

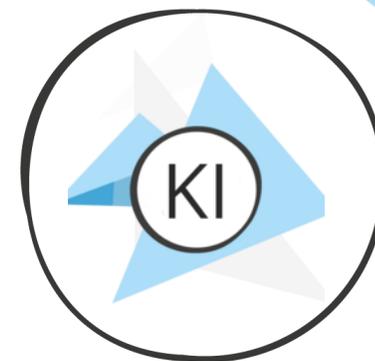
Der Weg zum autarken KI-Chatbot: Ein Praxisbericht

31.07.2024

Java Forum Stuttgart

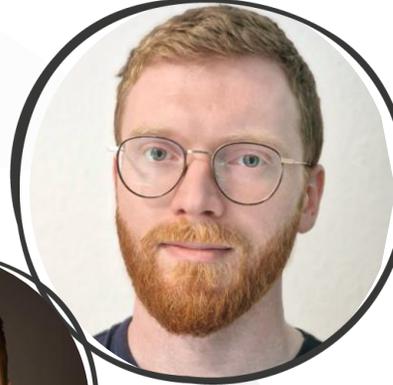
Dr. Anja Kleebaum
anja.kleebaum@andrena.de

Jacques Huss
jacques.huss@andrena.de



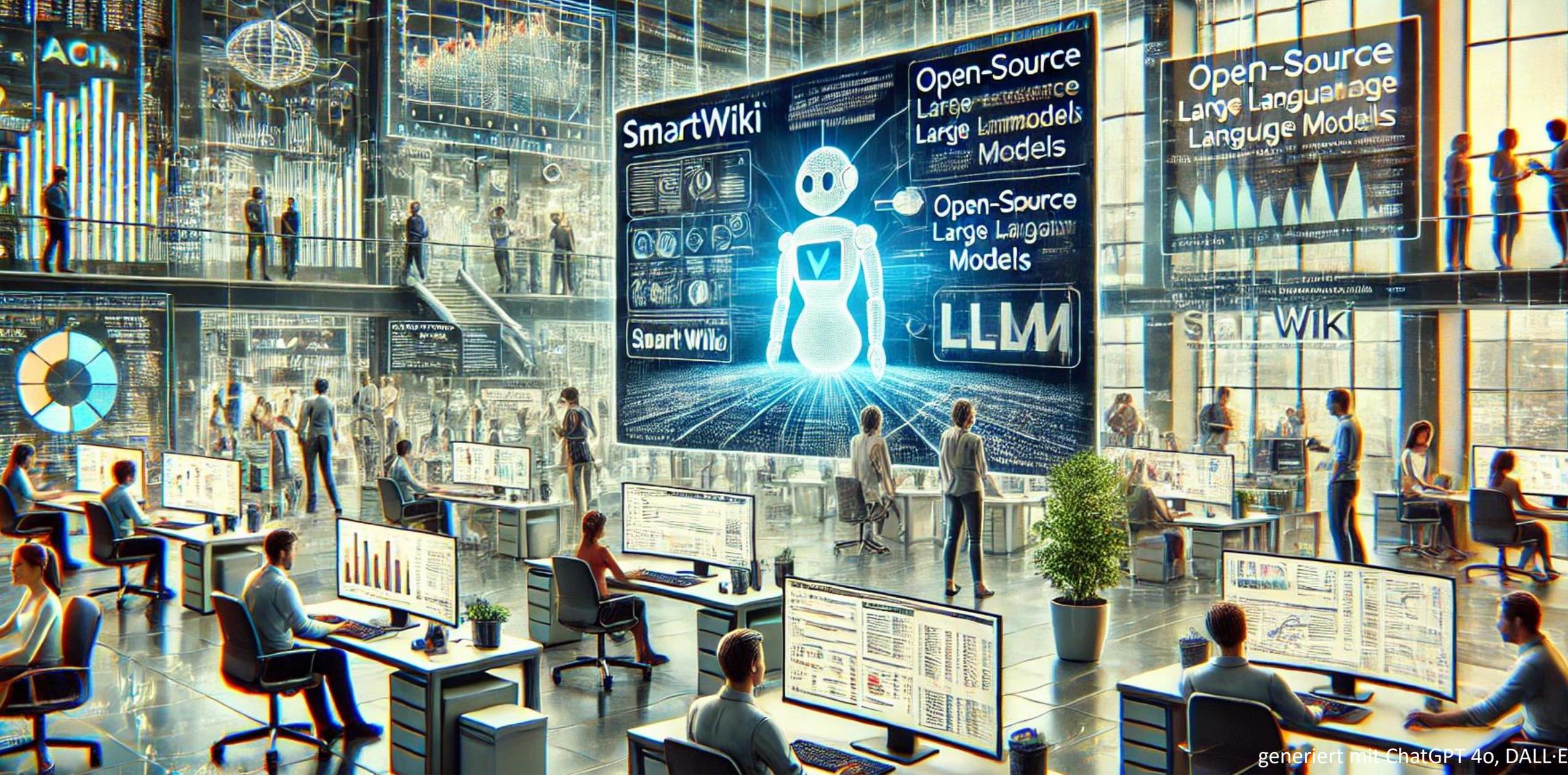
Wer sind wir?

