

Wie testet man automatisiert die Antwortqualität einer lokalen RAG-Anwendung?

10.07.2024

Mannheim

Jacques Huss
jacques.huss@andrena.de

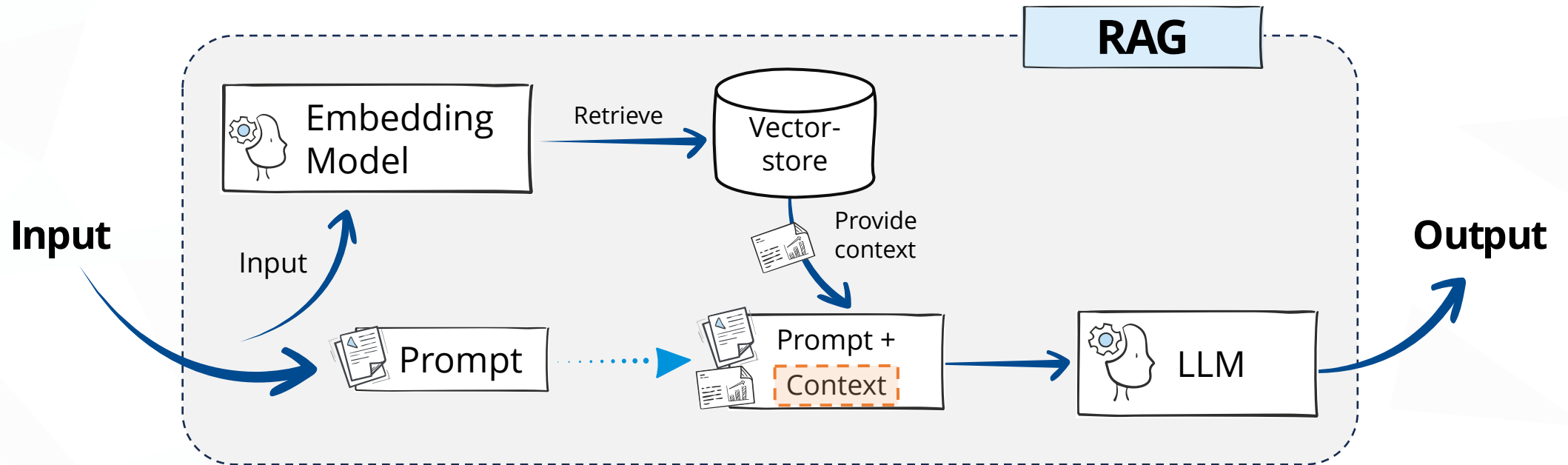
Dr. Anja Kleebaum
anja.kleebaum@andrena.de



Agenda: Test der Antwortqualität einer lokalen RAG-Anwendung

- Vorstellung lokale RAG-Anwendung
- Übersicht Testarten für KI-Anwendungen
- Test der Antwortqualität

Retrieval Augmented Generation (RAG)



Unsere lokale RAG-Anwendung: KI-ChatBot „SmartWiki“

andrena
OBJECTS

SmartWiki

Hallo! Welche Frage zum andrena-Wiki kann ich dir beantworten?
(Die Fragen und Antworten werden anonymisiert aufgezeichnet.)

SmartWiki ist ein Projekt, das den Ziel hat, einen Chatbot zu entwickeln, der Nutzerfragen mit datenbasierten Antworten und Quellenangaben versorgt. Der Chatbot wird mit internem Wissen angereichert und alle KI-Komponenten laufen komplett lokal und sind selbst gehostet.

