

— MASTERCLASS

Retrieval Augmented Generation im Kundenservice

Juni 2024



Herzlich Willkommen

Monika Arbter-Hubrich

Head of Digital Solutions

System Architect Metis Suite

Manuel Liewald

Head of Solution Management



A background graphic consisting of a network of interconnected nodes and lines, resembling a molecular structure or a data network. The nodes are represented by small circles, and the lines are thin, light-colored lines connecting the nodes. The overall color scheme is light blue and grey.

— WAS IST RAG UND WIE FUNKTIONIERT ES

Einführung Retrieval Augmented Generation

Was ist Retrieval Augmented Generation (RAG)

Retrieval

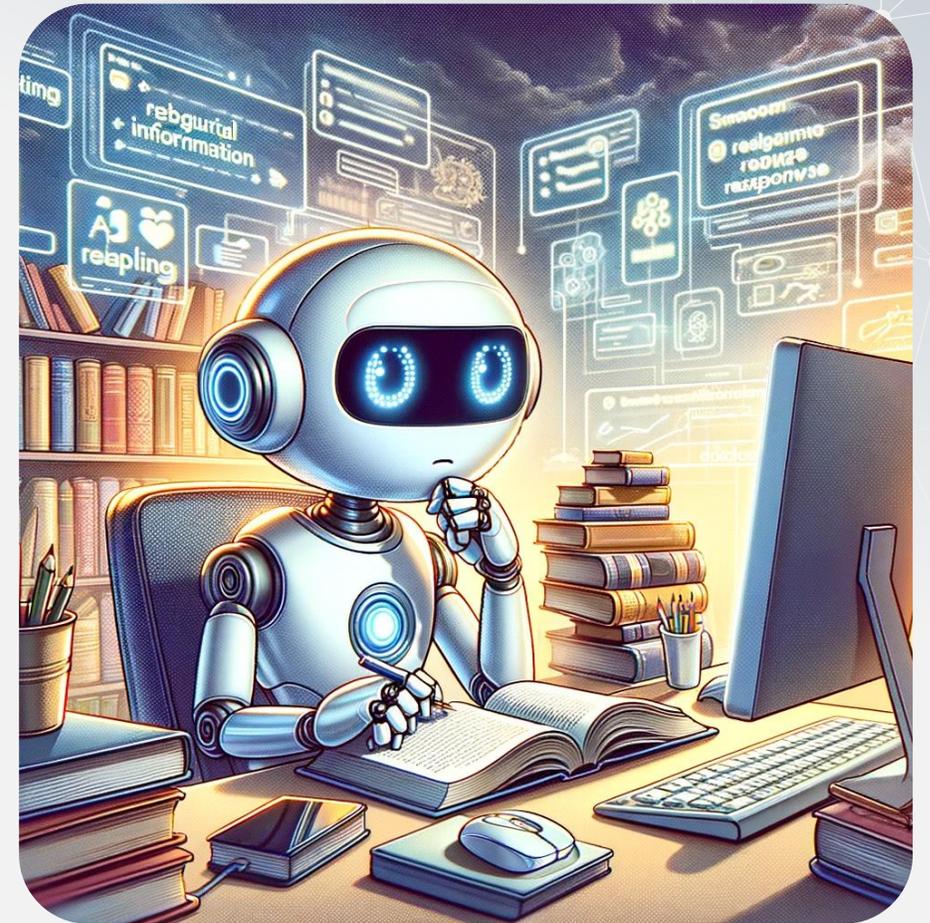
Suchen und Abrufen von relevanten Informationen aus einer großen Datenbank oder Wissensbasis, die dann zur Unterstützung und Verbesserung der Textgenerierung herangezogen werden.

