Retrieval Augmented Generation (RAG)

Agenda

1. Use Case
2. Requirement
3. Pure-LLM vs Fine-Tuning-LLM vs RAG
4. Architektur
5. Implementation
6. Evaluation
7. Results

Use Cases

- Document Question Answering
- Customize Question-Answering

Anforderungen:

- Abdeckung von alle möglichen User-Fragen über eigene Domain
- Korrekte Antwort auf Fragen
- Unternehmensspezifische Antwort
- Naturliche(menschliche) Antwort
- Lokal laufen lassen
- Möglich kostenlos

Lösungsmöglichkeiten

- 1.) Base LLM
- Keine spezielle Info über Domain
- Halluzination
- Antworten ohne Quelle
- Ihre Wissens meistens nicht aktuell

- 2.) Fine-Tune LLM
- 3.) Retrieval Augmented Generation

RAG vs Base LLM

User-Fragen

Bieten Sie eine telefonische Beratung zur Kfz-Versicherung an?

Kontext

Nein, es wird keine telefonische Beratung für die Kfz-Versicherung angeboten. Die Kommunikation und Angebotserstellung erfolgt ausschließlich online über das Internet.

Pure-LLM
(ohne Kontext)

```
model.generate("Bieten Sie eine telefonische Beratung zur Kfz-Versicherung an?")
✓ 50.5s

'\n\nWir bieten Ihnen gerne eine kostenlose und unverbindliche Beratung zu Ihrer KfZ-Versicherung an.
```

RAG
(mit Kontext)

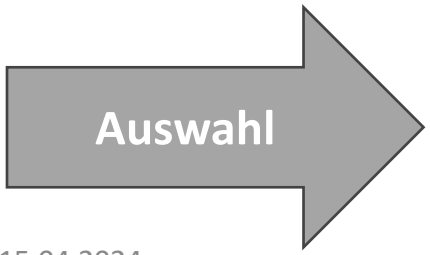
```
res = qa(
    "Bieten Sie eine telefonische Beratung zur Kfz-Versicherung an?"
)
✓ 5m 59.8s

Nein, es wird keine telefonische Beratung für die Kfz-Versicherung angeboten.
```

Fine-Tune LLM :

- ✓ teuer(Zeit, Geld, Daten)
- ✓ erfordert technische und fachliche Kenntnisse
- ✓ hat laut Studien schlechtere Ergebnisse als RAG
- ✓ eine besttimmte Fähigkeit verleihen
- ✓ nicht geeignet für Informationsaktualisierung

RAG vs Fine-tunning

Feature	RAG	Fine-tunning
Knowledge Updates	Direkt Update	Erfordert retraining
Reduzieren Halluzination	weniger anfällig für Halluzinationen	zeigen immer mehr Halluzinationen
Computational Resources	wenig	viel
Interpretability	Tracking möglich	Black box