Quellen ^

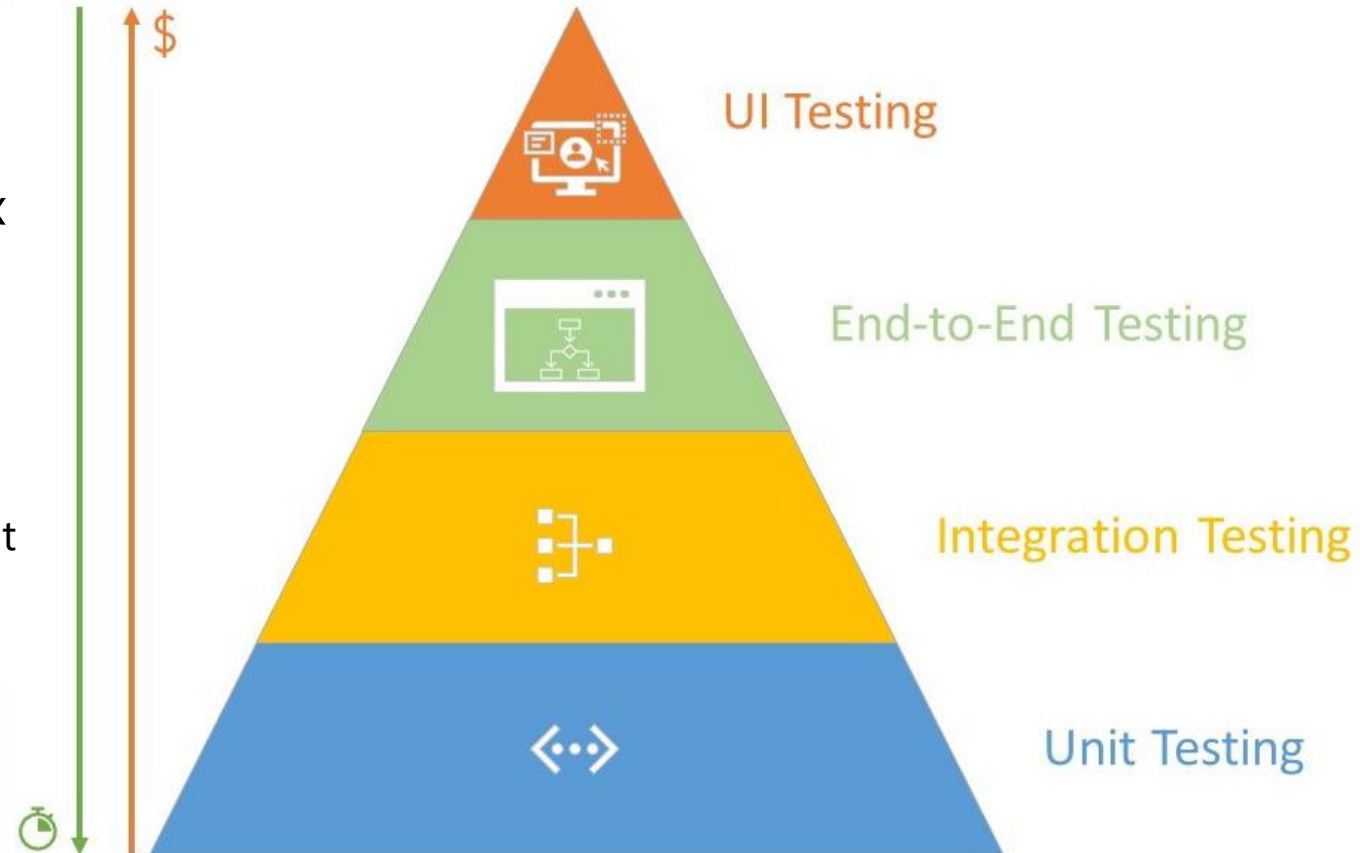
- <https://mail.andrena.de/wiki/Andrena/MERCURYII-SmartWiki>
- <https://mail.andrena.de/wiki/Mannheim/WebLeftBar>

Was ist SmartWiki?

Frage mich etwas!

Übersicht Testarten für KI-Anwendungen

- Testpyramide
- Tests von Modellen: End-to-End, Blackbox
 - Performance-Tests
 - Zeit bis zum ersten Token und Antwort
 - Lastentests: Antwortzeit in Abhängigkeit von Nutzeranzahl
 - Test der Antwortqualität



Welche Evaluationsplattformen gibt es?

- CI-Plattform?
- Langfuse
- LangSmith
- ...

