
Zuverlässiger Einsatz von Large Language Models (LLMs)

Teil 2: Retrieval Augmented Generation (RAG)

Daniel Seifert, Lisa Jöckel, Thorsten Honroth

Ausblick

- **Fraunhofer IESE, Abteilung Data Science**
- **Motivation: Warum RAG?**
- **RAG Grundprinzip**
- **RAG evaluieren**
- **Ist RAG das Richtige für meinen Anwendungsfall?**
- **Ein individuelles RAG-System implementieren – Was ist zu beachten?**
- **RAG vs. Fine-Tuning**
- **Unterstützung durch unsere Data Science Abteilung**

Fraunhofer IESE

Eines von 76 Instituten, die sich mit angewandter Forschung beschäftigen



Gegründet **1996** 
mit Sitz in Kaiserslautern

Über **270**  Mitarbeitende

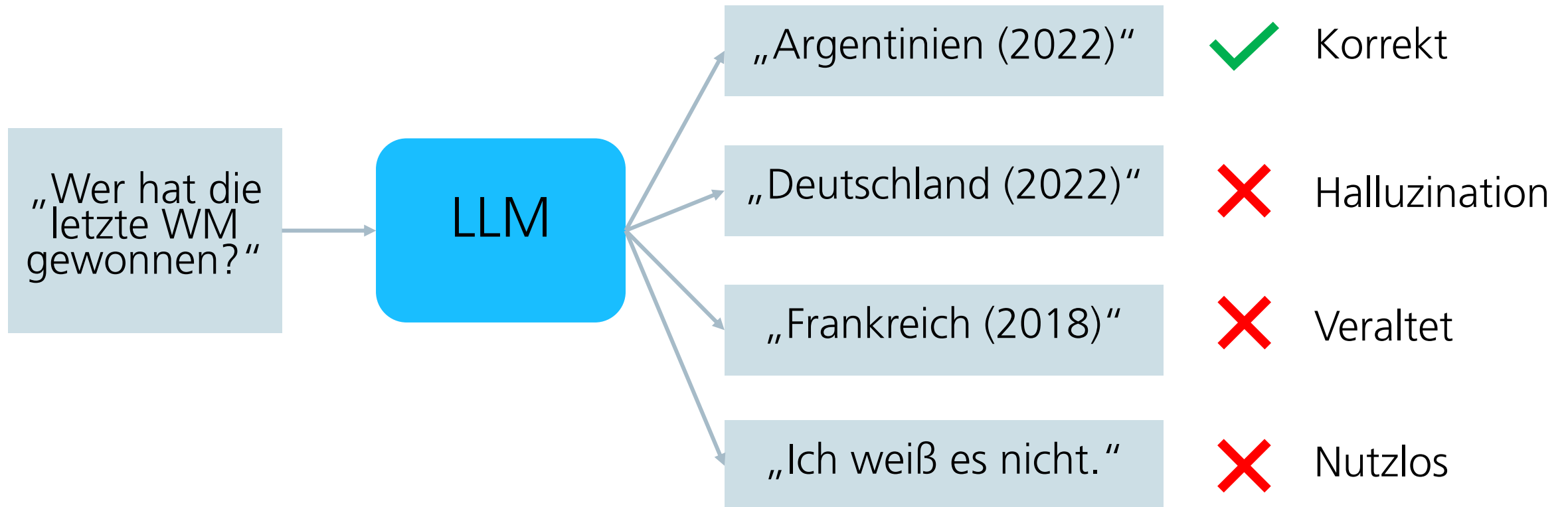


RAG: Einstieg

©iStock.com | Greyfebruary

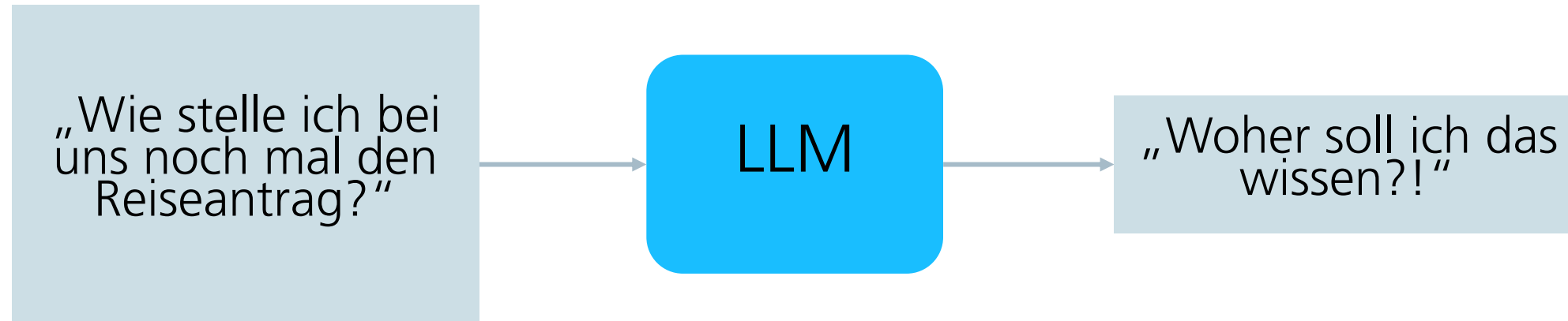
Motivation

Die Antwort eines LLMs kann aus verschiedenen Gründen inkorrekt sein



Motivation

LLMs können nur das Wissen haben, das in den Trainingsdaten steckt



Motivation

Problem

Wir wissen nicht, ob die Antwort des LLM korrekt ist.

Das LLM kann die Antwort gar nicht kennen.

Ursache

↓
Nicht nachvollziehbar, wie das LLM auf seine Antwort kommt.

↓
Das LLM wurde nicht auf den (privaten) Daten trainiert.

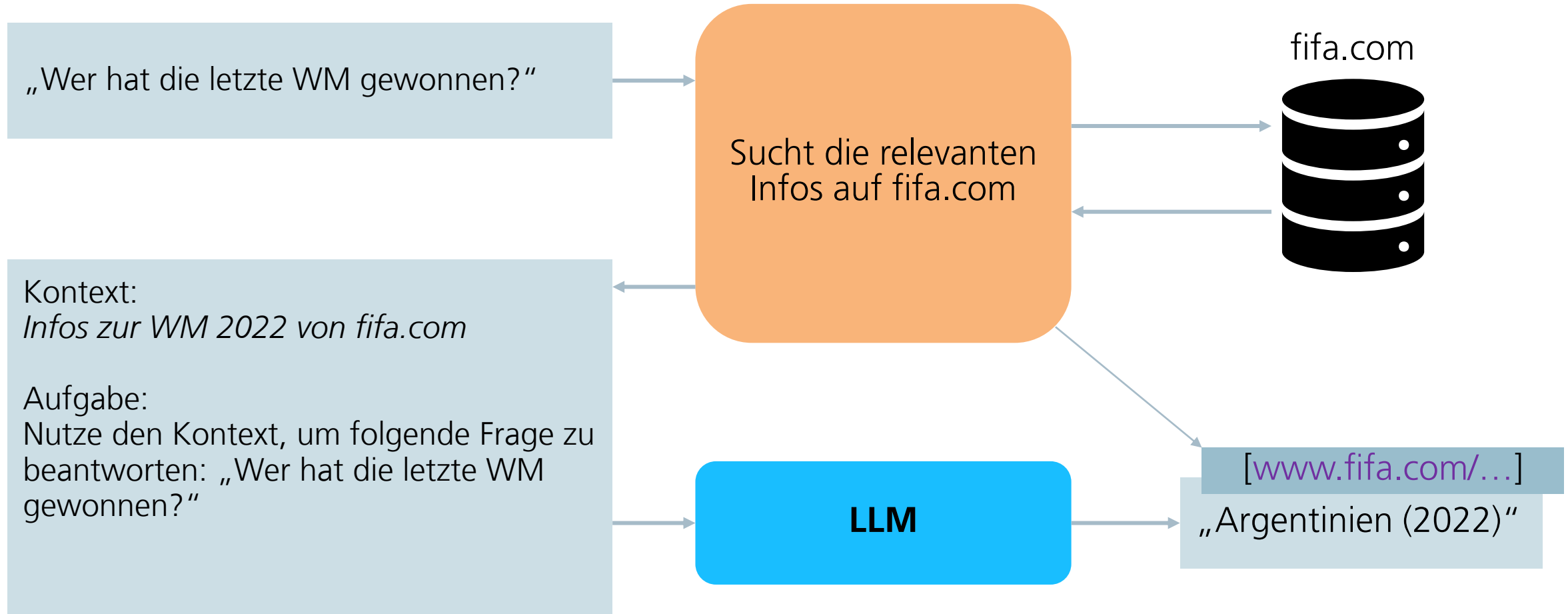
Lösung

Wir können die Antwort nicht überprüfen. Keine Quellenangaben.

↓
Dem LLM selbst die Informationsquelle bereitstellen, damit es die Anfrage verlässlich bearbeiten kann.

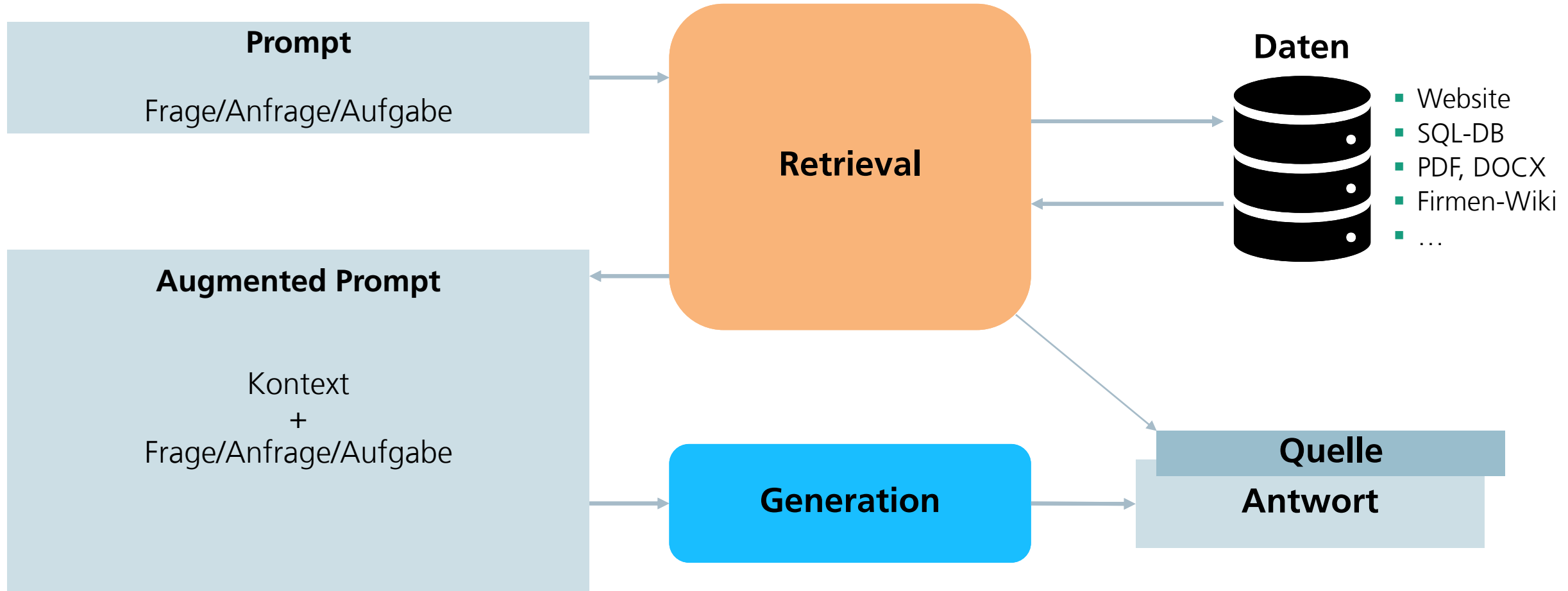
Motivation

Wie stellt man dem LLM eine Informationsquelle bereit?