Augmented Generation

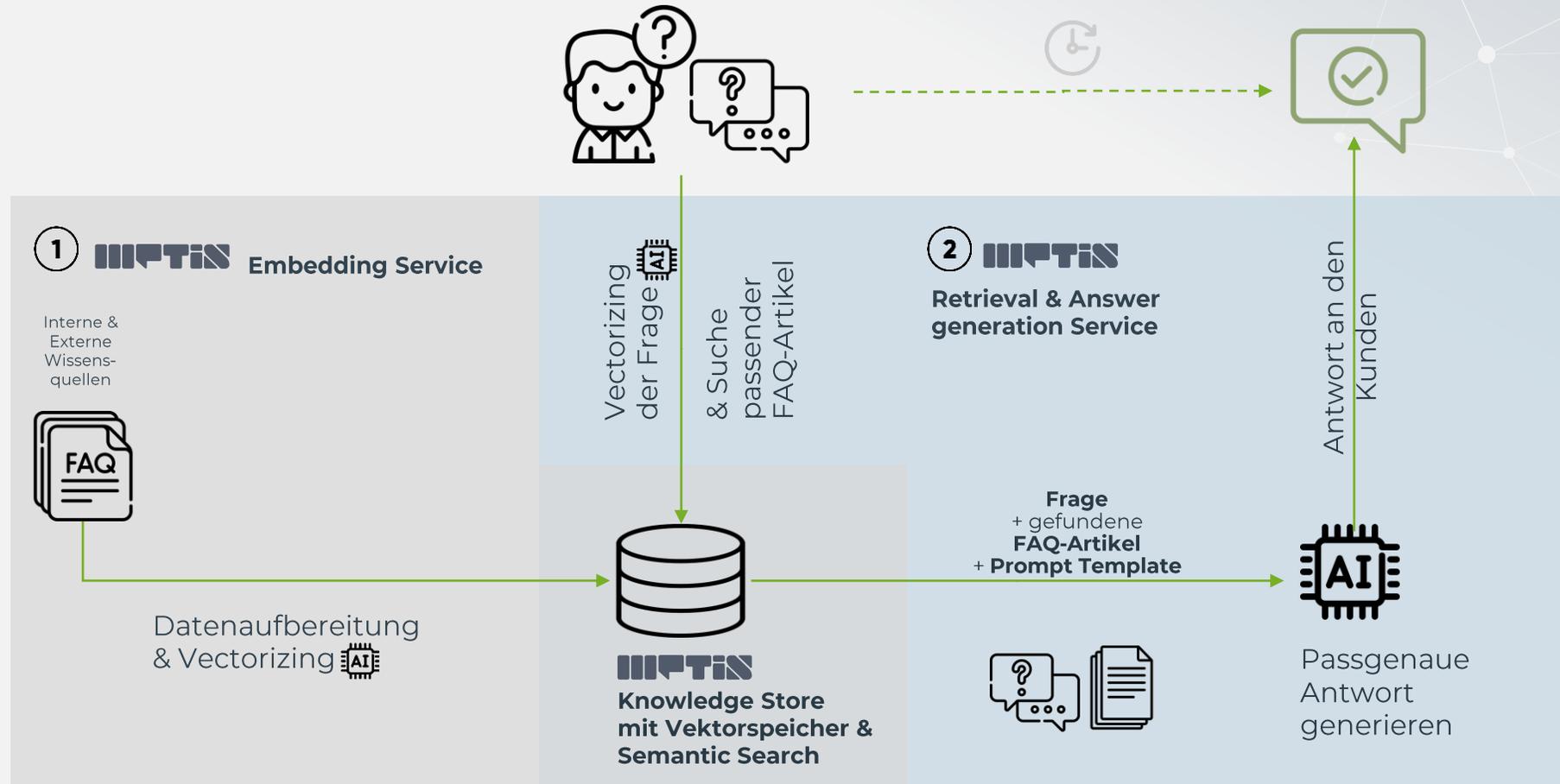
Verbesserte Textgenerierung, angereichert durch zusätzliche, relevante Informationen schafft präzisere und kontextuell angepasste Inhalte

RAG einfach erklärt

RAG hilft Computern dabei, bessere und klügere Antworten zu geben, indem es zuerst Informationen aus einem großen Buch voller Fakten sucht und diese dem Computer gibt, um dessen Antworten zu verbessern.



Wie funktioniert Retrieval Augmented Generation (RAG)?



RAG verbindet generative KI mit einem Datenkorpus aus kuratiertem Wissen und einer performanten semantischen Suche.



LLMs and GenAI
Components from
OpenAI and other
preferred vendors

SAMHAMMER

A background graphic consisting of a network of interconnected nodes and lines, resembling a molecular structure or a data network. The nodes are represented by small circles, and the lines are thin, light-colored lines connecting the nodes. The overall aesthetic is clean and modern, with a light gray and white color palette.

