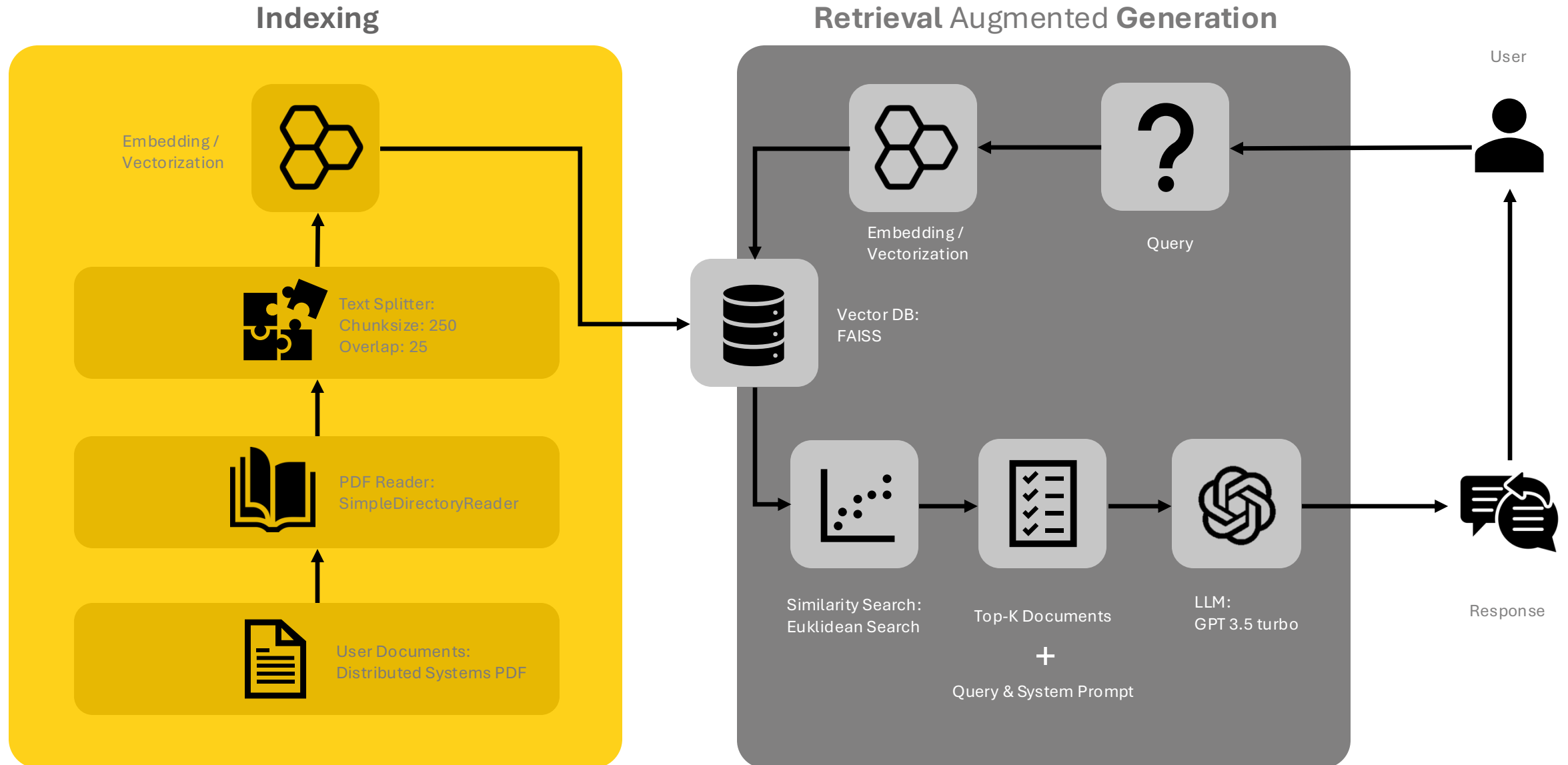


### Liefert präzisere Antworten

Lernunterlagen werden als Grundlage genutzt



Lernen mit RAG ist wie ein persönlicher Tutor, der **Lernmaterialien versteht**  
und daraus präzise Antworten ableitet