Motivation

Wie stellt man dem LLM eine Informationsquelle bereit?



RAG

„Manuelles RAG“ (Copy & Paste)

Kontext:

Dependable AI / Verlässliche KI

Dependability oder Verlässlichkeit eines Systems beschreibt die Fähigkeit, nicht akzeptable Fehler zu vermeiden (Jean-Claude Laprie). Dependability beinhaltet dabei z.B. die Verfügbarkeit oder funktionale Sicherheit (Safety), um katastrophale Folgen für die Benutzer und die Umgebung als Folge von Fehlern). Aber auch die Integrität und Zugriffssicherheit (Security) eines Systems im Kontext von Dependability genannt werden.

Beim Engineering verlässlicher KI-Systeme geht es entsprechend darum, mittels Prinzipien aus dem Bereich der Verifikation/Validation und dem Betrieb des KI-Systems die Verlässlichkeit systematisch zu gewährleisten. Normative Vorgaben bzgl. der Verlässlichkeit von Beginn an zu berücksichtigen.

Frage: Was ist Dependable AI? Antworte in Form einer Stichpunktliste.

Antworte präzise auf Grundlage des Kontextes. Benutze ausschließlich Informationen aus dem gegebenen Kontext. Falls die Antwort nicht im Kontext enthalten ist, antworte "ANSWER NOT IN CONTEXT".

Antwort:



Run 1: Nous Hermes 2 - Mixtral 8x7B - DPO (float16, Backend: vLLM, License: Apache 2.0)

1. Dependability oder Verlässlichkeit eines Systems: Fähigkeit, nicht akzeptable Fehler bei der Nutzung zu vermeiden (Jean-Claude Laprie).
2. Dependability beinhaltet: a. Verfügbarkeit oder Zuverlässigkeit eines Dienstes b. Funktionale Sicherheit für Benutzer und Umgebung c. Integrität und Zugriffssicherheit (Security) eines Systems d.
3. Engineering verlässlicher KI-Systeme: a. Anwendung von Prinzipien aus dem Systems- und Betriebsbereich bei Konstruktion, Verifikation/Validation und Betrieb des KI-Systems c. Berücksichtigung der Verlässlichkeit von Beginn an

<https://www.iese.fraunhofer.de/de/leistungen/dependable-ai.html>

RAG

Existierende Komplettlösungen: Beispiel AnythingLLM (Open Source)

The screenshot displays the AnythingLLM web interface. On the left sidebar, there is a 'New Workspace' button, an 'ollama' provider selection, a 'default' workspace, and a 'New Thread' button. The main chat area shows a user query: 'Please give me a quick summary of the Saturn project'. The AI response provides a detailed summary of the Saturn project, mentioning its goal to support rare disease diagnosis and treatment using AI and clinical data, and its use of Eclipse BaSyx middleware for sustainable production. Below the response are icons for copy, edit, refresh, like, and comment, along with a 'Hide Citations' dropdown. Two citation links are visible: 'ar-2022-eng-fraun...ese.docx' and 'ar_2021-dt-fraunh...ese.docx'. At the bottom, there is a 'Send a message' input field with a right-pointing arrow and icons for slash, mention, and text formatting.

RAG

Eigenständige RAG-Umsetzung

Enter text:

Bitte fasse das Projekt SATURN zusammen. Erstelle mir hierfür 5 Stichpunkte.

Submit

Clear all

(Soft) word limit

150

1

2000

Number of search results (candidates)

8

10



Prompt: Bitte fasse das Projekt SATURN zusammen. Erstelle mir hierfür 5 Stichpunkte.

Extracted question: Was sind die fünf wichtigsten Punkte des Projekts SATURN?

Generated answer:

1. SATURN ist ein Projekt, das vom Fraunhofer IESE geleitet wird und die Diagnose seltener Erkrankungen unterstützt.
2. Künstliche Intelligenz (KI) wird verwendet, um bei geringen Datenmengen nachvollziehbare und transparente Verdachtsdiagnosen zu erstellen.
3. Das Projekt wird bis 2024 vom Bundesministerium für Gesundheit (BMG) gefördert.
4. SATURN arbeitet mit dem Universitätsklinikum Frankfurt, der Goethe-Universität Frankfurt und der Technischen Universität Dresden zusammen.
5. Das Projekt zielt darauf ab, die Diagnose seltener Erkrankungen aus strukturierten und unstrukturierten Daten vorherzusagen und eine passendere Behandlung der Patienten zu ermöglichen.

Sources:

Score: 60.0%, File name: ar_2022-dt-fraunhofer_iese.pdf, Page: 34, Used for generation: True

Score: 51.2%, File name: ar_2019dt-fraunhofer_iese.pdf, Page: 62, Used for generation: True

Score: 26.2%, File name: ar_2017-dt-fraunhofer_iese.pdf, Page: 47, Used for generation: False

Score: 26.2%, File name: ar_2021-dt-fraunhofer_iese.pdf, Page: 46, Used for generation: True

Score: 17.9%, File name: ar_2021-dt-fraunhofer_iese.pdf, Page: 11, Used for generation: False

Score: 17.9%, File name: ar_2018-dt-fraunhofer_iese.pdf, Page: 46, Used for generation: True

Score: 17.5%, File name: ar_2022-dt-fraunhofer_iese.pdf, Page: 35, Used for generation: True

Score: 13.7%, File name: ar_2013-dt-fraunhofer_iese.pdf, Page: 105, Used for generation: False



RAG: Evaluation

©iStock.com | Greyfebruary

Halluzinationen

RAG adressiert faktenbezogene Halluzination in generierten Ausgaben

Halluzinationen: Generierte Inhalte, die unsinnig bzw. *fehlerhaft* sind, oder *Inkonsistenzen zu den (Nutzer-) Eingaben* aufweisen

Factuality hallucinations sind Diskrepanzen zwischen generierten Inhalten und *verifizierbaren Fakten*

Faithfulness hallucinations sind Abweichungen der generierten Inhalte von den *Nutzeranweisungen*, den bereit gestellten Kontexten oder die generierten Inhalte enthalten Inkonsistenzen in sich



RAG kann Halluzinationen reduzieren, welche dadurch entstehen, dass

- Informationen während des Modelltrainings *(noch) nicht vorlagen bzw. nicht mehr aktuell* sind (z.B. aktuelle Geschehnisse)
- *Spezialwissen* zur Beantwortung der Anfrage erforderlich ist (z.B. unternehmensinterne Informationen, Domänenwissen)

Effektivität von RAG

RAG behebt faktenbezogene Halluzinationen nicht komplett

Untersuchung zu Halluzinationen in RAG-basierten juristischen Recherchertools (im Vergleich zu GPT-4)¹:

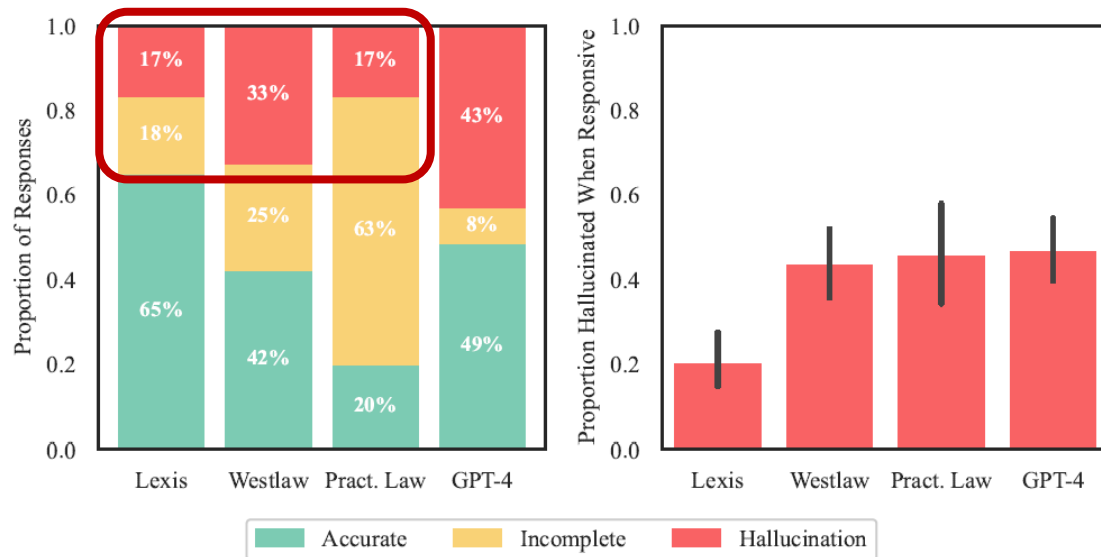


Figure 4: *Left panel:* overall percentages of accurate, incomplete, and hallucinated responses. *Right panel:* the percentage of answers that are hallucinated when a direct response is given. Westlaw AI-AR and Ask Practical Law AI respond to fewer queries than GPT-4, but the responses that they do produce are not significantly more trustworthy. Vertical bars denote 95% confidence intervals.

Untersuchung inwieweit die Bereitstellung der benötigten Informationen dem Modell hilft inkorrekte Antworten zu korrigieren („Tausziehen“ zwischen externem und internem Wissen)²:

- Bevorzugung der extrahierten Informationen ist *invers korreliert mit der Modellkonfidenz* bei gleicher Anfrage ohne Bereitstellung der Referenzdokumente
- Bei stärkerer *Modifikation der Fakten* in den Referenzdokumenten wird eher auf das *interne Wissen* zurückgegriffen

➔ **RAG verhindert Halluzinationen nicht komplett**

➔ **Nutzung der Informationen aus den Referenzdokumenten kann nicht garantiert werden**

¹ V. Magesh, et al., 2024, Hallucination-Free? Assessing the Reliability of Leading AI Legal Research Tools, arXiv:2405.20362 [Ref. Image]

² K. Wu, et al., 2024, ClashEval: Quantifying the tug-of-war between an LLM's internal prior and external evidence, arXiv:2404.10198

RAG Evaluationsmetriken

Komponenten der RAG Pipeline werden separat und zusammen evaluiert

Evaluation des Retrievers

Context precision: Anteil des abgerufenen Kontexts, welcher nützlich zur Beantwortung der Anfrage ist

Context recall: Sind alle benötigten Information zur Beantwortung der Anfrage im Kontext enthalten?

Retrieval

Suche



Evaluation des Generators

Generation

LLM



Faithfulness: Anteil der korrekten Aussagen in der generierten Antwort

Answer relevancy: Beantwortet die generierte Antwort die Anfrage?

End-to-end Evaluation

RAG-System

Answer semantic similarity: Semantische Ähnlichkeit der generierten zur erwarteten Antwort (ground truth)

Answer correctness: Berücksichtigt semantische Ähnlichkeit sowie sachliche Korrektheit (factual accuracy)

RAG Benchmarks

Benchmarkdaten basieren häufig auf Question-Answering Aufgaben

Benchmarks enthalten Testdatensätze um Aspekte der RAG-Pipeline basierend auf Metriken zu prüfen

- Häufig bei RAG: **Question-Answering**
- Verwendung verschiedener Metriken (häufig Accuracy, teils basierend auf datengetriebenen Modellen)
- Nutzung bei der Auswahl von LLMs bzw. RAG-Lösungen oder zur Evaluation des eigenen RAG-Systems (falls Testdaten verfügbar sowie automatisierte Evaluation)

**Comprehensive RAG Benchmark
(CRAG)**

**RAG Benchmark
(RGB)**

RAGTruth Benchmark

Comprehensive RAG Benchmark (CRAG)

CRAG umfasst unterschiedliche Fragetypen und prüft Einzelkomponenten

Table 2: Definition of CRAG question types.