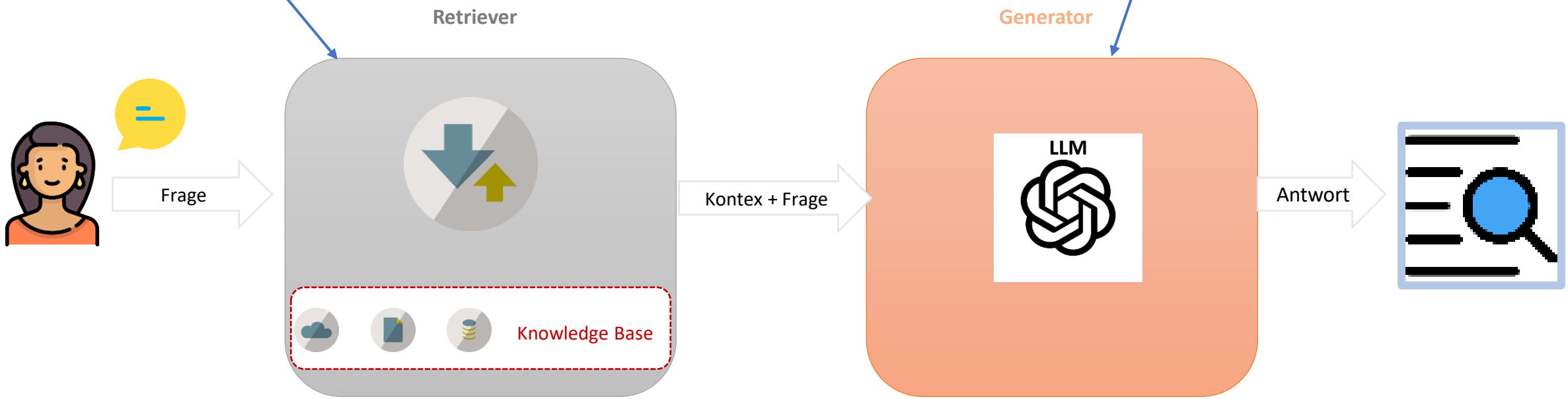


RAG

RAG

- A type of language generation model that combines generator(LLM) and retriever for language generation.