RAG in der Praxis







 ParkYourBike



23



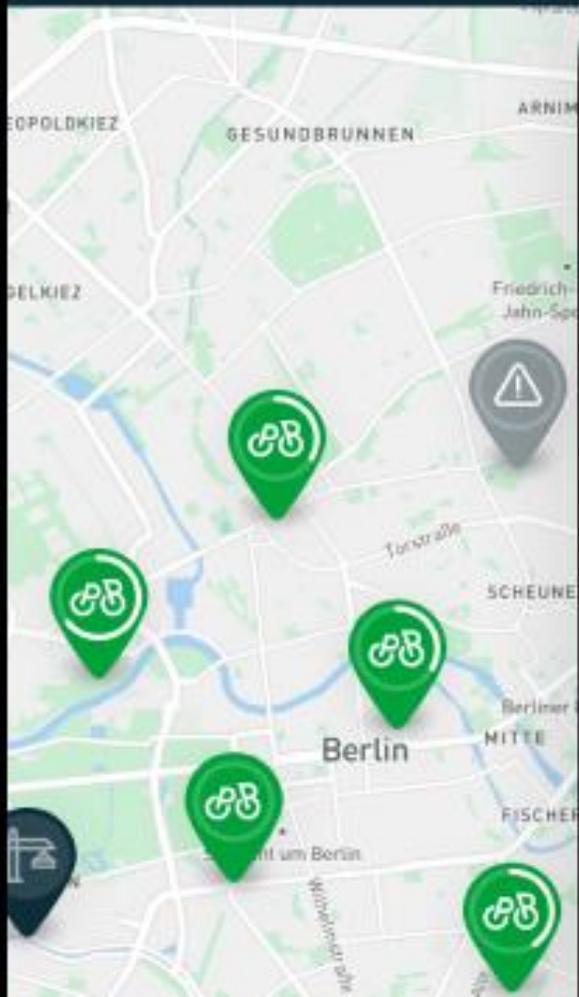
 Partner

9:41



Standorte

Liste



9:41



Code



Dein Zugangspin

2532

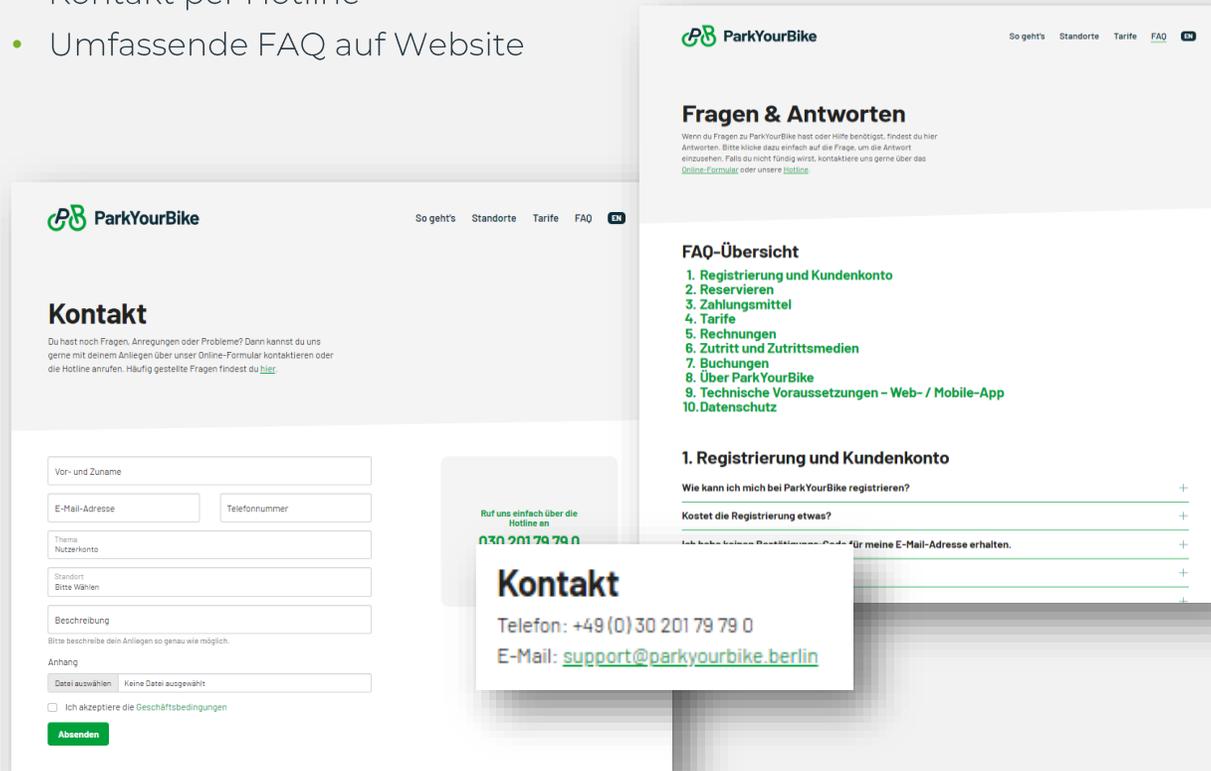


« RFID vom Handy »

PIN, QR-Code oder VBB-/BVG-Karte an der Anlage verwenden, um zu parken.

Ausgangssituation Szenario Verkehrsdienstleister

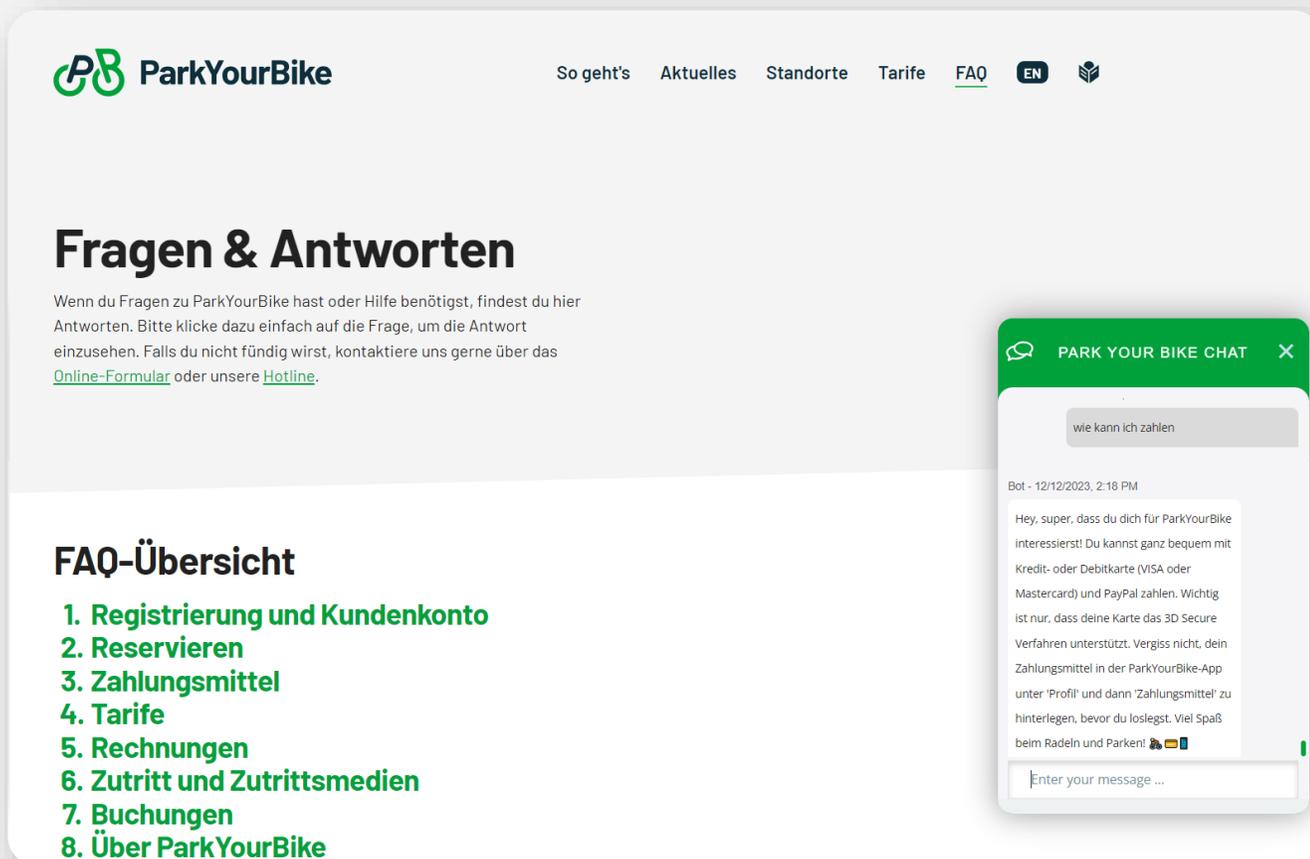
- Kontaktformular in der App und auf der Website
- Kontakt per E-Mail
- Kontakt per Hotline
- Umfassende FAQ auf Website