Question type	Definition
Simple	Questions asking for simple facts that are unlikely to change overtime, such as the birth date of a person and the authors of a book.
Simple w. Condition	Questions asking for simple facts with some given conditions, such as stock prices on a certain date and a director's recent movies in a certain genre.
Set	Questions that expect a set of entities or objects as the answer (e.g., "what are the continents in the southern hemisphere?").
Comparison	Questions that compare two entities (e.g., "who started performing earlier, Adele or Ed Sheeran?").
Aggregation	Questions that require aggregation of retrieval results to answer (e.g., "how many Oscar awards did Meryl Streep win?").
Multi-hop	Questions that require chaining multiple pieces of information to compose the answer (e.g., "who acted in Ang Lee's latest movie?").
Post-processing heavy	Questions that need reasoning or processing of the retrieved information to obtain the answer (e.g., "how many days did Thurgood Marshall serve as a Supreme Court justice?").
False Premise	Questions that have a false preposition or assumption (e.g., "What's the name of Taylor Swift's rap album before she transitioned to pop?" (Taylor Swift has not yet released any rap album)).

X. Yang, et al., 2024, CRAG -- Comprehensive RAG Benchmark, arXiv:2406.04744 [Ref. Image]
<https://www.aicrowd.com/challenges/meta-comprehensive-rag-benchmark-kdd-cup-2024>

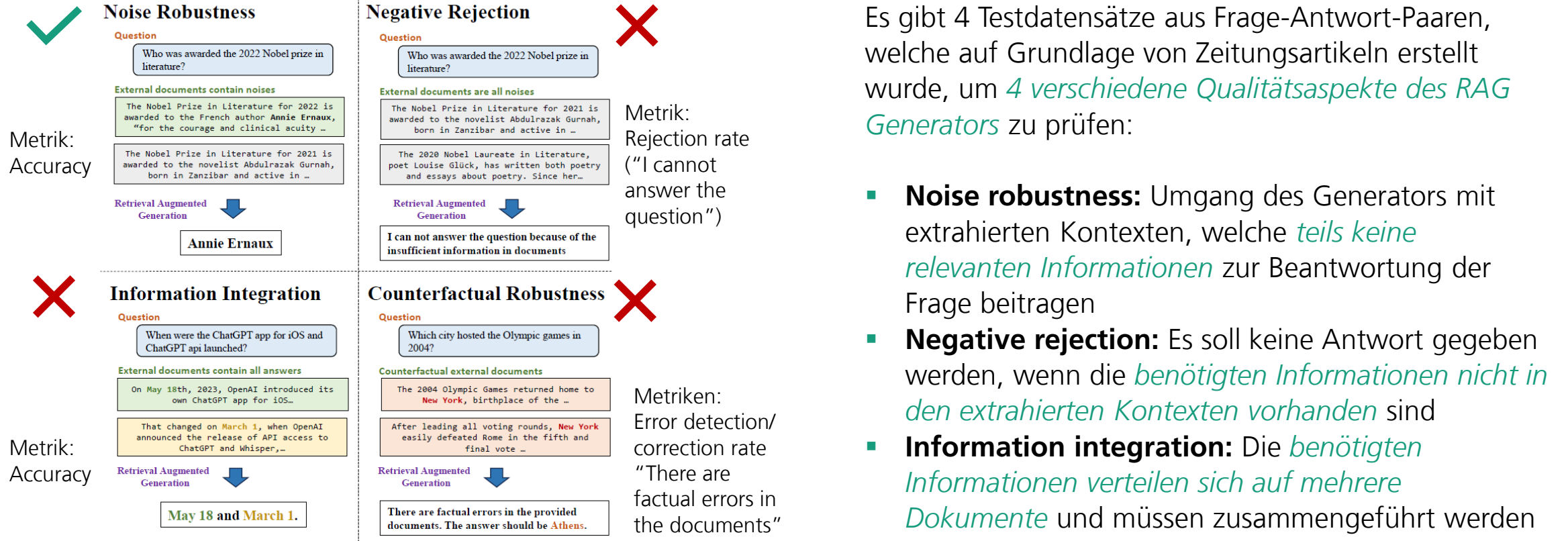
Zur Beantwortung der Fragen werden *unterschiedliche externe Informationen* zur Verfügung gestellt, um verschiedene Aspekte der RAG Pipeline zu prüfen:

- **Retrieval summarization:** 5 Webseiten pro Frage zur *Evaluation des Generators*
- **Retrieval augmentation:** Zusätzlicher Zugang zu Wissensgraphen zur *Evaluation des Retrievers*
- **End-to-end RAG:** 50 Webseiten und Zugang zu Wissensgraphen zur *Evaluation des Rankings einer größeren Menge an Informationen* aus dem Retrieval

Human and LLM-based evaluation of answer (accurate, incorrect, missing)

RAG-Benchmark (RGB)

RGB evaluiert den Generator hinsichtlich Nutzung der Kontexte



Es gibt 4 Testdatensätze aus Frage-Antwort-Paaren, welche auf Grundlage von Zeitungsartikeln erstellt wurde, um *4 verschiedene Qualitätsaspekte des RAG Generators* zu prüfen:

- **Noise robustness:** Umgang des Generators mit extrahierten Kontexten, welche *teils keine relevanten Informationen* zur Beantwortung der Frage beitragen
- **Negative rejection:** Es soll keine Antwort gegeben werden, wenn die *benötigten Informationen nicht in den extrahierten Kontexten vorhanden* sind
- **Information integration:** Die *benötigten Informationen verteilen sich auf mehrere Dokumente* und müssen zusammengeführt werden
- **Counterfactual robustness:** Umgang mit *Fehlern in den Dokumenten*

Figure 1: Illustration of 4 kinds of abilities required for retrieval-augmented generation of LLMs.

J. Chen, et al., 2023, Benchmarking Large Language Models in Retrieval-Augmented Generation, arXiv:2309.01431 [Ref. Image]

<https://github.com/chen700564/RGB>

RAGTruth Benchmark

RAGTruth evaluiert 6 LLMs hinsichtlich Halluzinationen mittels Detektors

QUESTION: how to prepare to get an ultrasound?

REFERENCE: Passage 1: 1 If you are just interested in checking the gender, you can go as early as 17 weeks. 2 Any earlier, and the results may not be accurate, leading to an extra coat of paint for your bedroom. 3 To get good pictures, 3D ultrasounds are best performed between 20 and 32 weeks, and the best pictures are between 24 and 30 weeks.

...

RESPONSE: Based on the given passages, here's how to prepare for an ultrasound:

* For a gender reveal ultrasound, wait until at least 17 weeks of pregnancy to ensure accurate results.

* For 3D ultrasounds, schedule the appointment between 20 and 32 weeks of pregnancy for the best pictures ...

ANNOTATION

Span: between 20 and 32 weeks of pregnancy for the best pictures

Type: Evident Conflict

Reason: Original: "the best pictures are between 24 and 30 weeks", Generative: "between 20 and 32 weeks of pregnancy for the best pictures"

Table 1: An example of RAGTruth data from the question answering task. It contains context, response generated by LLM with and span-level annotation.

RAGTruth fokussiert auf *Halluzinationsdetektion* und enthält Antworten zu

- *3 Aufgabentypen* (Question answering, Data-to-text writing, News summarization), welche
- von jeweils *6 LLMs* generiert wurden,
- *annotiert durch Menschen* auf der Ebene von *Textspannen* innerhalb der generierten Antworten,
- deren Inhalte den bereitgestellten Informationen *widersprechen oder frei erfunden* sind.

Neuer **Detektor für Halluzinationen** durch Finetuning von einem LLM (Llama-2-13B) auf dem Trainingsplit der RAGTruth-Daten

Hallucination suppression: Nutzung des Detektors um aus mehreren generierten Antworten zu wählen

C. Niu, et al., 2023, RAGTruth: A Hallucination Corpus for Developing Trustworthy Retrieval-Augmented Language Models, arXiv:2401.00396 [Ref. Image]

<https://github.com/ParticleMedia/RAGTruth>

Evaluationsframeworks

RAGAs unterstützt die Evaluation von RAG-Systemen auf eigenen Testdaten

RAGAs (Retrieval-Augmented Generation Assessment):

Framework zur Evaluation des Retrievers, des Generators, sowie End-to-end

Erwartet die folgenden Eingaben:

- **question:** Die Nutzeranfrage and das RAG-System
- **answer:** Die generierte Antwort des RAG-Systems
- **contexts:** Die extrahierten Texte aus den Referenzdokumenten
- **ground_truth:** Annotierte erwartete Antwort zur question (nur für bestimmte Metriken benötigt)

Auf eigenen Testdaten, welche z.B. mithilfe eines LLMs automatisch generiert wurden

Beispiel

question = An welchen Normen orientiert sich das Fraunhofer IESE bei der Durchführung von IT-Sicherheitsüberprüfungen?

answer = Das Fraunhofer IESE orientiert sich bei der Durchführung von IT-Sicherheitsüberprüfungen an der ISO27000-Reihe und weiteren einschlägigen Best Practices.

contexts = ... die IT-Sicherheitsüberprüfungen inner -nhalb der Fraunhofer-Gesellschaft durch [...] Seit einigen Jahren bietet das IESE diese Dienst -nleistung auch Unternehmen aus der Wirtschaft an – stets orientiert an der ISO27000-Reihe und \nweiteren einschlägigen Best Practices. ...

ground_truth = ISO27000-Reihe und weiteren einschlägigen Best Practices

Evaluation results:

- Context precision: 1.0, Context recall: 1.0
- Faithfulness: 1.0, Answer relevancy: 0.73
- Answer similarity: 0.7, Answer correctness: 0.92

Evaluationsframeworks

RAGAs nutzt LLMs zur Bestimmung der Metriken, was zu Ungenauigkeiten der Evaluation führen kann

Beispiel

question = Wie viele Auszubildende und duale Studierende sind im Überblick aufgeführt?

answer = 10 Auszubildende und duale Studierende sind im Überblick aufgeführt.

contexts = [IESE im Überblick, 71 Personal (IN KÖPFEN)\n201\nSTAMMPERSONALIESE im Überblick\n10\nAUSZUBILDENDE &\nDUAL STUDIERENDE\n61\n ...

ground_truth = 10

Evaluation results:

- Context precision: 0.5, Context recall: 1.0
- Faithfulness: 0.0**, Answer relevancy: 0.79
- Answer similarity: 0.48, Answer correctness: 0.87

Wie wird Faithfulness in RAGAs berechnet?

Generierte Antwort

llm

Given a question and answer, create one or more statements from each sentence in the given answer.

question: [question]

answer: [answer]

Extrahierte Aussagen aus der generierten Antwort

llm

Consider the given context and following statements, then determine whether they are supported by the information present in the context. Provide a brief explanation for each statement before arriving at the verdict (Yes/No). Provide a final verdict for each statement in order at the end in the given format. Do not deviate from the specified format.

statement: [statement 1]

...

statement: [statement n]

$$\text{Faithfulness} = \frac{\# \text{ im Kontext begründete Aussagen}}{\# \text{ Aussagen}}$$

Faithfulness: Anteil der korrekten Aussagen in der generierten Antwort
→ Sollte höher oder sogar 1.0 sein!!