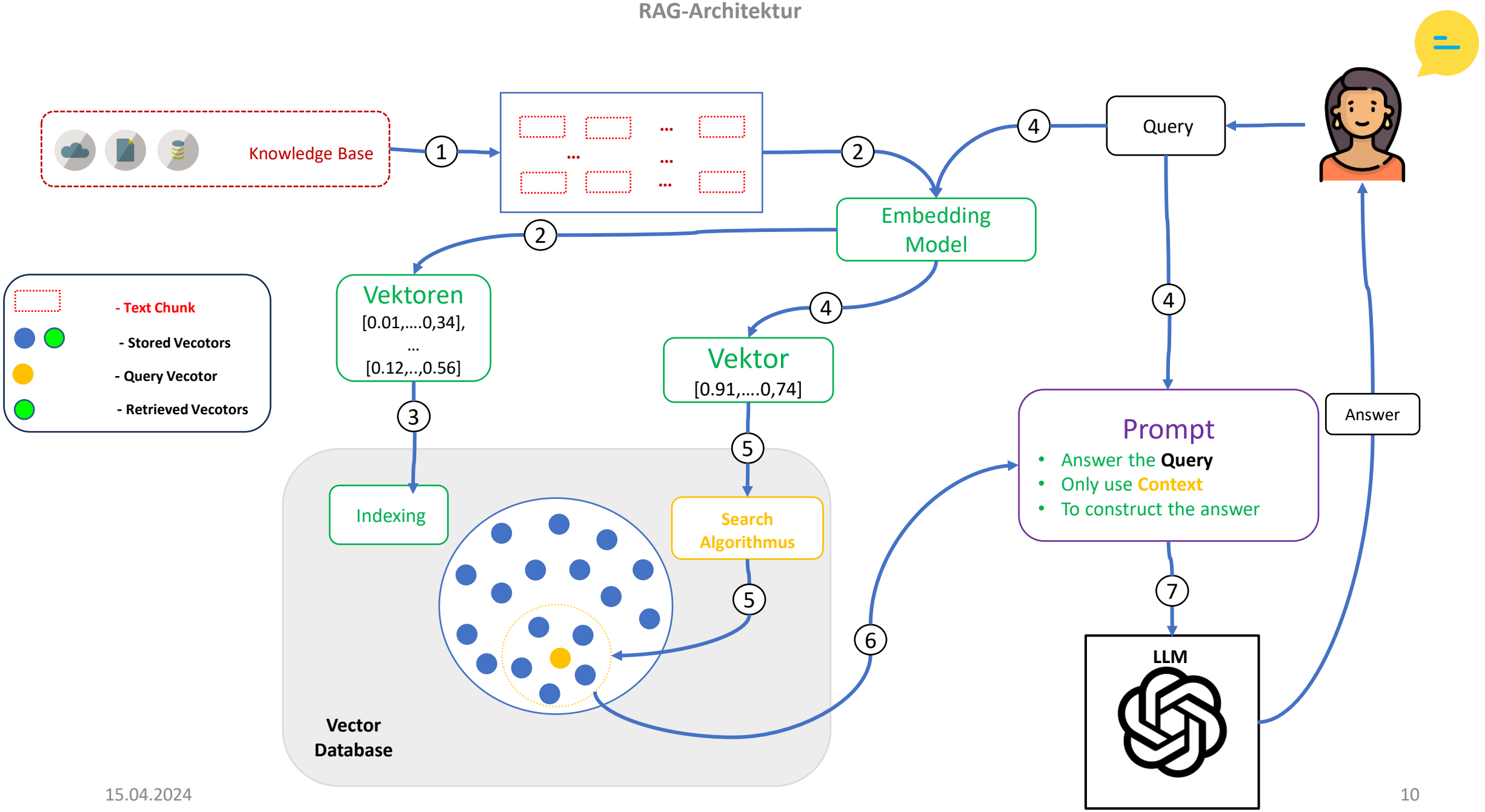
Lewis et al., 2021



RAG-High-Level Architektur, Quelle: Eigene Darstellung

RAG Architektur, Quelle: Eigene Darstellung

RAG-Architektur



Vorteile von RAG

- Reduziert Halluzination
- Verifizierung von Antworten möglich
- Scalable
- Anpassungsmöglichkeiten nach Domaine
- Genauere Antwort als base LLM und Fine-Tune LLM

Implementierung

- 1.) Loading Data
- 2.) Chunking
- 3.) Embeddings and Save in DB
- 4.) Retriever
- 5.) Prompt for LLM

TEST

- Ziel: Die Messung der Genauigkeit der von der Anwendung generierten Antworten
- Datensatz: Question, Ground Throuth, Contex
- Methode: RAGAS-Framework, eventuell Experten-Evaluation

Test-Methoden

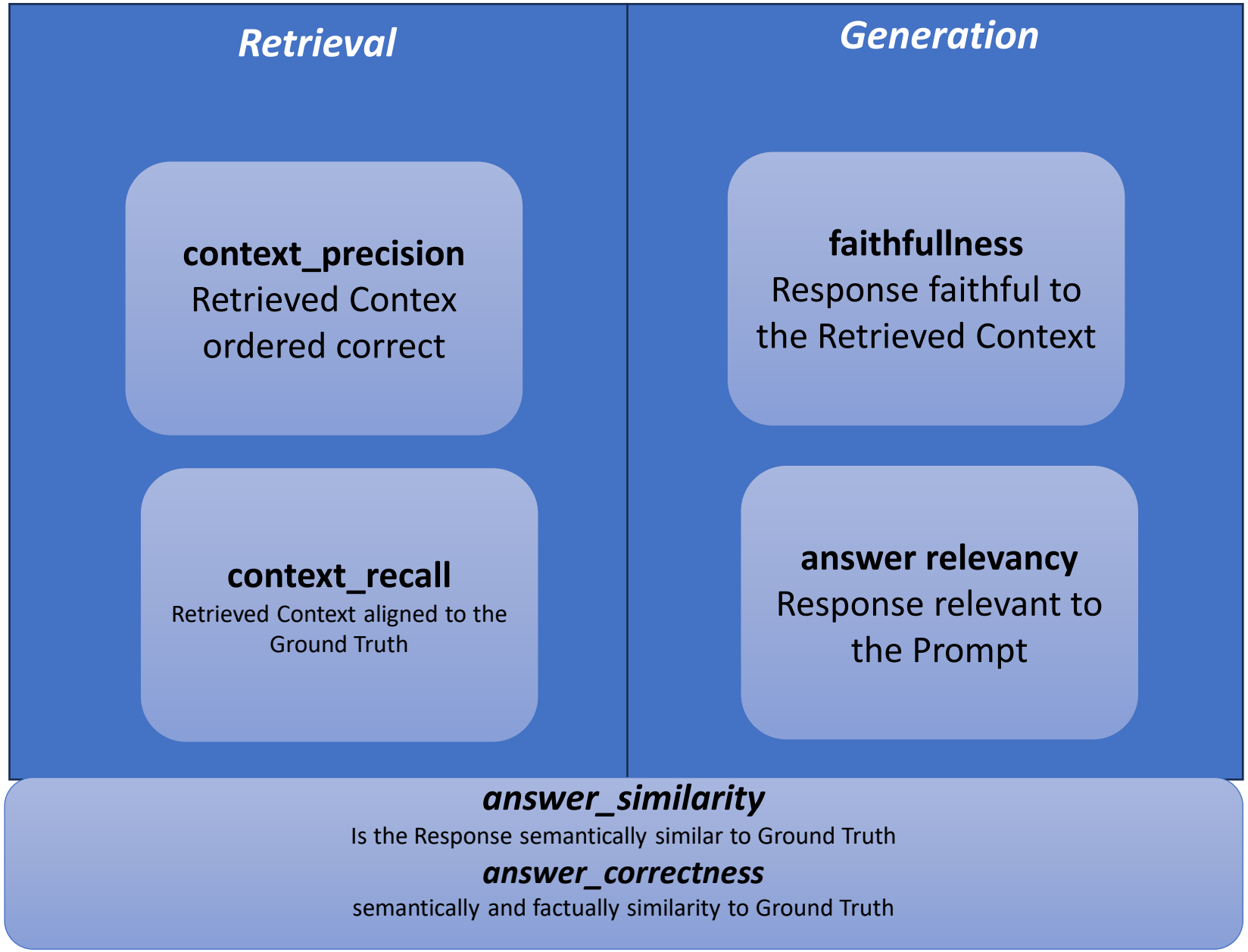
- Human Centric Evaluation
  - Beste Method