The screenshot displays the Langfuse v2.60.4 dashboard for a project named 'Smartwiki - RAG - PROD'. The interface is divided into several sections:

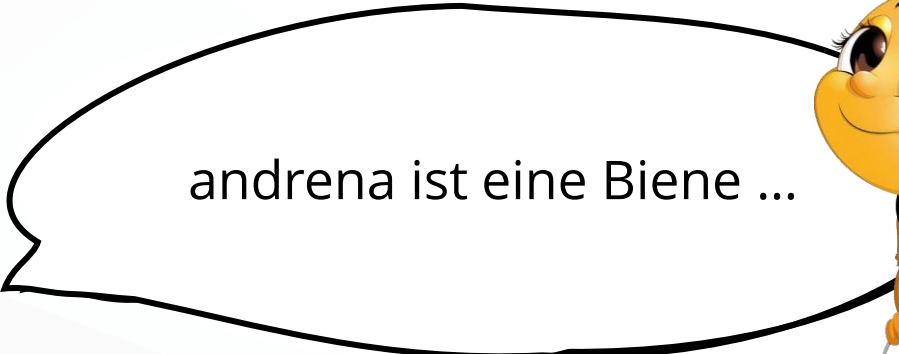
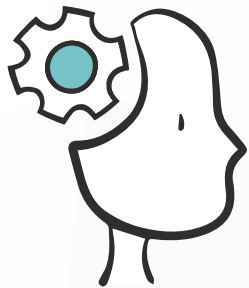
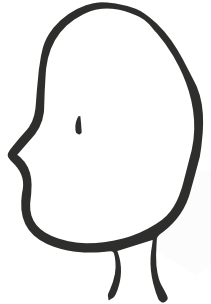
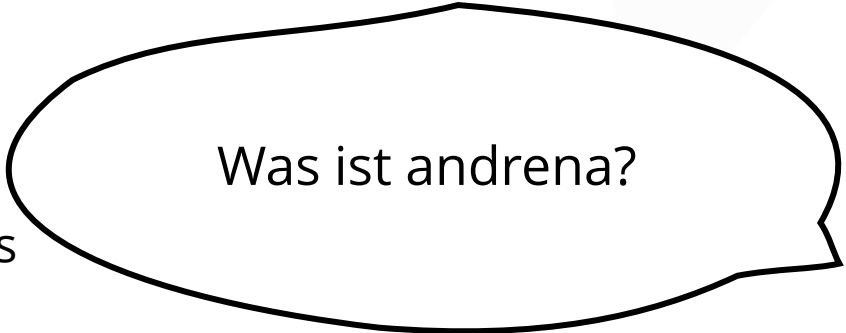
- Traces:** A list of 785 total traces tracked. A table shows the top 5 logs with their respective counts: 'Log: Was ist Mercury?' (27), 'Log: Wann ist das Kickerturnier 2024?' (13), 'Log: Wie kann ich meine Reisekosten einreichen?' (12), 'Log: Was ist andrena?' (11), and 'Log: Was ist SmartWiki?' (10).
- Model costs:** A summary showing a total cost of \$0.00. A table lists the models used: 'llama3:8b-instruct-q8_0' and 'solar:10.7b-instruct-v1-q8_0', both with 0 tokens and \$0 cost.
- Scores:** A summary showing 60 total scores tracked. A table lists the scores: 'user_feedback' with a count of 60, an average score of -0.07, and 0 out of 1 scores.
- Traces (Time Series):** A line chart showing the number of traces tracked over time from 4/9/24 to 7/8/24. The y-axis ranges from 0 to 60. The chart shows a fluctuating but generally increasing trend in the number of traces.
- Scores (Time Series):** A line chart showing the average score per name over time from 4/9/24 to 7/8/24. The y-axis ranges from -1 to 1. The chart shows a highly volatile score for 'user_feedback', with several peaks near 1 and troughs near -1.

A yellow callout box with the text 'ermöglicht Monitoring über Zeit' (enables monitoring over time) is overlaid on the dashboard, pointing to the time-series charts.

Wie baut man einen Evaluationsdatensatz (Goldstandard/ground truth) auf?

Wichtige Begriffe

- True positives (TP)
 - False negatives (FN)
 - False positives (FP)
- Included terms
- Excluded terms



Wie baut man einen Evaluationsdatensatz (Goldstandard/ground truth) auf?

○ **Manuelle Erzeugung**

- Domäne Experten befragen
- Was wollen mögliche Nutzer wissen?
- Kann im späteren Verlauf auf Basis von Nutzeranfragen geschehen
- Nachteil: teuer & zeitintensiv
- Vorteil: verlässlich & anwendungsspezifisch

Wie baut man einen Evaluationsdatensatz (Goldstandard/ground truth) auf?