RAG Evaluation

RAG Evaluation ist wichtig, bringt aber keine vollständigen Garantien

Empfehlung die *Komponenten sowohl einzeln zu evaluieren als auch ihr Zusammenspiel*

RAG-Pipelines können einen *Halluzinations-Detektor* nutzen

- zur Evaluation der generellen Halluzinationsanfälligkeit des RAG-Systems
- als nachgelagerte Prüfung für Einzelausgaben
- zur Selektion aus mehreren generierten Antwortoptionen (hallucination suppression)



Die Evaluationsmetriken nutzen selbst häufig LLMs oder Embeddingmodelle zur Berechnung.

Öffentliche Benchmarkdaten können Teil der Trainingsdaten sein. Daher zusätzlich auf eigenen Testdaten evaluieren.

Automatisch generierte Testdaten können fehlerhaft sein.

→ Gute Evaluationsergebnisse sollten nur als Indikator und nicht als Garantie für eine gute Performance des RAG-Systems gesehen werden!



RAG: Umsetzung

©iStock.com | Greyfebruary

Vom Ausprobieren zur eigenen Produktivlösung

- Keine Black-Box
- Einfachste Möglichkeit, Wissensfragen zu bekannten Quellen beantworten zu lassen

- Black-Box
- Automatisches Retrieval
- Sehr einfache Möglichkeit, RAG für eigene Dokumente zu nutzen
- Begrenzte Veränderungsmöglichkeiten

- Keine Black-Box
- On-premises-Betrieb möglich
- Optimale Anpassung an Use Case
 - Datenquellen
 - Sprache
 - Algorithmische Variationen

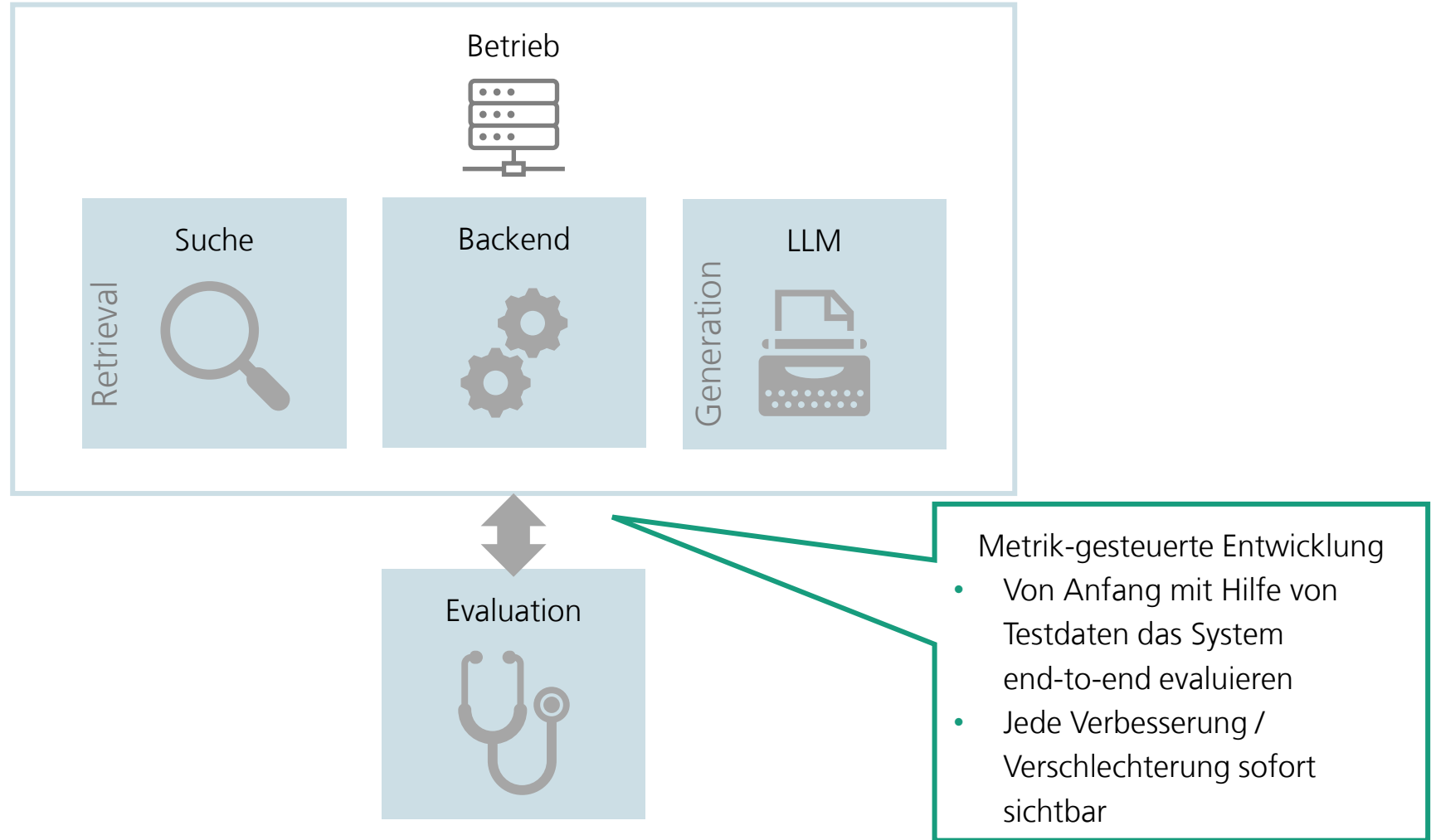
„Manuelles RAG“
(Copy & Paste)

AnythingLLM, ...

Eigenständige RAG-
Umsetzung

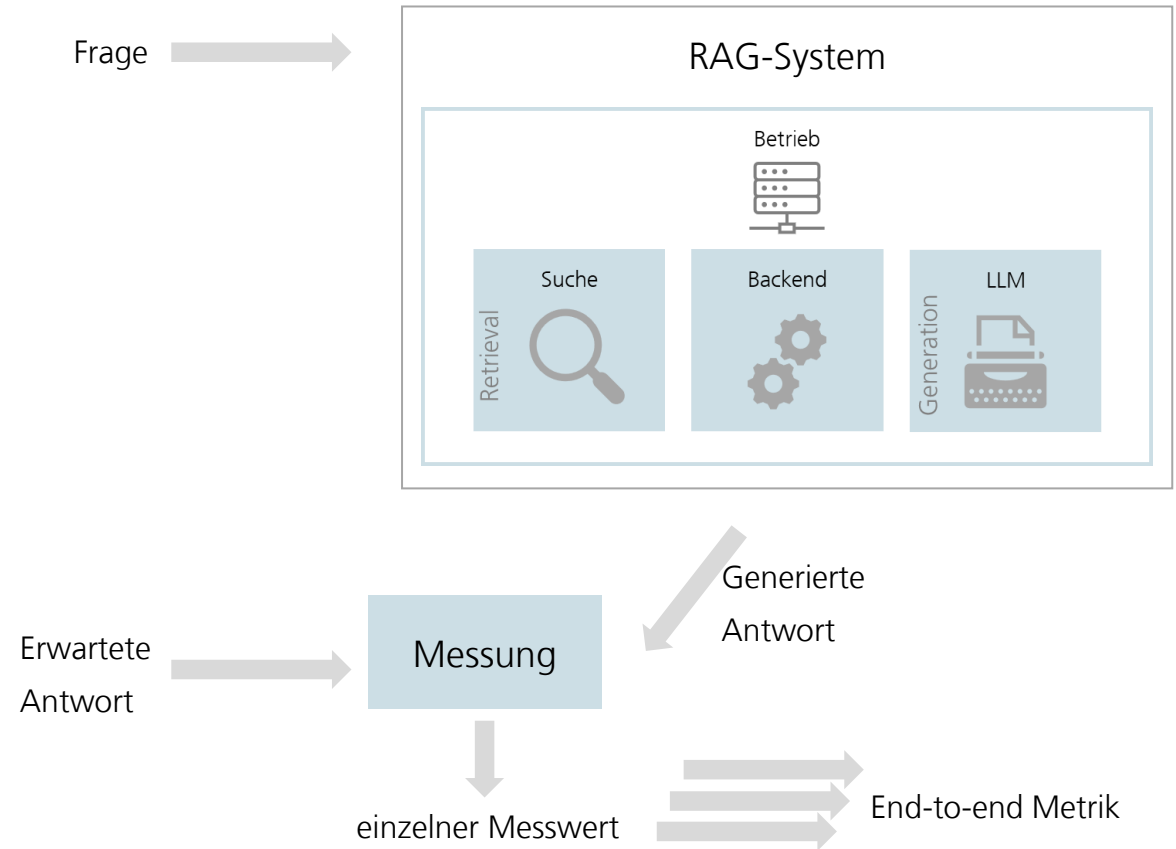
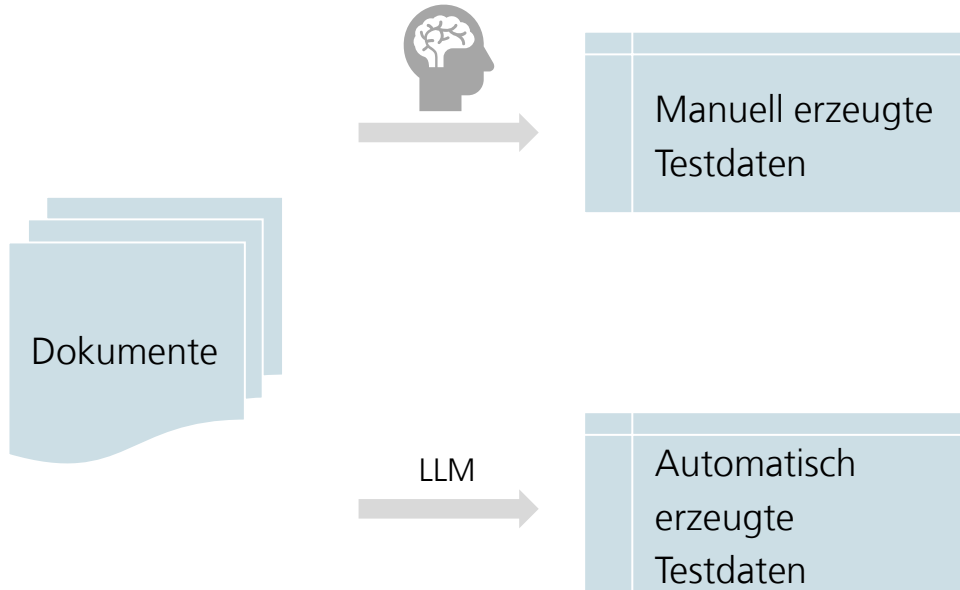
RAG