- Teil von internem KI-Projekt „Mercury“
bei andrena objects



Agenda: Der Weg zum autarken KI-Chatbot: Ein Praxisbericht

- Vorstellung lokale RAG-Anwendung: KI-Chatbot „SmartWiki“
- Betrieb von Open-Source-LLMs
- Modellauswahl
- Testen und Evaluation
- Nutzerfeedback
- RAG-Optimierung

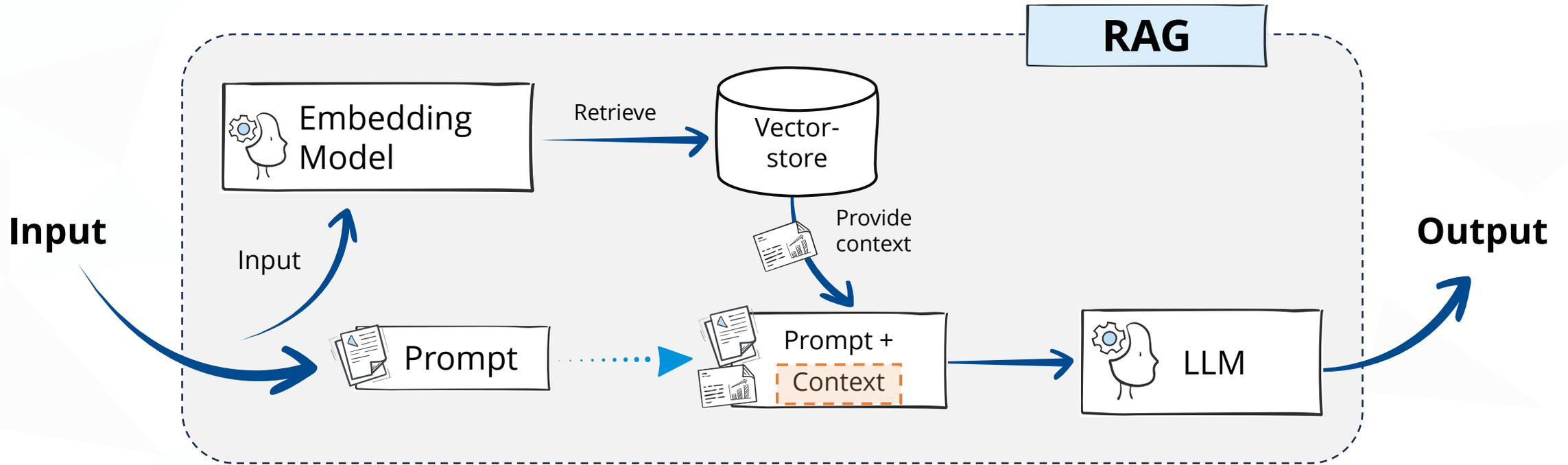


generiert mit ChatGPT 4o, DALL·E

Vorstellung lokale RAG-Anwendung: KI-Chatbot „SmartWiki“



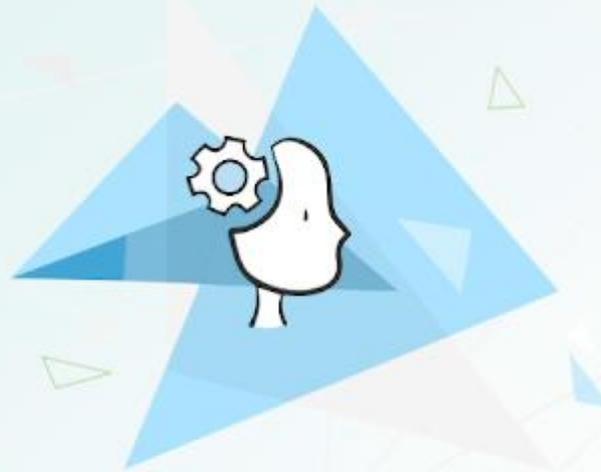
Retrieval Augmented Generation (RAG)