- Liste von Fragen mit erwarteten Antworten, included/excluded terms, Quellen

Langfuse v2.60.4

Smartwiki - RAG(e) - DEV > Datasets > smart-wiki > Items

smart-wiki

+ New item

(4/4) Runs Items

Item id	Source	Status	Created At	Input	Expected Output	Metadata	Actions
...24gkmzq		● ACTIVE	2.4.2024, 11:58:05	{ "questions": ["Wie kocht man Lasagne?"] }	{ "answer": "Ich weiß es nicht, da die gegebenen Kontextinformationen..." }		⋮
...0l0w2kk		● ACTIVE	2.4.2024, 11:56:52	{ "questions": ["Gib mir Empfehlungen für Hotels in Mannheim"] }	{ "answer": "Für Hotels in Mannheim gibt es folgende Empfehlungen: ..." }		⋮
...lardie1		● ACTIVE	2.4.2024, 11:32:54	{ "questions": ["Wie viele Tage pro Jahr für Weiterbildungen bekomme..."] }	{ "answer": "Nach Absolvierung des ASE-Trainingsprogramms erhalten ..." }		⋮
...unojtg		● ACTIVE	2.4.2024, 11:28:16	{ "questions": ["Welche Kurse besucht man im ASE Trainingsprogram..."] }	{ "answer": "Im ASE Trainingsprogramm Junior besuchen die Teilnehm..." }		⋮
...t8rjil6		● ACTIVE	2.4.2024, 11:02:11	{ "questions": ["Wie oft findet der Open Friday statt?"] }	{ "answer": "Der Open Friday findet monatlich statt.", "exclude": [], "inc..." }		⋮
...jenhu6f		● ACTIVE	2.4.2024, 11:02:11	{ "questions": ["Wo fand das Stuttgarter Sommerfest 2023 statt?"] }	{ "answer": "Das Stuttgarter Sommerfest 2023 fand im Waldheim Hesla..." }		⋮
...f97qfw3		● ACTIVE	2.4.2024, 10:55:20	{ "questions": ["Was ist der Open Friday?"] }	{ "answer": "Der Open Friday ist eine monatliche Mini-Konferenz für A..." }		⋮
...zup20b7		● ACTIVE	2.4.2024, 10:50:39	{ "questions": ["Welche Außeneinsätze fanden im Jahr 2023 statt?"] }	{ "answer": "2023 fanden folgende Außeneinsätze statt: 1. Kunst, Kultur..." }		⋮
...phknh75		● ACTIVE	2.4.2024, 10:42:21	{ "questions": ["Was ist ein Außeneinsatz?"] }	{ "answer": "Ein Außeneinsatz bei Andrena bezieht sich auf eine außer..." }		⋮
...03aqa7x		● ACTIVE	2.4.2024, 10:26:08	{ "questions": ["Wie lautet die aktuelle Adresse des Kölners?"] }	{ "answer": "Die Adresse des Kölner Büros ist Josef-Lammerting-Allee 2..." }		⋮

schauen wir uns genauer an...

Wie baut man einen Evaluationsdatensatz (Goldstandard/ground truth) auf?

Smartwiki - RAG(e) - DEV > Datasets > smart-wiki > Items > dui4s7f000ls6m367phknh75

Dataset Item

Input

```
1 {
2   "questions": [
3     "Was ist ein Außeneinsatz?"
4   ]
5 }
```

Expected output

```
1 {
2   "answer": "Ein Außeneinsatz bei Andrena bezieht sich auf eine außerhalb stattfindende Veranstaltung oder
Aktivität, die von diesem Unternehmen organisiert und angeboten wird. Diese können Wandern, Paddeln, Malen,
Mountainbiken, Survival-Trainings oder kulturelle Ausflüge umfassen. Die genauen Termine, Inhalte und
Ansprechpartner sind auf den entsprechenden Webseiten dokumentiert.",
3   "exclude": [
4     {
5       "variants": [
6         "Girls Day"
7       ]
8     }
9   ],
10  "include": [
11    {
12      "variants": [
13        "Veranstaltung",
14        "Aktivität"
15      ]
16    },
17    {
18      "variants": [
19        "Wandern",
20        "Paddeln",
21        "Malen",
22        "Mountainbiken",
23        "Survival-Training",
24        "kulturelle Ausflüge"
25      ]
26    }
27  ],
28  "sources": [
29    "https://mail.andrena.de/wiki/Andrena/ArchivBisherigerAusseneinsaetze"
30  ]
}
```

Wie baut man einen Evaluationsdatensatz (Goldstandard/ground truth) auf?