Komponenten



Evaluation

Metrik-gesteuerte Entwicklung

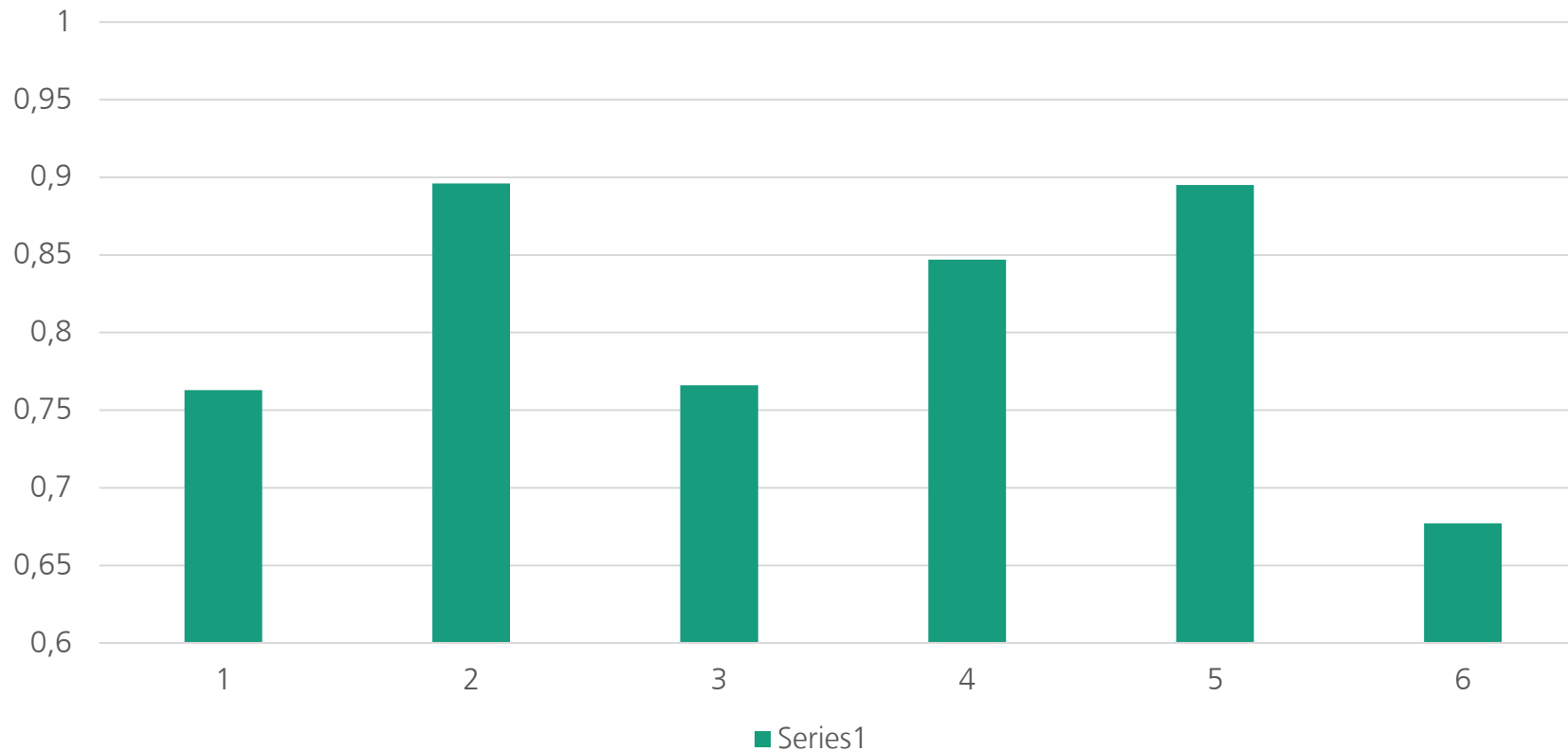


Evaluation

Metrik-gesteuerte Entwicklung

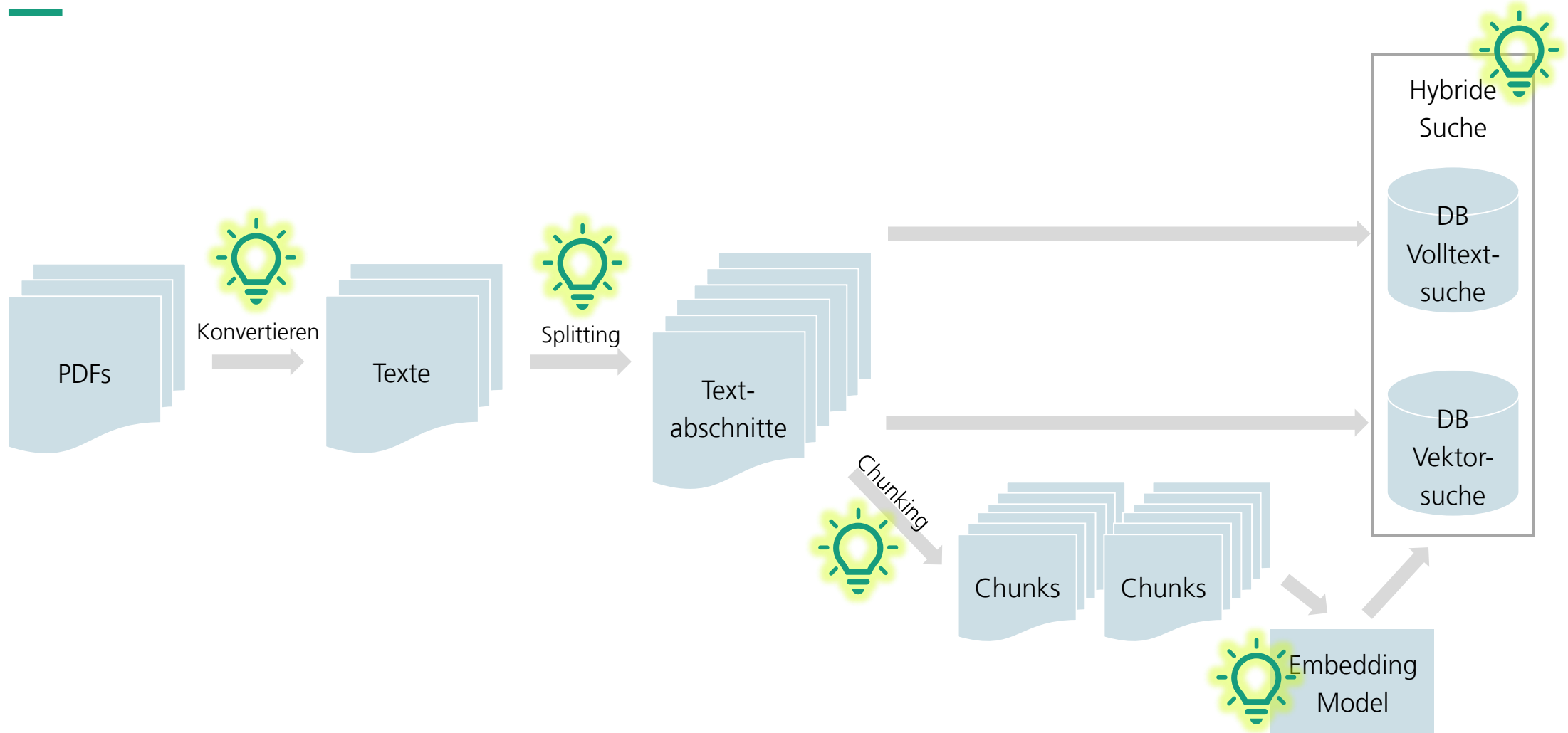


Jahresbericht 2022 (PDF), n=159, k=6



Suche

Eigene PDFs: Import

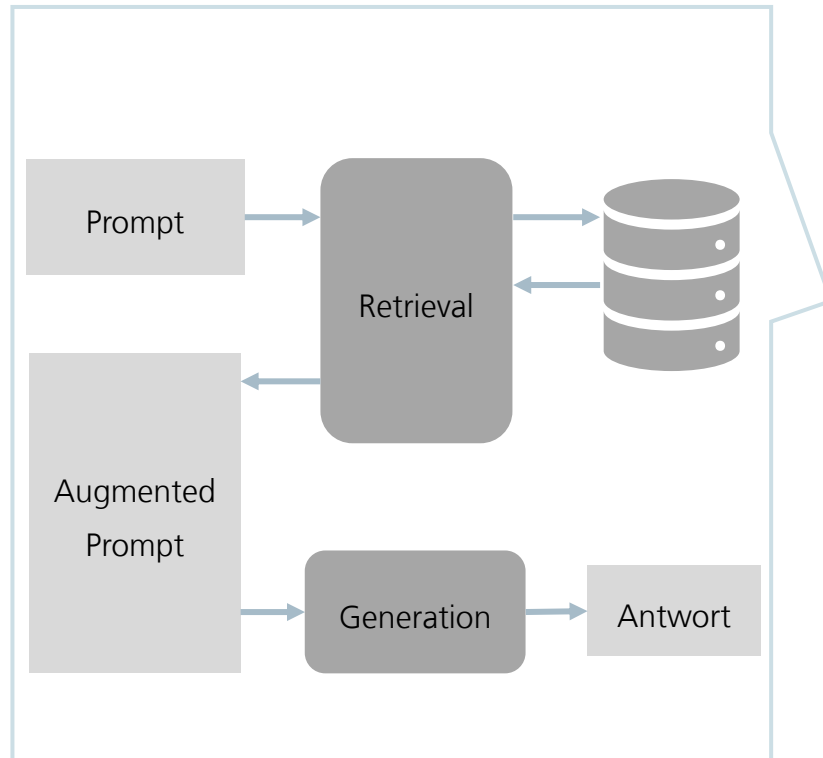


Suche

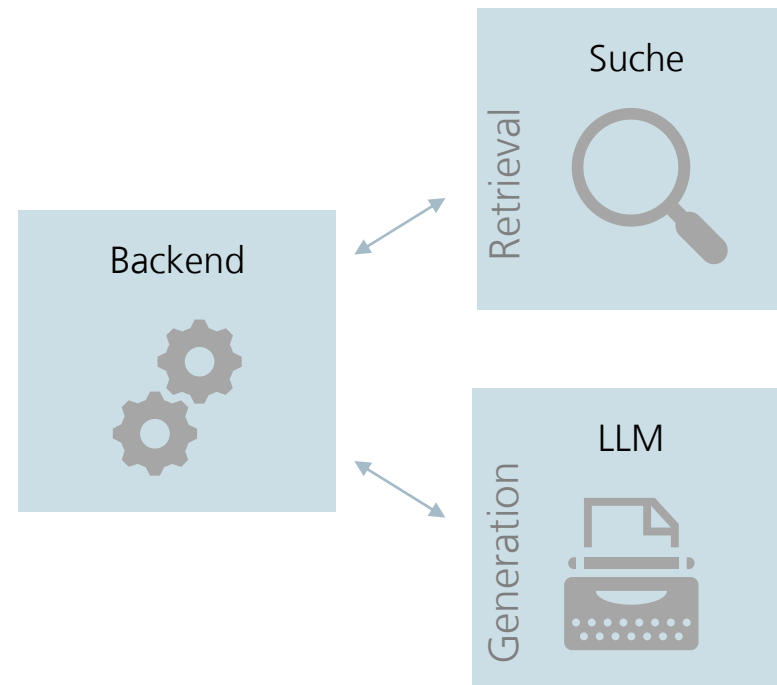
Eigene PDFs: Retrieval



Algorithmus



Komponenten anbinden





Backend

Libraries zum Entwickeln von LLM-Applikationen

Pro

- **Eigener Code wird lesbarer / übersichtlicher**
- **Einfache Anbindung**
- **Schneller Austausch von RAG-Komponenten (LLM, Datenbank)**
- **Einarbeitung hilfreich auch für andere LLM-basierte Projekte**