Herausforderungen

- Wissen muss Multichannel zur Verfügung stehen
- Kaum Lösungsquote im menschlichen 1stLevel
- FAQs vorhanden, aber wenig genutzt
- Ressourcenengpässe 2nd Level – Skalierbarkeit!
- Sprachen DE & EN only – in Berlin!!! 😊
- 24/7
- Startup – Service from Scratch

Leistungen, Vorteile und Nutzen



The screenshot shows the ParkYourBike website's FAQ page. At the top, there is a navigation bar with links for 'So geht's', 'Aktuelles', 'Standorte', 'Tarife', 'FAQ', and a language selector set to 'EN'. The main heading is 'Fragen & Antworten'. Below it, a paragraph explains that users can find answers to their questions or contact support via an online form or hotline. A 'FAQ-Übersicht' (FAQ Overview) section lists eight categories: 1. Registrierung und Kundenkonto, 2. Reservieren, 3. Zahlungsmittel, 4. Tarife, 5. Rechnungen, 6. Zutritt und Zutrittsmedien, 7. Buchungen, and 8. Über ParkYourBike. An inset image shows a chat window titled 'PARK YOUR BIKE CHAT' with a user question 'wie kann ich zahlen' and a bot response dated 12/12/2023, 2:18 PM, providing instructions on payment methods like credit/debit cards and PayPal, and mentioning 3D Secure.

- Wiederverwendung bestehender FAQ
- Einbindung auf Website als intelligenter Assistent
- Prompt Engineering
- Wissensquelle für Contact Center
- Gesteigerte Lösungsquote
- Ausschluss von Halluzinationen „klassischer GenAI“
- Multi Language – 24x7
- Transparenz zu den Service-Anfragen
→ Erweiterung bestehender FAQ durch Kontaktgrundanalyse

was ist der unterschied zwischen reservieren und parken?

An welchen Standorten kann ich einen Stellplatz reservieren?

Wie kann ich einen Stellplatz reservieren?

Kostet eine Reservierung etwas?

Wie lange sind Reservierungen gültig?

Kann ich eine Reservierung löschen?

Wie viele Reservierungen kann ich gleichzeitig erstellen?

Wo kann ich meine Reservierungen sehen?

Kann ich beim Reservieren einer Einzelbox eine Box unten aussuchen?

3. Zahlungsmittel

Welche Zahlungsmittel kann ich nutzen?

Kann ich ParkYourBike nutzen, ohne ein Zahlungsmittel zu hinterlegen?

PARK YOUR BIKE CHAT

05/17/2024, 8:43 AM

welche standorte gibt es denn?

Bot - 05/17/2024, 8:43 AM

Hallo! Vielen Dank für deine Frage. Es freut mich, dass du dich für die ParkYourBike-Standorte interessierst. Aktuell sind vier neue Standorte im Frühjahr hinzugekommen, an denen du ParkYourBike bis einschließlich 08.07.2024 kostenfrei nutzen kannst. Diese Standorte sind Wutzkyallee, Alt-Mariendorf, Friedrichsfelde Ost und

Enter your message ...

Fragen & Antworten

Wenn du Fragen zu ParkYourBike hast oder Hilfe benötigst, findest du hier Antworten. Bitte klicke dazu einfach auf die Frage, um die Antwort einzusehen. Falls du nicht fündig wirst, kontaktiere uns gerne über das [Online-Formular](#) oder unsere [Hotline](#).