03

Ausblick



## Warum Cloud?

### Skalierbarkeit:

Automatische  
Anpassung an  
Nutzeranfragen

### Kostenoptimierung:

Pay-as-you-go-Modelle,  
nur zahlen, was genutzt  
wird

### Sicherheit & Compliance

Vereinfachte Absicherung der  
Zugriffe und Einhaltung  
unternehmensinterner Richtlinien  
als auch gesetzlicher Vorgaben wie  
der Datenhoheit oder DSGVO

### Endpunkt-



### Bereitstellung

Dauerhaft erreichbare  
Endpunkte vorhanden

### Einfache Wartung

Updates & Patches  
werden mit möglichst  
wenigen manuelle  
Eingriffe möglich

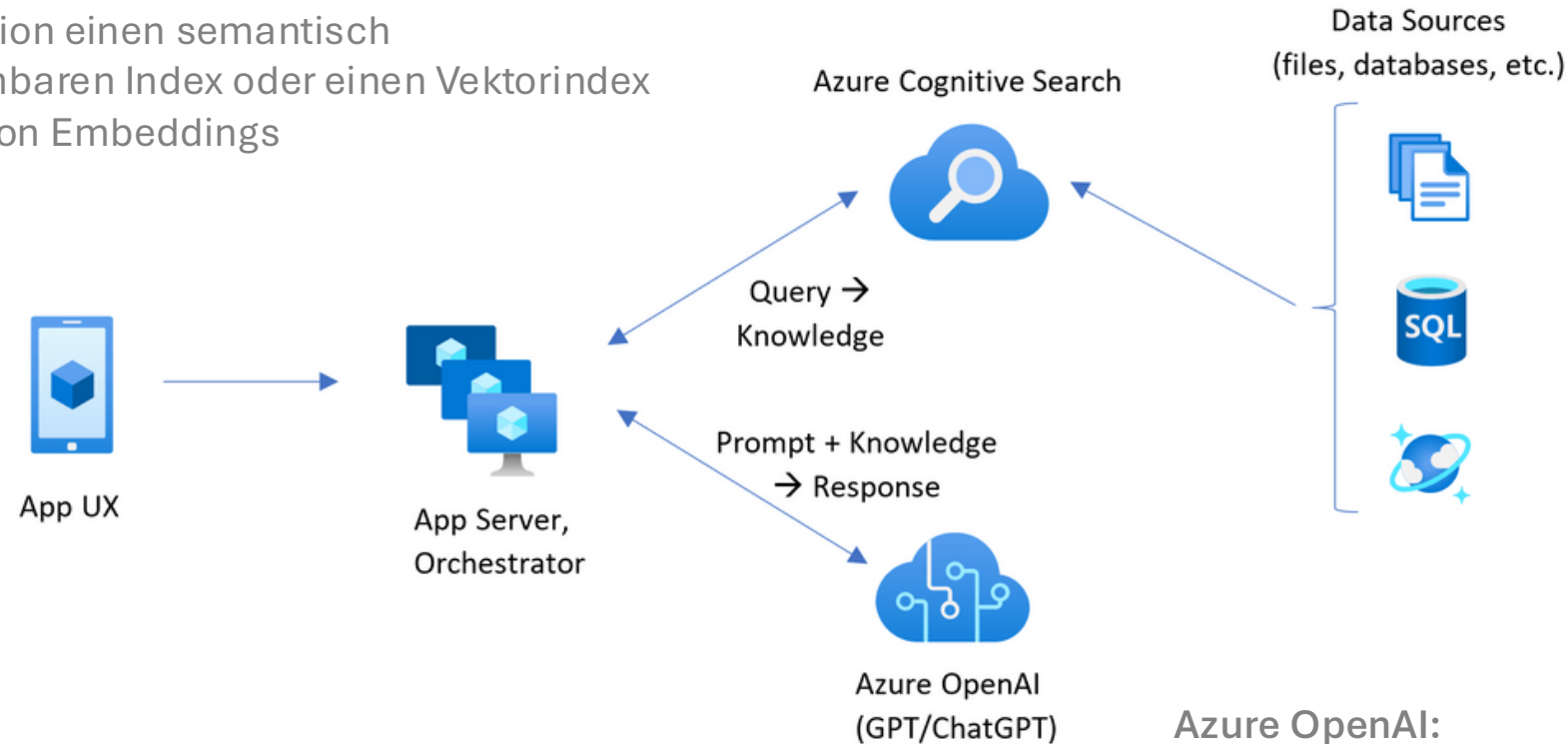
# Azure vs. AWS als Cloud-Anbieter

| Kriterium      |                   |  |
|----------------|--|---|
| LLM-Angebot    | Diverse KI-Ressourcen für unterschiedliche Use-Cases lassen sich hosten und verknüpfen               | Amazon Bedrock als zentrale Plattform für KI-Lösungen                               |
| Kostenstruktur | Flexible Tarife mit stündlicher oder abfragebasierter Abrechnung; LLMaaS                             | Breites und günstiges LLMaaS Angebot von unterschiedlichen Anbietern                |
| Sicherheit     | Viele EU-Rechenzentren und starken Enterprise-Fokus im Cybersecurity-Bereich                         | Größerer Fokus auf US-Markt mit starker Absicherung der genutzten Ressourcen        |
| Integration    | Nahtlose Verbindung mit vielen Microsoft Enterprise-Plattformen und Management-Systemen wie Entra ID | Weniger direkte Integrationen verfügbar   |