Dataset Item

Input

1	{
2	"questions": [
3	"Was ist ein Außeneinsatz?"
4]
5	}

Wie baut man einen Evaluationsdatensatz (Goldstandard/ground truth) auf?

Expected output

```
1 {  
2   "answer": "Ein Außeneinsatz bei Andrena bezieht sich auf eine außerhalb stattfindende Veranstaltung oder  
Aktivität, die von diesem Unternehmen organisiert und angeboten wird. Diese können Wandern, Paddeln, Malen,  
Mountainbiken, Survival-Trainings oder kulturelle Ausflüge umfassen. Die genauen Termine, Inhalte und  
Ansprechpartner sind auf den entsprechenden Webseiten dokumentiert.",
```

Wie baut man einen Evaluationsdatensatz (Goldstandard/ground truth) auf?

```
10  "include": [  
11    {  
12      "variants": [  
13        "Veranstaltung",  
14        "Aktivität"  
15      ]  
16    },  
17    {  
18      "variants": [  
19        "Wandern",  
20        "Paddeln",  
21        "Malen",  
22        "Mountainbiken",  
23        "Survival-Training",  
24        "kulturelle Ausflüge"  
25      ]  
26    }  
27  ],
```

Wie baut man einen Evaluationsdatensatz (Goldstandard/ground truth) auf?

```
28  "sources": [  
29    "https://mail.andrena.de/wiki/Andrena/ArchivBisherigerAusseneinsaetze"  
30  ]
```

Wie baut man einen Evaluationsdatensatz (Goldstandard/ground truth) auf?

```
3   "exclude": [  
4     {  
5       "variants": [  
6         "Girls Day"  
7       ]  
8     }  
9   ],
```

Wie baut man einen Evaluationsdatensatz auf?

- **Bestehende Benchmarks ([MTEB](#), [Chatbot Arena](#))**
- Können hilfreich sein zur Auswahl von LLM und Embedding-Model
- Kein Anwendungsbezug
- Nicht sehr aussagekräftig

Wie baut man einen Evaluationsdatensatz auf?

- **Automatische/synthetische Erzeugung** von Fragen und Antworten möglich
- z.B. durch RAGAS-Framework, Auto Evaluator, ...
- Man kann schnell viele Testdaten erzeugen
- Problem: Open source LLM's kommen an ihre Grenzen
- Generierte Fragen decken sich nicht immer mit Nutzerfragen
- Halluzinationen können Integrität des Datensatzes beeinflussen

Wie kann man bewerten, ob Antwort mit Erwartung übereinstimmt?

Abgleich von zurückgegebener mit erwarteter Antwort:

- **manuell** durch Menschen (human in loop)
- **reguläre Ausdrücke**
- Bewertung durch Modelle mit **Schüler-Lehrer-Ansatz (LLM as a judge)**
 - LLM findet true positives, false positives und false negatives
 - Retrieval Augmented Generation Assessment (RAGAS)



generiert mit ChatGPT 4o

Wie kann man bewerten, ob Antwort mit Erwartung übereinstimmt?

Die Evaluationsplattform (Langfuse) ermöglicht...

- Experimente zur Bewertung der Antwortqualität und -zeit des ChatBots
- Erfassung von **ChatBot-Konfiguration** wie Prompt, LLM, Chunk-Größe, Embedding-Model