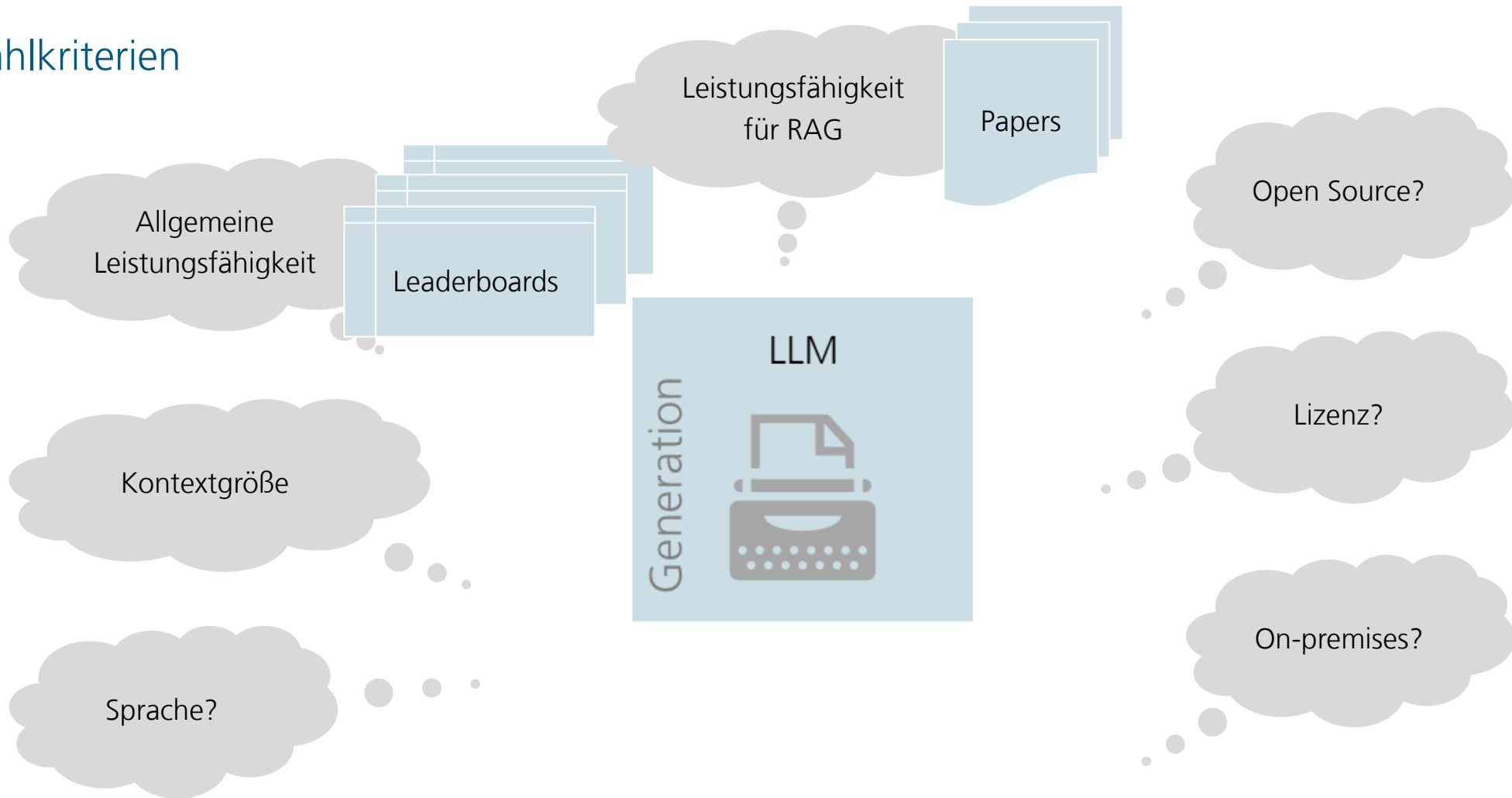
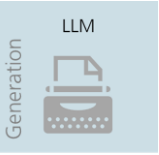
Contra

- **Z.T. viel „Magic“ / Black Box**
- **Nicht alle Stellschrauben (einfach) zugänglich**
- **Große Libraries, schwierig zu überblicken**
- **Einarbeitungszeit nicht unterschätzen**
- **Für eigene Algorithmus-Variationen muss Code „aufgebrochen“ werden**

<https://www.langchain.com/>
<https://www.llamaindex.ai/>

LLM

Auswahlkriterien



LLM

Sind mehrere Modelle erforderlich?

Typ 1: Beantworten / Generieren

- **Kontextgröße sehr wichtig**
- **„Gesprächigkeit“ in Ordnung**
- **Generieren in Zielsprache wichtig**
- **Vorhersagequalität wichtiger als Geschwindigkeit**
 - Reasoning, Zusammenhänge verstehen, ...

Typ 2: Schritte im Algorithmus

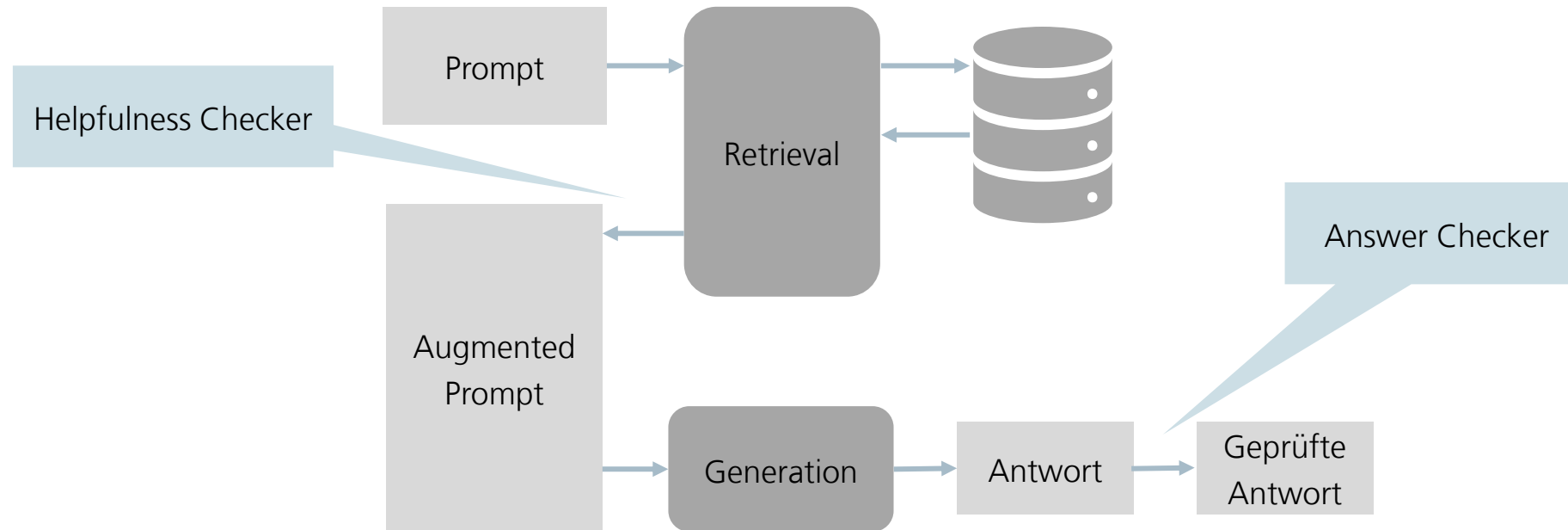
(Query ableiten, Kontexte bewerten, Routing, ...)

- **Sehr hohe Geschwindigkeit erwünscht**
- **Sehr gutes Instruction-Tuning nötig**
- **Kontextgröße von untergeordneter Bedeutung**
- **„Gesprächigkeit“ unerwünscht**
- **Generieren in Zielsprache irrelevant**

→ Eine eigene, Use-Case-spezifische Evaluation hilft bei der Modellauswahl!

LLM

Auswahlhilfe: Teilschritte evaluieren









Auswahlhilfe: Teilschritte evaluieren („Helpfulness Checker“)

 Datasets:  deutsche-telekom / Ger-RAG-eval  like 37  Dataset card  Viewer  Files and versions  Com

 Subset (4)
 task2 · 1k rows

 Split (1)
 test · 1k rows

Search this dataset

question string · lengths	choice_a string · lengths	choice_b string · lengths	choice_c string · lengths	choice_d string · lengths	target string · lengths
 83→97 8.1%	 828→970 14.6%	 691→831 13.4%	 683→824 13.4%	 540→679 12.1%	 A
Wann wurde der Verein "aDEvantgarde - Projekt...	Der Sender Waldshut ist ein Füllsender des...	Die "Scheune aus Gottstreu" ist ein ehemals unter Denkmalschut...	Kliemannsland ist ein Kreativprojekt des YouTubers...	A•DEvantgarde ist ein Zusammenschluss von Komponisten i...	D
Was versteht man unter Mehrfachvererbung bei...	Bei der Objektorientierten Programmierung handelt es...	Not Quite C (kurz NQC) ist eine Programmiersprache mit...	Als Audiostecker werden Steckverbinder für Audiokabe...	Ein Label, auch Static, Bezeichnungsfeld oder...	A
Was ist das offizielle Name des...	Das Trappistinnenkloster Boa Vista, offizieller...	Die Küste de "A Costa dos			
Welche Wirkung hat Mabinlin 4 auf die Zung...	Mabinlin 4 ist ein Protein aus der chinesischen...	Der englisch (MPMD, dt. e			
Wann fand das Tennisturnier Barcelona...	Mitar Novaković (kyrillisch: Митар...	Die Futsal-Ozeanienmeis			
Was war der "SIVOM de Rohrbach-lès-Bitche"?	Loktionowo () ist ein Dorf "(derewnja)" in der Oblas...	Der Kirchen ein Kirchen			
Unter welchem Pseudonym war Mac Miller als...	Mark Allan Hoppus (* 15. März 1972 in Ridgecrest,...	Jehor Wadymc Markelow (;			

```

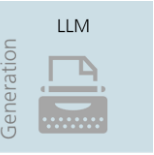
1 system_prompt_template = \
2 """You are a helpful AI assistant. You must answer very precisely.
3
4 Context:
5 {context}
6
7 Question:
8 {question}"""
9
10 instruction = """Is the given Context helpful for answering the given Question, YES or NO? You must only give your verdict as YES or NO. You must not add comments, explanations or notes."""

```

<https://huggingface.co/datasets/deutsche-telekom/Ger-RAG-eval>

LLM

Auswahlhilfe: Teilschritte evaluieren („Answer Checker“)