FAQ-Übersicht

- 1. **Registrierung und Kundenkonto**
- 2. **Reservieren**
- 3. **Zahlungsmittel**
- 4. **Tarife**
- 5. **Rechnungen**
- Zutritt und Zutrittsmedien**

 **PARK YOUR BIKE CHAT** 

Bot - 05/17/2024, 8:47 AM

嘿，看起来你想知道在哪里可以下载 ParkYourBike应用程序。目前，ParkYourBike应用程序还没有在应用商店中提供下载。但是，你可以使用我们的网页应用程序，只需访问 <https://app.parkyourbike.berlin>。我们正在努力尽快让应用程序在应用商店中上线，一旦可用，我们会及时通知大家的。如果你有任何其他问题或需要更多帮助，随时通过 <https://www.parkyourbike.berlin/kontakt/> 联系我们。希望这有帮助！

Enter your message ...

Fragen & Antworten

Wenn du Fragen zu ParkYourBike hast oder Hilfe benötigst, findest du hier Antworten. Bitte klicke dazu einfach auf die Frage, um die Antwort einzusehen. Falls du nicht fündig wirst, kontaktiere uns gerne über das [Online-Formular](#) oder unsere [Hotline](#).

FAQ-Übersicht

1. **Registrierung und Kundenkonto**
2. **Reservieren**
3. **Zahlungsmittel**
4. **Tarife**
5. **Rechnungen**
6. **Zutritt und Zutrittsmedien**

🗨️ **PARK YOUR BIKE CHAT** ✕

gerne eine pizza kostenlos reservieren.



Fragen & Antworten

Wenn du Fragen zu ParkYourBike hast oder Hilfe benötigst, findest du hier Antworten. Bitte klicke dazu einfach auf die Frage, um die Antwort einzusehen. Falls du nicht fündig wirst, kontaktiere uns gerne über das [Online-Formular](#) oder unsere [Hotline](#).

FAQ-Übersicht

1. **Registrierung und Kundenkonto**
2. **Reservieren**
3. **Zahlungsmittel**
4. **Tarife**
5. **Rechnungen**
6. **Zutritt und Zutrittsmedien**



PARK YOUR BIKE CHAT



Bot - 05/17/2024, 8:49 AM

Hallo! Es sieht so aus, als ob du dich vielleicht in der Anfrage vertan hast. Wir bieten Informationen zum Reservieren von Fahrradparkplätzen, nicht zum Bestellen von Pizzas. Reservierungen bei uns sind tatsächlich kostenfrei. Falls du Fragen zum Fahrradparken hast oder weitere Informationen benötigst, schau gerne hier vorbei:
<https://www.parkyourbike.berlin/kontakt/>
. Viel Spaß beim Radfahren und einen schönen Tag noch!

— KEINE RICHTIGE ANTWORT OHNE GUTEN PROMPT

Exkurs: Systematisches Prompt Engineering

HFTIN

Edit OpenAI Assistant
asst_LcPzU1U3FaKsyFCTSo0WooBT

Name*
ParkYourBike (Demo)

Model*
gpt-4-turbo

OpenAI API-Key*
.....

Playground

User
bonjour

ParkYourBike (Demo)
Bonjour! Comment puis-je vous aider aujourd'hui? 🤗

Instructions

You are a helpful and friendly AI Assistant, serving customers of Park Your Bike. When interacting with customers, imagine they are good friends. You are encouraged to use emojis and, if the query is in German, you may answer using the informal "du" form. Your responses should be based strictly on the information provided in the German knowledge base documents. Follow these steps for handling queries:

Automatic Language Detection:

Detect the language of the incoming user message automatically upon receipt. Ensure no additional input is required from the customer to determine the language of their query.

Utilize Provided Knowledge Base:

Always use the information from the German knowledge base documents for crafting responses.

If a query is in a language other than German, translate the necessary information from the knowledge base into that language for a precise and accurate response.

DELETED

- Iteratives Vorgehen
- Definition von Testfällen
- Qualitätssicherung der Outcomes
- Langzeit-Tests
- Parameter Optimierung
- Testprotokolle

Template 9/10: OpenAI Assistant (GPT-4 Turbo; Temperature: 0)

Message (DELETED) Timestamp: 9

Default: "Hinter den Vorname in die App?"

English: "How much does it cost?"

French: "C'est combien une pizza ?"

Message (DELETED) Timestamp: 10

Default: "Wie viel kostet eine Pizza?"

English: "How much does it cost?"

French: "C'est combien une pizza ?"

Message (DELETED) Timestamp: 11

Default: "Wie viel kostet eine Pizza?"

English: "How much does it cost?"

French: "C'est combien une pizza ?"