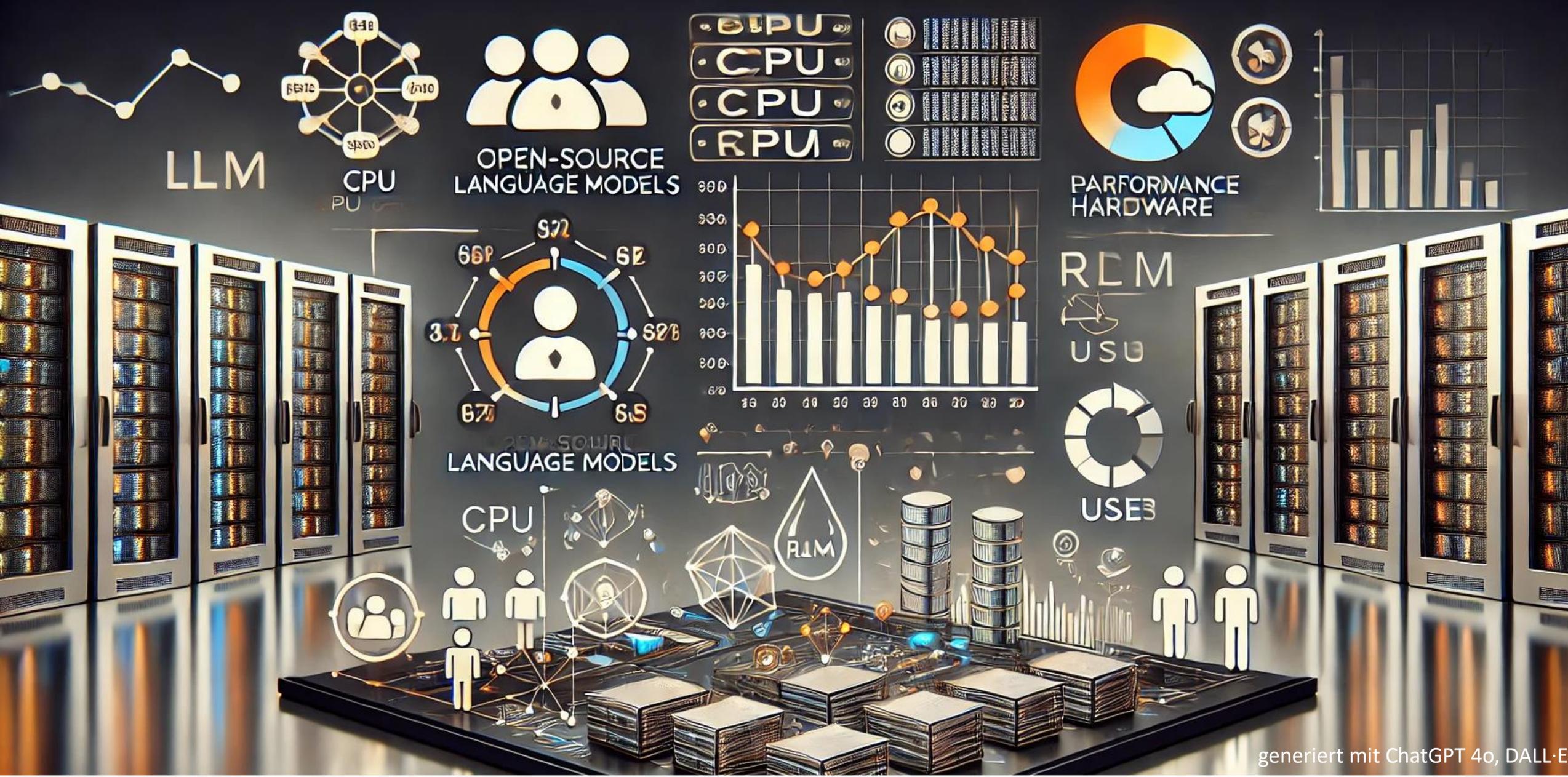
Hallo! Welche Frage zum andrena-Wiki kann ich dir beantworten?
(Die Fragen und Antworten werden anonymisiert aufgezeichnet.)