Datasets: deutsche-telekom / Ger-RAG-eval like 37

Dataset card **Viewer** Files Community **2**

Subset (4)
task3 · 1k rows

Split (1)
test · 1k rows

Search this dataset

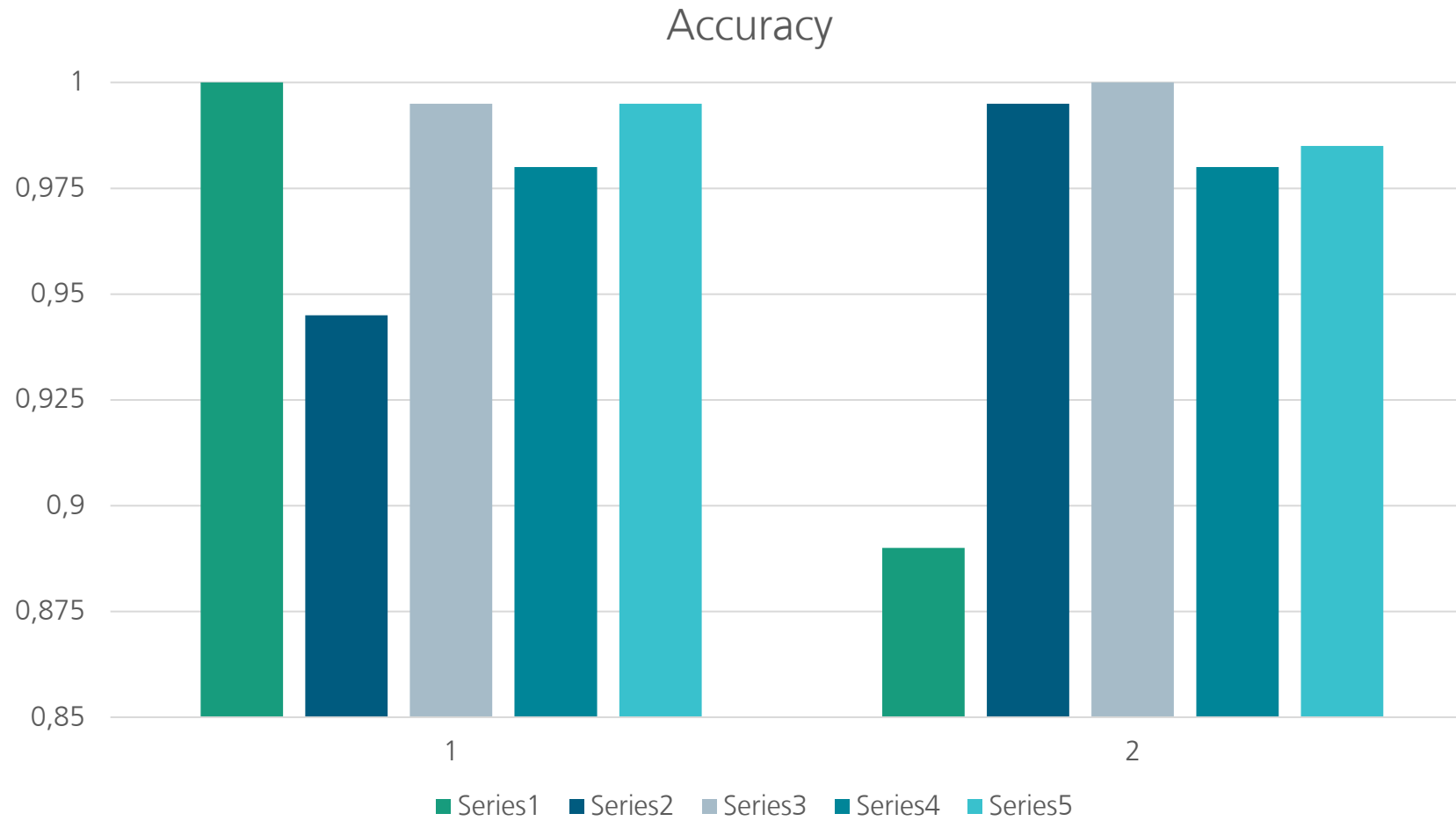
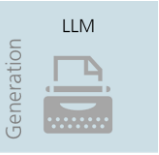
question string · lengths	wiki_id int64	context_uuid string · lengths	answer string · lengths	target string · c
14 161	4.67k 12.5M	36 36	25 921	2 values
Seit welcher Kernel Version ist CUBIC der Standard TCP Überlastalgorithmus in...	12,462,563	a7582384-6873-4238-99ea-1c4b011452c0	Seit Kernel Version 2.6.19 ist CUBIC der Standard TCP Überlastalgorithmus in...	J
Was passiert beim ersten Aufruf der Methode codice_5 in der Klasse codice_1?	9,544,619	ee58...	Beim ersten Aufruf der Methode codice_5	
Welche Themen stehen im Mittelpunkt des Programms von Bergblick?	11,908,915	096e...		
Was sind die bekanntesten Themen, über die McCorduck schrieb?	9,972,432	f0d2...		
Was ist der Crop Trust?	2,295,013	a161...		
Wo kann der Versatz der Legebarren mechanisch ausgeführt werden?	10,631,430	b953...		
Was sind die Hauptziele der PPLU?	5,618,929	10dbi3dc-d7cf-4f8f-be1a-2838d7243e71	für Bürgerrechte, verbesserter...	J

```
1 system_prompt_template = \  
2 """You are a helpful AI assistant. You must answer very precisely."""  
3  
4 instruction = """Question: {question}  
5 Answer:  
6 {answer}  
7  
8 Is the given Answer an answer to the given Question, YES or NO? You must only give your  
9 verdict as YES or NO. You must not add comments, explanations or notes."""
```

<https://huggingface.co/datasets/deutsche-telekom/Ger-RAG-eval>

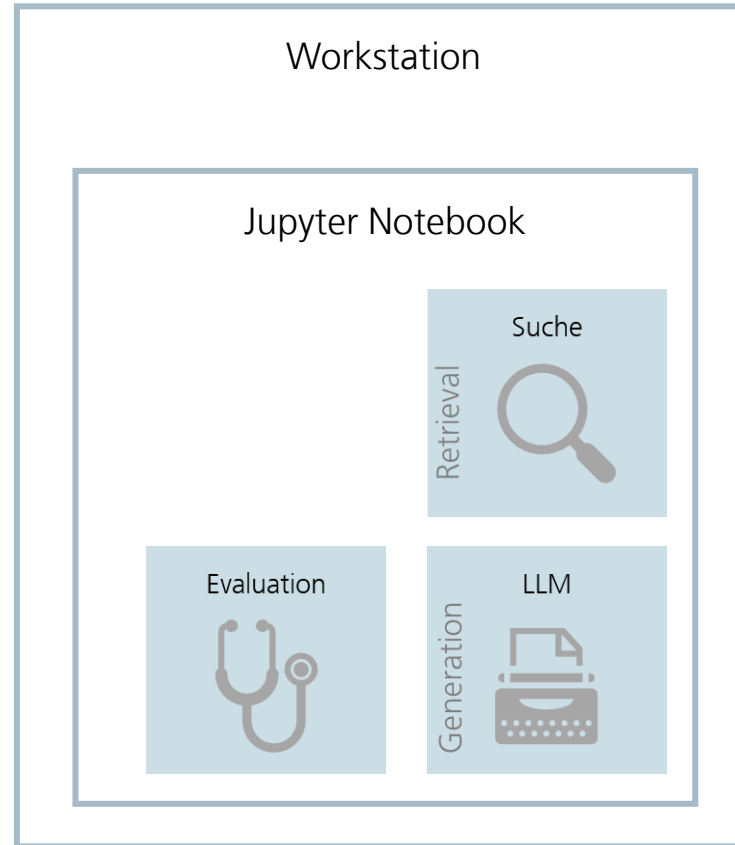
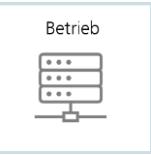
LLM

Auswahlhilfe: Ergebnisse (je n=200)



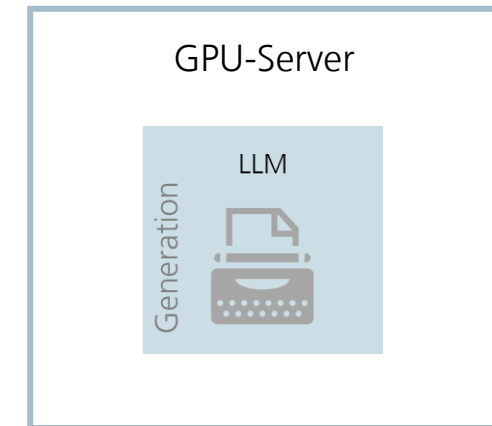
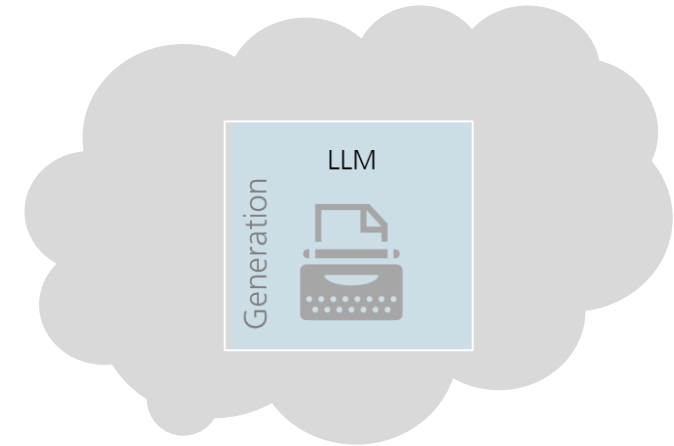
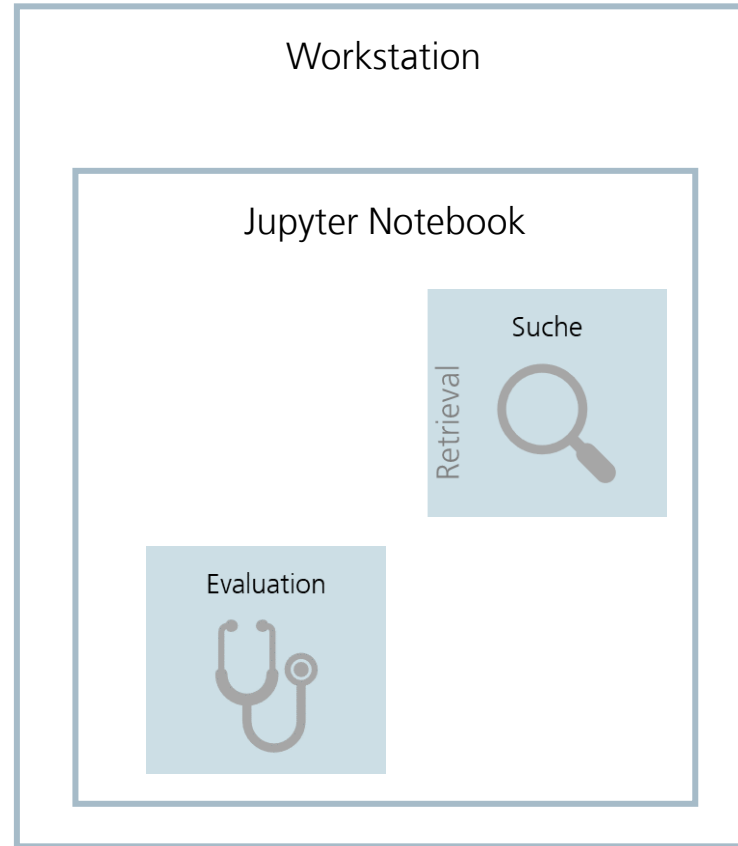
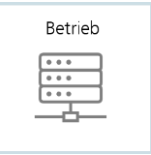
Betrieb