# Mögliche (vereinfachte) Architektur einer Azure Umsetzung

## Azure Cognitive Search:

Erzeugt aus den unsortierten Daten je nach Konfiguration einen semantisch durchsuchbaren Index oder einen Vektorindex mit Hilfe von Embeddings



## Azure OpenAI:

Stellt Basismodelle von OpenAI für die Interpretation und Verarbeitung der erhaltenen Daten bereit



## Azure Computer Vision

OCR und Image Recognition für die Erkennung von Dokumenten und Inhalten in Form von Bilddateien



## Azure Speech

Menschliche Spracheingabe und -Ausgabe (Speech-to-Text, Text-to-Speech)



## Azure Bot Services

Einfache Bereitstellung der Chat-Applikation als Teams-App, Web-App, über Copilot etc.



## Azure ML Workspace

Optimierung und Training von diversen KI-Modellen auf leistungsfähigen GPUs

Prompting Patterns sind Methoden zur Eingabenverarbeitung in LLMs, um präzisere Antworten zu erhalten.

## Limitierungen von LLMs mit RAG

- ✗ Fehlinterpretationen
- ✗ Abhängig von hinterlegten Dokumenten
- ✗ Schwierige Feinabstimmung
- ✗ Fehlendes Gedächtnis

## Vorteile von Prompting Patterns

- ✓ Verbesserte Antwortqualität durch Erkennen von Kontext
- ✓ Reduktion von Halluzinationen
- ✓ Erhöhte Transparenz
- ✓ Nutzung von zusätzlichen Daten sofern sinnvoll
- ✓ Reduzierter Kontext notwendig

## Beispiele

### *Reason & Act (ReAct):*

- Nachdenken
- Zusätzliche Quellen nutzen
- Anpassung der Handlung / Antwort
- Handeln

### *Chain of Thought (CoT):*

- Nachdenken
- Aufteilen der eingegeben Informationen
- Antwort mit Gedankengang

# Danke