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



Frage mich etwas!





generiert mit ChatGPT 4o, DALL-E

Betrieb von Open-Source-LLMs

Welche Infrastruktur verwenden wir für den Betrieb des Chatbots?

○ **Einfache Virtual Machine** (VM ohne GPU)



Pro: geringe Kosten für Anschaffung und Betrieb



Contra: keine GPU, Antwort von LLM mit Kontext dauert mehrere Minuten

○ **Betrieb in Cloud** (z.B. bei Hugging Face, Amazon Titan, NVIDIA NeMo, ...)



Pro:

- leicht skalierbar, einfacher aufzusetzen
- ermöglicht auch Nutzung performanter proprietärer Modelle, Finetuning



Contra: Kontrollverlust



NVIDIA.



Welche Infrastruktur verwenden wir für den Betrieb des Chatbots?

● **Eigene Hardware mit GPU** (z.B. Lenovo ThinkSystem, MacStudio, ...)



Pro:

- geringe Kosten für Betrieb
- Datensicherheit



Contra:

- Upfront Invest: Kosten für Anschaffung (z.B. MacStudio ~ 3k Euro)
- aufwändig in Wartung
- Skalierbarkeit begrenzt
- bei Mac: Dienste laufen nativ, docker kann GPU nicht nutzen

Wie betreibt man ein LLM lokal?

○ LM Studio



Pro:

- einfach zu benutzen für erste Versuche
- ermöglicht es Modelle von HuggingFace herunterzuladen
- man kann Server starten mit OpenAI-REST-API



Contra: unnötige GUI, closed source, nur für MacOS und Windows

○ Selbst-gebauter Service mit CTransformers, LlamaCpp, vllm, ...



Pro: mehr Flexibilität bei der Entwicklung



Contra: aufwändig, Streaming und Umgang mit parallelen Fragen nicht unterstützt



LM Studio

LLaMA C++

Wie betreibt man ein LLM lokal?

LocalAI



Pro:

- Server mit OpenAI-REST-API
- ermöglicht auch Bildgenerierung, Audiotranskription



Contra: benötigt viel Speicherplatz (>30 GByte)

Ollama



Pro:

- unterstützt parallele Anfragen
- schnellste Antwort bei Benchmark-Tests auf MacStudio



Wie betreibt man ein Embedding-Model lokal?



○ Ollama



Pro: betreiben wir schon für LLM



Contra: bietet noch keine Funktionalität für Scraping, Chunking, Kontext-Retrieval, ...

○ Selbst-gebauter Service mit HuggingFaceEmbeddings



Pro:

- stellt Funktionalität für Scraping, Chunking, Kontext-Retrieval bereit
- bietet Schnittstelle zu Vektordatenbank
- Retrieval geht schnell auch ohne GPU



Contra: viel manuelle Arbeit, RAG-Optimierung muss man selbst durchführen

Modellauswahl

- **Bestehende Benchmarks** hilfreich für Auswahl

- Embedding-Model: [Massive Text Embedding Benchmark \(MTEB\)](#)
- LLM: [Chatbot Arena](#)

- **Nachteile:** 

- schnell bewegend => leichte Austauschbarkeit wichtig
- kein Anwendungsbezug => eigene Evaluation essenziell!

- **Guter Startpunkt:**

- Embedding-Model: multilingual-e5-large
- LLM: llama3:8b





LUSTALLY
OPERATED
METRICS

PRECISION

3:3.2

Precision

RECALL

3:3.6

CSV L-Score

F-SCORE

3:3.8

Custom Lexical Metrics

LAS-SCORE



EVALUATION

Model	Score	Notes
Model A	0.95	High precision
Model B	0.92	Good recall
Model C	0.90	Stable performance
Model D	0.88	Needs tuning
Model E	0.85	Low accuracy