Stufe 1: Prototyping mit Jupyter



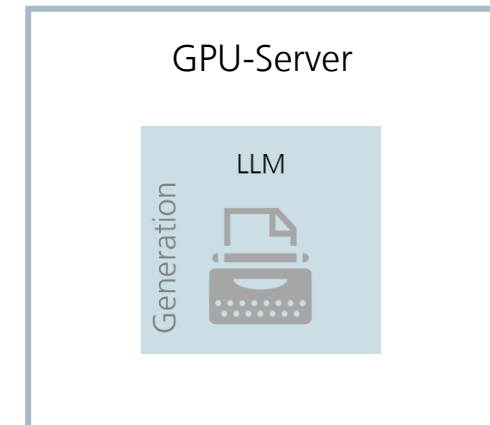
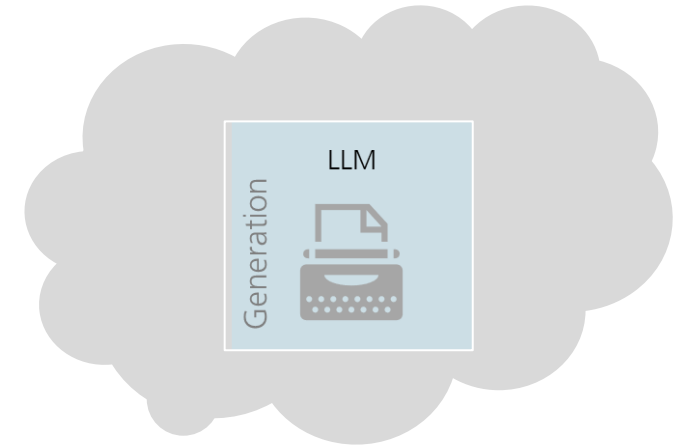
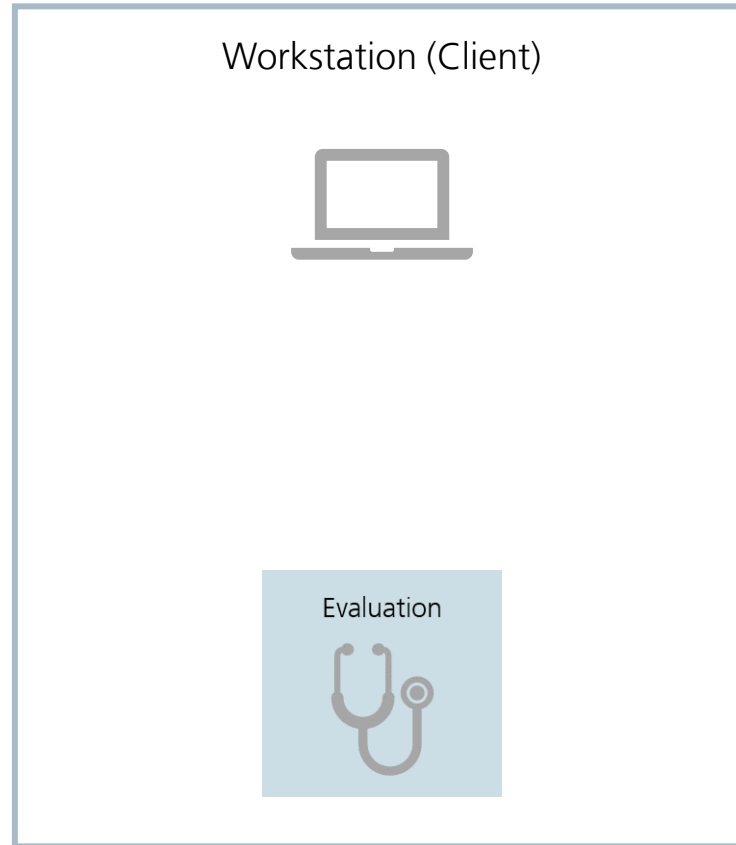
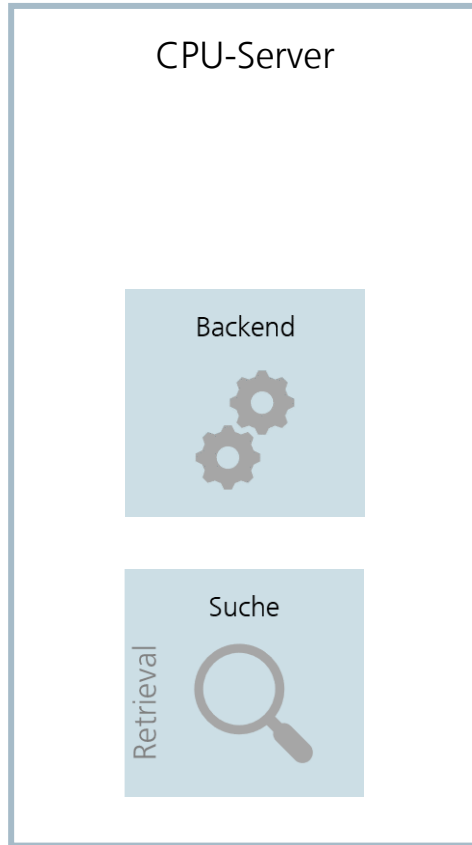
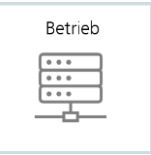
Betrieb

Stufe 2: Ausgelagertes LLM



Betrieb

Stufe 3: Server-basiert



RAG vs. Fine-Tuning

Was ist Fine-Tuning?

- Weiter-Trainieren eines ML-Modells (supervised, unsupervised)
- Im Kontext von LLMs auch: Trainieren eines Zusatzmoduls (z.B. LoRA)

Wann ist Fine-Tuning eines LLMs hilfreich?

- Feinabstimmung für sehr spezielle Aufgaben, z.B. Text-to-SQL
- Feinabstimmung für spezielle Sprachen bzw. Programmiersprachen

Herausforderungen von Fine-Tuning

- Fine-Tuning erfordert Trainingsdaten
- Fine-Tuning erfordert GPU-Rechenzeit
- Fine-Tuning erfordert Experimente, Optimierung (z.B. Aufgabe eines Data Scientists)
- Fine-Tuning für eine Aufgabe kann die Performance in anderen Bereichen verschlechtern

RAG vs. Fine-Tuning

Ovadia et al. (2024): Fine-Tuning or Retrieval? Comparing Knowledge Injection in LLMs

- “Our findings reveal that **while unsupervised fine-tuning offers some improvement, RAG consistently outperforms it**, both for existing knowledge encountered during training and entirely new knowledge.”
- “Moreover, we find **that LLMs struggle to learn new factual information through unsupervised fine-tuning**”

Lakatos et al. (2024): Investigating the performance of Retrieval-Augmented Generation and fine-tuning for the development of AI-driven knowledge-based systems

- “[...] we demonstrate that **RAG-based constructions are more efficient than models produced with FN.**”
- “We point out that **connecting RAG and FN is not trivial**, because connecting FN models with RAG can cause a decrease in performance”
- “Furthermore, we outline a simple RAG-based architecture [...] **This shows the significant advantage of RAG over FN in terms of hallucination**”

Barnett et al. (2024): Fine-Tuning or Fine-Failing? Debunking Performance Myths in Large Language Models

- „Our findings indicate that **fine-tuning resulted in a decline in performance compared to the baseline models**”
- “**The base models of Mixtral and Llama2 without fine-tuning outperformed any of their fine-tuned counterparts across all datasets**, except for the NQ dataset”
- “Similar observations were made with the baseline **GPT-4 model, which performed better than any of the fine-tuned models across all datasets.**”

Unser Fazit: Zunächst alle Optimierungsmöglichkeiten ausreizen, die kein Fine-Tuning erfordern!

RAG vs. Fine-Tuning

Beispielhafte Optimierungsmöglichkeiten, die kein Fine-Tuning benötigen



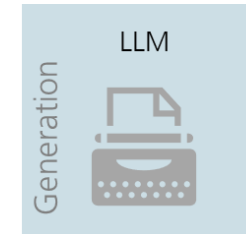
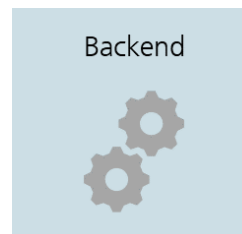
- Art der Datenbank, mehrere Datenbanken?

- Dokumentensammlung

- Suchtechnologie: Vektorsuche, Keywordsuche, Hybrid
- Konvertierung nach Text, Pre-Processing, Splitting
- Reranking, checking

- Algorithmus

- Query ableiten (evtl. mehrere?)
- Mehrere Quellen einbeziehen? Agenten-basiert?
- Mit Teilantworten arbeiten? Diese zusammenfassen?
- Checking?



- LLM

- Prompts variieren
- Prompting-Techniken, z.B. Few-Shot-Prompting
- Wahl des LLM
- Mehrere LLMs? Ensembles?

Zusammenfassung

- RAG erlaubt die Beantwortung von Wissensfragen durch ein LLM bei reduziertem Halluzinations-Risiko
- RAG erlaubt die Verbindung von LLM und eigenen Daten („Chatten mit den eigenen Daten“)
- Für eine qualitätsorientierte Entwicklung eines RAG-Systems ist eine use-case-spezifische Evaluation (von Beginn an) wichtig.
- Fine-Tuning ist (unserer Einschätzung nach) nicht geeignet zum Erlernen von Wissen.
- Bevor Fine-Tuning in Erwägung gezogen wird, sollten (unserer Einschätzung nach) alle anderen Optimierungsmöglichkeiten der RAG-Pipeline ausgereizt werden.

Ist RAG das Richtige für meinen Anwendungsfall?

Dürfen wirklich *keine Ungenauigkeiten* im Ergebnis enthalten sein?

→ Alternativen: z.B. semantische Suche, Knowledge Graph

Passt die *Art der Fragen* zu (einfachem) RAG? („Wie viele ...?“, „Liste alle ...“, Multi-Hop-Fragen)

→ Ergänzung z.B. durch Knowledge Graph, SQL-Datenbank, Routing, Agentic RAG, ...

Suche ich eine on-premises-Lösung, aber steht mir *nur eingeschränkt GPU-Hardware* zur Verfügung?

→ Alternativen: z.B. semantische Suche, Wissensextraktion mit LLM + semantische Suche

Enthalten meine Referenzdokumente *Tabellen, Abbildungen oder Formeln*?

→ Alternativen: Multimodales RAG, Semantische Suche

Leistungen Team Data Science

Wo finden Sie weiterführende Informationen?

IESE-Blogartikel zu LLM / Generative KI und Zuverlässigkeit von KI

- [Was sind Large Language Models? Und was ist bei der Nutzung von KI-Sprachmodellen zu beachten?](#)
- [Open Source Large Language Models selbst betreiben](#)
- [Retrieval Augmented Generation \(RAG\): Chatten mit den eigenen Daten](#)
- [Die Zukunft des Sprachassistenten: Datenhoheit durch Spracherkennung mit eigenem LLM Voice Bot](#)
- [Prompt Engineering: wie man mit großen Sprachmodellen kommuniziert](#)
- [Large Action Models \(LAMs\) nutzen neurosymbolische KI – Die nächste Stufe im Hype rund um Generative AI](#)
- [Wie funktionieren LLMs? Ein Blick ins Innere großer Sprachmodelle](#)

Informationsflyer

- [LLM Innovation Labs \(oder als PDF\)](#)
- [Open-Source Large Language Models \(LLMs\) erfolgreich einsetzen \(oder als PDF\)](#)

Podcasts

- [Morgen Denker Podcast Folge 33: Large Language Models – der Hammer für alle Schrauben?!](#)
- [Morgen Denker Podcast Folge 31: Künstliche Intelligenz – aber sicher und zuverlässig!](#)

Leistungen Team Data Science

Wie können wir Sie rund um das Thema Large Language Model (LLM) unterstützen?

Passend für Ihr Unternehmen und Ihren Use Case, können wir Sie bei folgende Themen unterstützen:

- **Mehrwert:** Beurteilung des Nutzens durch den Einsatz der neuen LLM-Technologien
- **Anwendungsfall:** Identifizieren von sinnvollen Anwendungsfällen für den Einsatz von LLMs in Ihrem Unternehmen
- **Geschäftsmodell:** Auswirkungen auf Ihr Geschäftsmodell und Ihre Kunden
- **Modellauswahl:** Das passende Open-Source LLM für den eigenen Use Case finden
- **Modellbetrieb:** Hilfe beim Betreiben von Open-Source LLMs (On-Premises)
- **Integration:** Einbindung eines Open-Source Large Language Models in eigene Applikationen
- **Unternehmensdaten:** Nutzung eines Open-Source LLMs, um Unternehmensdaten besser zugänglich zu machen, z.B. mithilfe von Retrieval Augmented Generation (RAG)
- **Individualisierung:** Fine-Tuning von Open-Source LLMs für eigene Use Cases bzw. Daten (z. B. mit LoRA)
- **Sprache und Software Engineering:** Einsatz von LLMs für deutschsprachige Anwendungen oder Code/Software Engineering
- **Evaluation:** Evaluation von Open-Source LLMs für eigene Use Cases
- **Qualität:** Beurteilung hinsichtlich Zuverlässigkeit und Vertrauenswürdigkeit
- **Kosten:** Beurteilung der erforderlichen Investitionen und Identifikation von einsparbaren Kosten

Kontakt

Daniel Seifert
Abteilung Data Science
daniel.seifert@iese.fraunhofer.de

Lisa Jöckel
Abteilung Data Science
lisa.joeckel@iese.fraunhofer.de

Thorsten Honroth
Abteilung Data Science
thorsten.honroth@iese.fraunhofer.de

Fraunhofer IESE
Fraunhofer-Platz 1
67663 Kaiserslautern
www.iese.fraunhofer.de



Fraunhofer-Institut für Experimentelles
Software Engineering IESE