Smartwiki - RAG(e) - DEV > Datasets > smart-wiki-history

smart-wiki-history ¹

Evaluationsdurchläufe mit Durchschnittsmetriken

(7/7) v

Name	Description	Run Items	Latency (avg)	Total Cost (avg)	Scores (avg)							
2024-07-02-16-42-08-FLUSH-TEST		10	0.00s	\$0.00	c__answer_f1	c__answer_precision	c__answer_recall	c__context_f1	c__context_precision	c__context_recall	c__context_sources_recall	c__is_german_answer
					0.34	0.40	0.33	0.59	0.56	0.56	0.74	1.00
2024-07-02-13-36-37-TEST		10	0.00s	\$0.00	c__answer_f1	c__answer_precision	c__answer_recall	c__context_f1	c__context_precision	c__context_recall	c__context_sources_recall	c__is_german_answer
					0.34	0.40	0.33	0.59	0.56	0.50	0.74	1.00
exp-2024-07-01-21-17-53-llama3:8b-instruct-q8_0		3	0.00s	\$0.00	c__answer_f1	c__answer_precision	c__answer_recall	c__context_f1	c__context_precision	c__context_recall	c__context_sources_recall	c__is_german_answer
					0.67	0.67	0.67	0.83	0.75	0.75	0.67	1.00

- Automatisierung in CI/CD-Pipeline

Welche Metriken kann man für Retrieval und Generation erheben?

○ Metriken für Retrieval:

- context recall, context precision, context F1
- context sources recall, ...

○ Metriken für Generation:

- answer recall, answer precision, answer F1
- RAGAS: answer similarity, answer correctness, faithfulness, answer relevancy
- friendliness, harmfulness, conciseness, ...

$$recall = \frac{\#true\ positives}{\#true\ positives + \#false\ negatives}$$

$$precision = \frac{\#true\ positives}{\#true\ positives + \#false\ positives}$$

$$F1 = 2 \frac{recall \cdot precision}{recall + precision}$$

Wie kann man Nutzerfeedback sammeln und zur Verbesserung verwenden?

Was ist SmartWiki?

SmartWiki ist ein Projekt, das den Ziel hat, einen Chatbot zu entwickeln, der Nutzerfragen mit datenbasierten Antworten und Quellenangaben versorgt. Der Chatbot wird mit internem Wissen angereichert und alle KI-Komponenten laufen komplett lokal und sind selbst gehostet.

Quellen ^

– <https://mail.andrena.de/wiki/Andrena/MERCURYII-SmartWiki>

Nutzer gibt Feedback

Langfuse v2.60.4

Traces

Search by id, name, user id

Filter Timestamp > 1.7.2024

ID	Timestamp	Name	User ID	Session ID	Latency	Usage	Cost	Scores
...c812fae	8.7.2024, 20:32:46	Log: Was ist SmartWiki?		U5KEzsf5ksLyMX8rAAAD	18.30s		\$0.00	user_feedback 1.00

Feedback wird in Langfuse erfasst

Langfuse zeigt Fragen und Antworten des ChatBots (anonymisiert)

Wie können Chatverläufe (nicht nur Einzelfragen) getestet werden?

Langfuse v2.60.4

Smartwiki - RAG(e) - DEV > Datasets > smart-wiki-history >

Dataset Item

Input	Expected output
<pre> 1 { 2 "questions": [3 "Wie lautet die aktuelle Adresse des Karlsruher Büros?", 4 "Und die von Köln?" 5] 6 } </pre>	<pre> 1 { 2 "answer": "Die Adresse des Kölner Büros ist Josef-Lammerting-Allee 25, 50933 Köln.", 3 "exclude": [4 { 5 "variants": [6 "Widdersdorfer Straße 262, 50933 Köln" 7] 8 } 9], 10 "include": [11 { 12 "variants": [13 "Josef Lammerting Allee 25, 50933 Köln", 14 "Josef-Lammerting-Allee 25, 50933 Köln" 15] 16 } 17], 18 "sources": [19 "https://mail.andrena.de/wiki/Koeln/Kontaktdaten" 20] 21 } </pre>

2 Fragen in Folge

erwartete Antwort auf Frage 2

Zusammenfassung: Test der Antwortqualität einer lokalen RAG-Anwendung

- Testen von KI-Anwendungen speziell wegen **Blackbox-Modellen**
- **Goldstandard** nötig
 - Fragen und erwartete Antworten für ChatBot
 - Synthetisch oder manuell erstellt
- Bei RAG: Test von **Retrieval und Generation**
- Viele **Metriken**: precision, recall, user feedback, ...
- **Bewertung** durch Menschen, reguläre Ausdrücke oder **Model-as-a-judge**
- **Evaluations- und Monitoringplattform** wie Langfuse oder LangSmith nützlich
- Ausblick: **Automatische Optimierung** auf Basis der Evaluation



generiert mit ChatGPT 4o