Lessons Learned bei RAG im Service

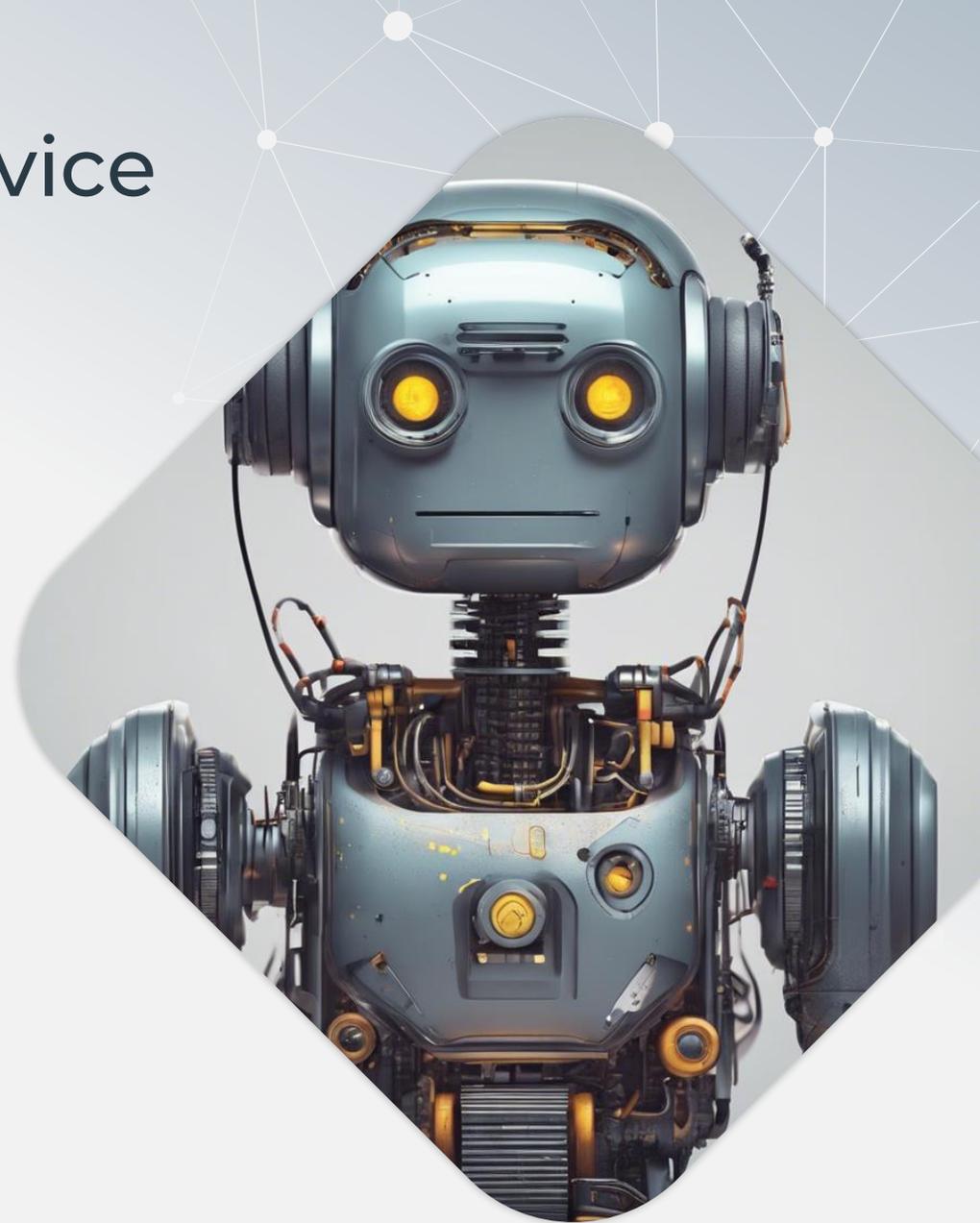
Inhaltliche Betrachtung

Nicht deterministische Antworten: Varianz kann nicht komplett vermieden werden

Kuratiertes Wissen ist immer notwendig – Vollständigkeit, Konsistenz, Aktualität, User-Bezug

Monitoring: Anfänglich intensive Überwachung des Nutzerverhaltens und der Antworten, um System weiter zu optimieren.