    Aber,
  - Subjektif
  - Zeitaufwändig
  - Teuer
  - Erfordert Expert
- Retrieval Augmented Generation Assessment(RAGAS)
  - ein reference-free evaluation Framework
  - Retriever und Generator Evaluation

RAGAS

- Input Parameter:
- Question: anhand derer Ihre RAG-Pipeline bewertet wird
- Response: generiert von RAG-System
- Contexts: an das LLM übergeben wurden, um die Frage zu beantworten.
- Ground_truths: Die Ground Truth Antwort auf die Fragen(nur bei contex-recall)

Ragas Score


Retrieval
Generation
context_precision
Retrieved Contex
ordered correct
faithfullness
Response faithful to
the Retrieved Context
context_recall
Retrieved Context aligned to the
Ground Truth
answer relevancy
Response relevant to
the Prompt
answer_similarity
Is the Response semantically similar to Ground Truth
answer_correctness
semantically and factually similarity to Ground Truth

	question	ground_truths	answer	contexts	context_relavency	faithfulness	answer_relevancy
0	How to deposit a cheque issued to an associate...	[Have the check reissued to the proper payee.J...	\nThe best way to deposit a cheque issued to a...	[Just have the associate sign the back and the...	0.867	1.0	0.922
1	Can I send a money order from USPS as a business?	[Sure you can. You can fill in whatever you w...	\nYes, you can send a money order from USPS as...	[Sure you can. You can fill in whatever you w...	0.855	1.0	0.923
2	1 EIN doing business under multiple business n...	[You're confusing a lot of things here. Compan...	\nYes, it is possible to have one EIN doing bu...	[You're confusing a lot of things here. Compan...	0.768	1.0	0.824
3	Applying for and receiving business credit	["I'm afraid the great myth of limited liabili...	\nApplying for and receiving business credit c...	[Set up a meeting with the bank that handles y...	0.781	1.0	0.830
4	401k Transfer After Business Closure	[You should probably consult an attorney. Howe...	\nIf your employer has closed and you need to ...	[The time horizon for your 401K/IRA is essenti...	0.737	1.0	0.753

Fazit

- RAG ist momentan beste QA-Methode
- Testmethod-RAGAS
- Knowledge Base
- Testdaten
- Sehr dynamisher Bereich
- Große Theorie Gebiet(Question-Answering und LLM)

Ausblick

- Advance RAG
- Moduler RAG

Quelle

- https://docs.ragas.io/en/stable/
- https://python.langchain.com/docs/use_cases/question_answering/
- https://arxiv.org/abs/2312.10997
- https://arxiv.org/abs/2005.11401
- https://arxiv.org/abs/2309.15217
- https://www.youtube.com/watch?v=WUxksE41woY&t=1083s
- https://www.youtube.com/watch?v=pTszM3YN7_8&t=1s
- https://ronxin.github.io/wevi/
- https://www.chatpdf.com/
- https://cbarkinozer.medium.com/almayla-artt%C4%B1r%C4%B1lm%C4%B1%C5%9F-%C3%BCretim-y%C3%B6ntemi-rag-nedir-e0ac458de13f#:~:text=Retrieval%20Augmented%20Generation%20(RAG)%20modeli,sa%C4%9Flamak%20%C3%BCzere%20tasarlanm%C4%B1%C5%9F%20bir%20mimaridir.

Danke für Ihre Aufmerksamkeit

FRAGEN? ☺