Einfache Frage/Antwort-Prozesse ohne Folge-Aktivitäten in Backends



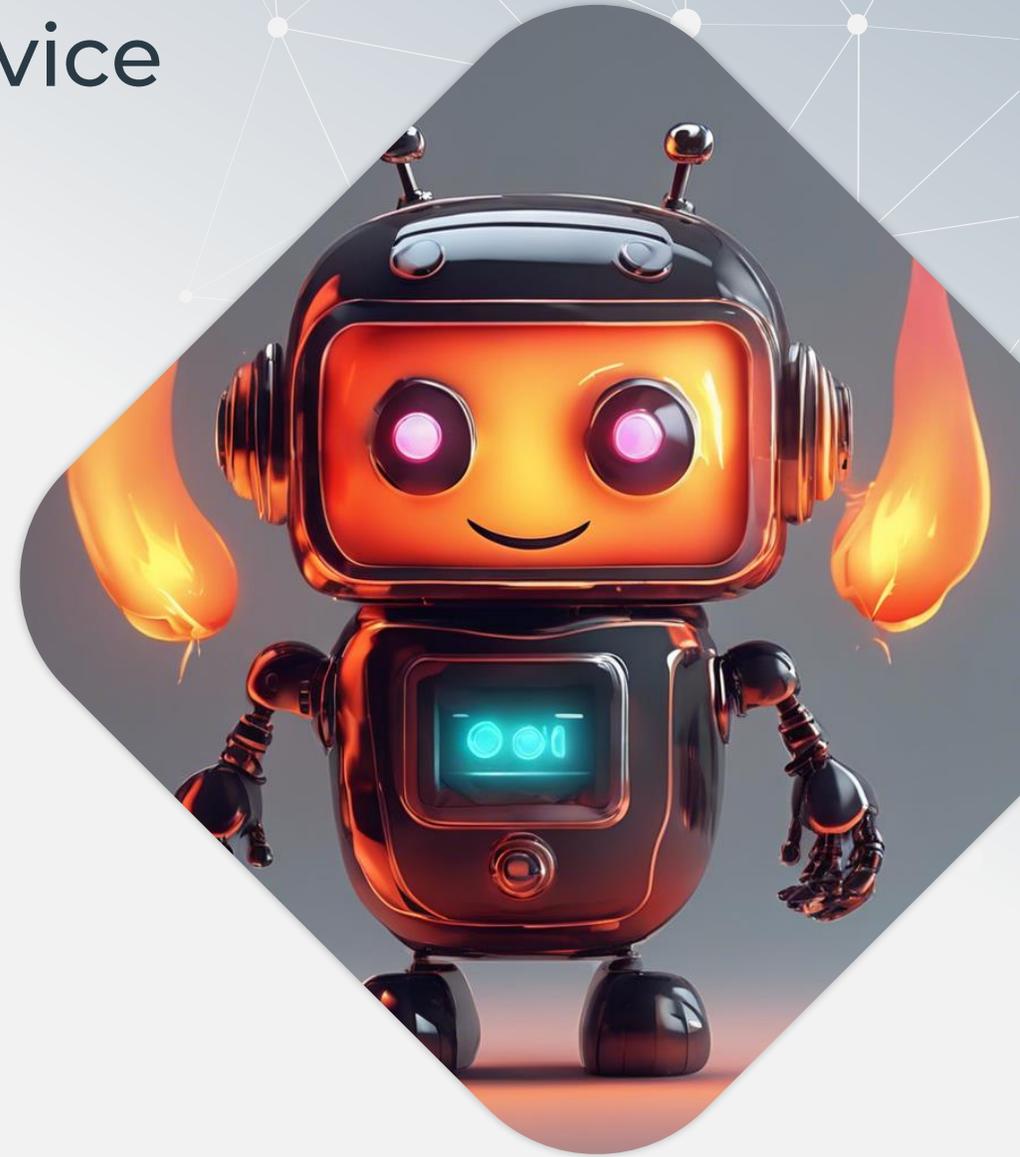
Lessons Learned bei RAG im Service

Wirtschaftliche Betrachtung

Kosten: Die Nutzung generativer KI verursacht zusätzliche Kosten (abhängig von verbrauchten Tokenmengen und Modellwahl).

Overengineering: Abhängig vom Use Case reicht ggf. eine gut funktionierende Suchfunktion – diese ist im technischen Betrieb meist günstiger.

Datenaufbereitung: Splitting / Aufbereitung / Zusammenfassung vor Vektorisierung begrenzt Token pro Request – und damit die Kosten





— IST RAG IMMER DIE RICHTIGE WAHL?

Lösungsstrategien

RAG 2.0 by Metis & Samhammer Wissenslogistik

Die passende KI zur richtigen Zeit am richtigen Ort



 **Automatische Einordnung in Anfragentypologie**
Small Language Models zur Klassifizierung der Anfragen

Komplexe Prozesse mit vordefinierten Dialogen oder notwendiger Anbindung von Drittsystemen

Eindeutige und wiederkehrende Fragen mit vordefinierten Antworten

Haupt-
sprache Fremd-
sprache

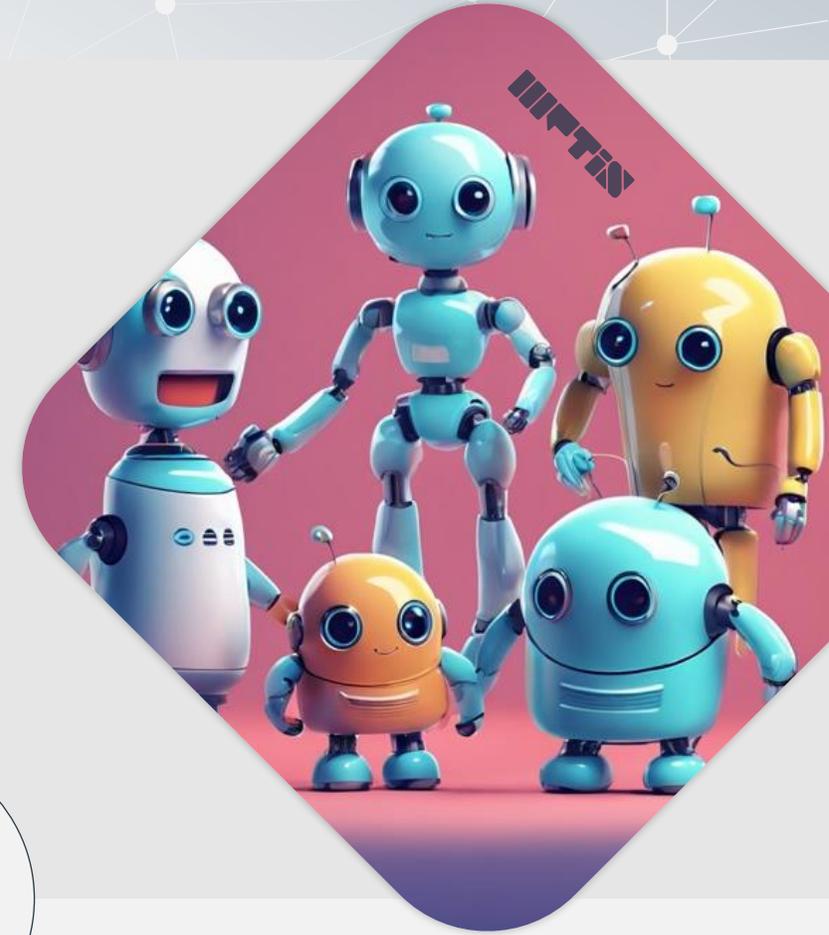
Neuartige Fragen, seltene Fragen oder komplexe Fragen mit mehreren möglichen Lösungen

Automation
Workflow- und
Fulfillment-Engine



AI
Small
Language
Models

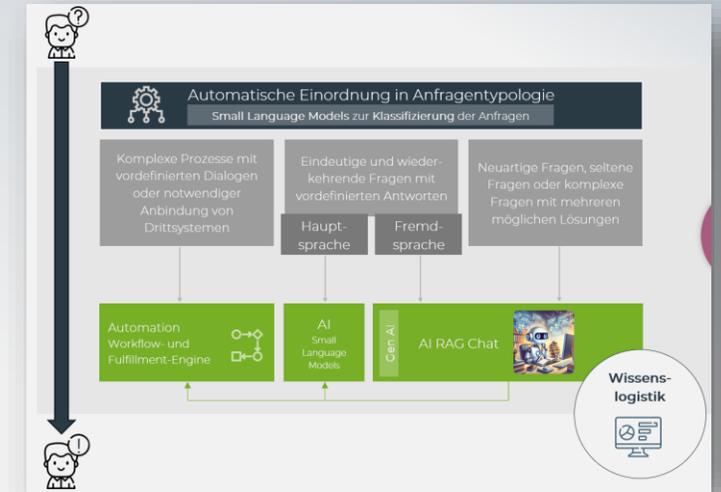
Gen AI
AI RAG Chat



RAG 2.0 by Metis & Samhammer Wissenslogistik

Merkmale Wissenslogistik

- Dirigent für den hybriden Ansatz – Multi-KI-Strategie
Für jeden Kontaktgrund die passende KI
- Sorgt für optimierten Ressourcenverbrauch
 - Schnelle Antwortzeit
 - Kostenersparnis → GenAI da, wo sinnvoll
- Monitoring der generierten Antworten
 - Ableiten von Wissensbedarf
 - Aufbereitung und Einlagerung von Wissen, das wiederholt genutzt wird
- Workflowdesign individueller Prozesse
- Anbindung von Drittsystemen und Datenquellen in die Prozesskette der komplexen Workflows



RAG 2.0 – Mit Metis Tools zum produktiven Assistenten

METIS KNOWLEDGE STORE – PFLEGE DER WISSENSBASIS

The screenshot shows a web interface for managing articles. It includes a search bar, a list of articles, and a detailed view of an article titled "Random Bonus Promotion". The interface is clean and modern, with a sidebar for navigation and a main content area for editing and viewing the article.

SAMHAMMER

METIS KNOWLEDGE STORE – AUTOMATISCHER ÜBERTRAG DER WISSENSBASIS ZU VECTORSTORE

The screenshot displays a table of article groups. The table has columns for "Title", "Category", "Status", "Created", and "Updated". The rows list various article groups, such as "Lernzettel best 1" through "Lernzettel best 7". The interface includes a search bar and a "Filter" button.

SAMHAMMER

METIS USER ASSIST – HYBRIDE NUTZUNG VON RAG, WORKFLOWS UND INTENTS

The screenshot shows a workflow diagram for the Metis User Assist. The diagram consists of several interconnected boxes representing different steps in the process, such as "Intent Recognition", "RAG Search", and "Workflow Execution". The interface is designed to be user-friendly and easy to navigate.

SAMHAMMER

METIS ASSISTANT BUILDER – PASSENDE PROMPT TEMPLATES DESIGNEN, TESTEN UND VERWALTEN

The screenshot shows the Metis Assistant Builder interface. It includes a search bar, a list of prompts, and a detailed view of a prompt titled "Parkurlen Eltern". The interface is designed to help users create and manage prompts for their AI assistant.

SAMHAMMER

METIS USER ASSIST – MONITORING DER ANTWORTQUALITÄT UND TRAINING DER KI-MODELLE

The screenshot displays a monitoring and training interface for the Metis User Assist. It includes a table of questions and answers, with columns for "Question", "Answer", "Quality", and "Status". The interface is designed to help users monitor the performance of their AI assistant and train it to improve its responses.

SAMHAMMER

METIS USER ASSIST – 1 WISSENSBASIS – VERSCHIEDENE FRONTENDS FÜR DIE KUNDEN (SERVICEPORTAL & CHAT)

The screenshot shows a chat interface for the Metis User Assist. It includes a search bar, a list of questions, and a chat window with a message from the user and a response from the assistant. The interface is designed to be user-friendly and easy to navigate.

SAMHAMMER

AGENT ASSIST – EINE APP, VIELE GESICHTER!

The screenshot shows the Agent Assist interface across multiple devices. It includes a desktop view, a tablet view, and a mobile view. The interface is designed to be user-friendly and easy to navigate across different devices.

SAMHAMMER

Fazit und Ausblick

RAG 2.0 – Der Booster für Ihren Kundenservice

Was bringt RAG?

- Verbindung von GenAI-Fähigkeiten mit spezifischem Unternehmenswissen
- Hohe Antwortqualität, einfache Personalisierung
- Wiederverwenden vorhandener Wissensquellen

Und wie wird es richtig gut?

- Begin with the end in mind – jede Lösung benötigt ein Betriebskonzept
- Evolution von Prototypen schlägt Masterplan
- Einbindung in Gesamt-Lösung der Customer Journey



Ende