LOCALLY OPERATED LM METRICS

2.9.9

LANG-METRICS



CSV DATA



CUSTOM LEXICAL METRICS

METR-METRICS

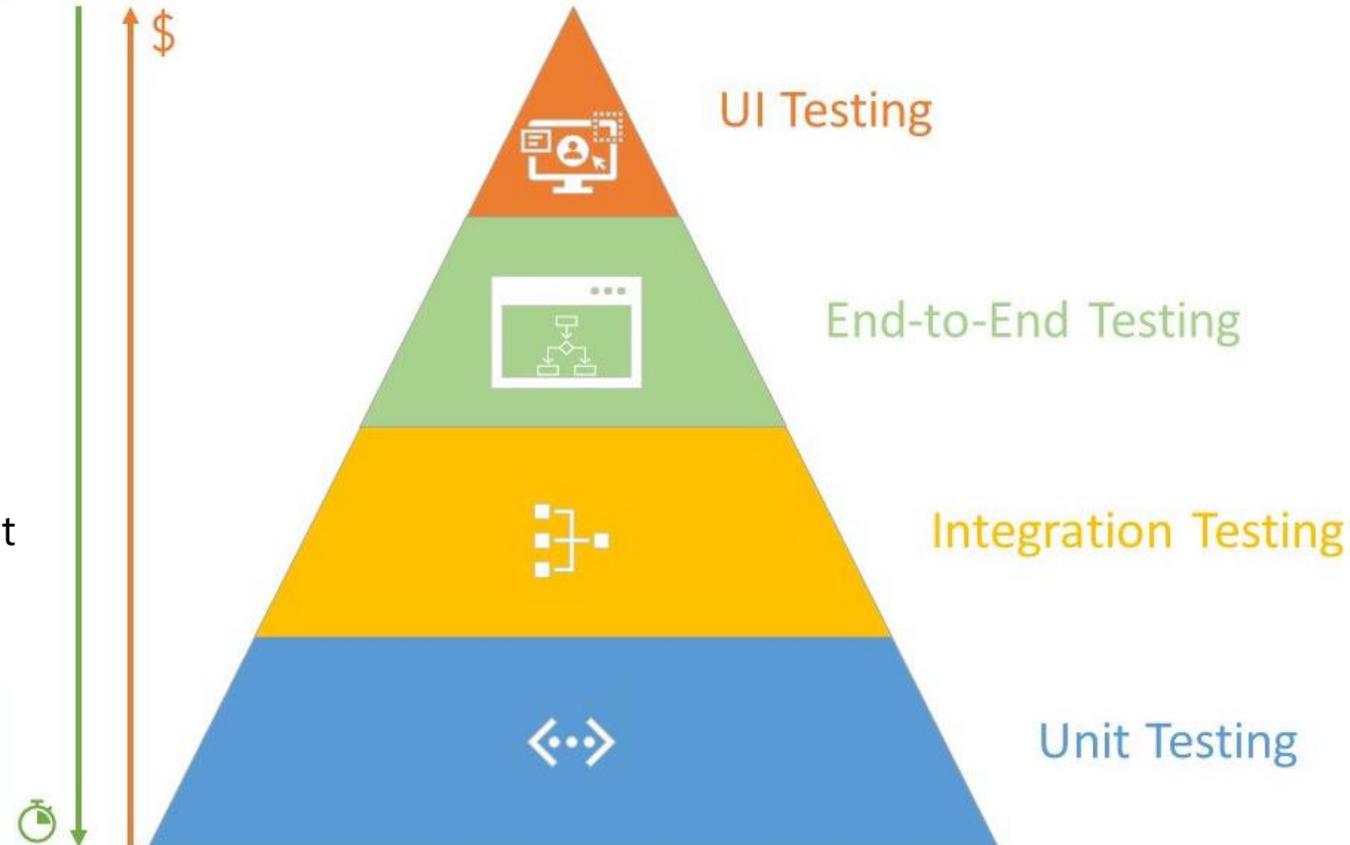


Testing und Evaluation

generiert mit ChatGPT 4o, DALL-E

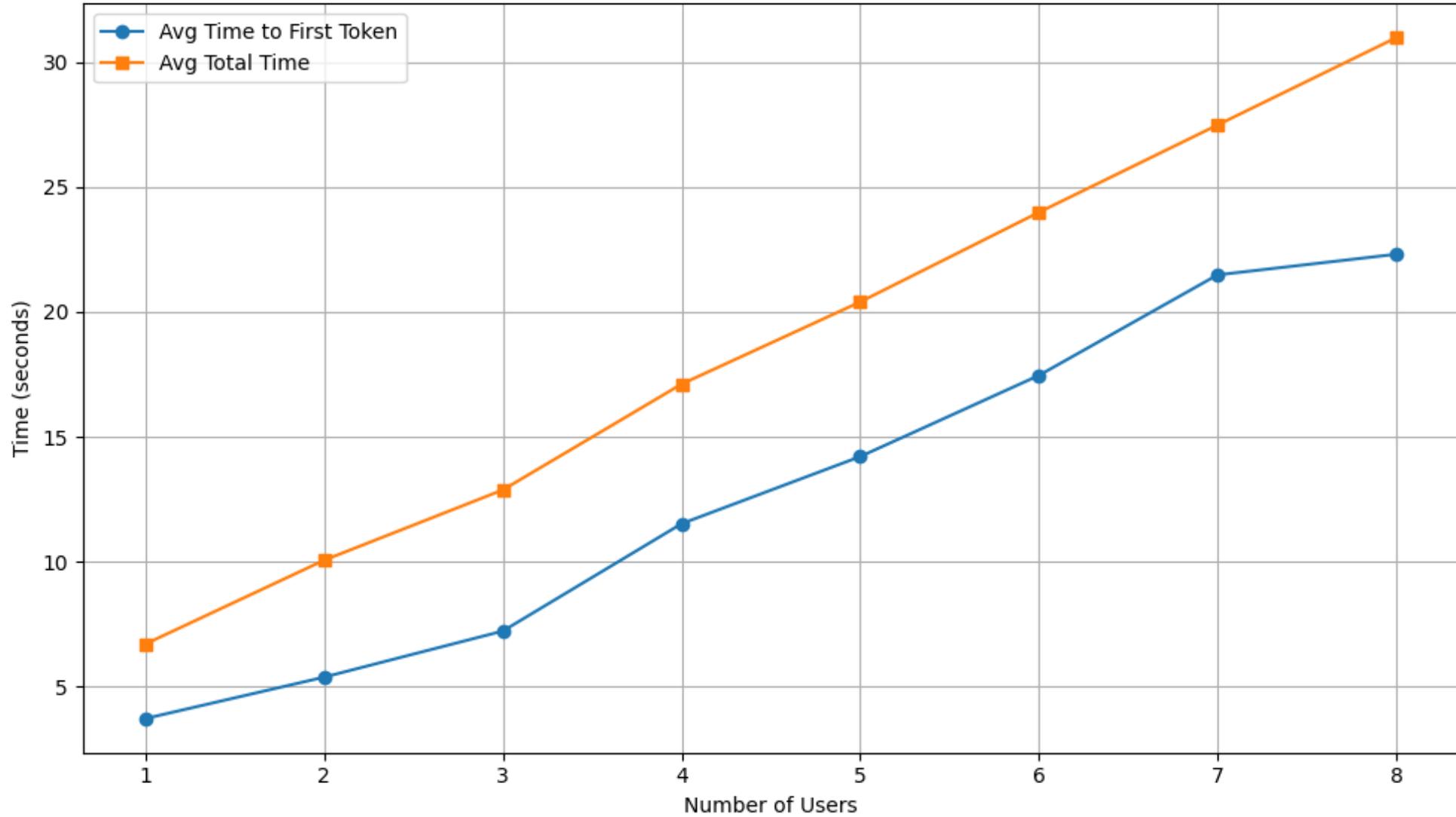
Übersicht Testarten für KI-Anwendungen

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



Performance-Tests

API Performance under Different User Loads



Welche Evaluationsplattformen gibt es?

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

Langfuse v2.60.4
Smartwiki - RAG - PROD
Apr 09, 24 : 20:12 - Jul 08, 24 : 20:12
3 months
Request Chart

- Dashboard
- Tracing
- Sessions
- Generations
- Scores
- Models
- Users
- Prompts
- Datasets

Traces

785 Total traces tracked

- 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
- Log: Was ist SmartWiki? 10

Model costs

\$0.00 Total cost

Model	Tokens	USD
llama3:8b-instruct-q8_0	0	\$0
solar:10.7b-instruct-v1-q8_0	0	\$0

Scores

60 Total scores tracked

Name	#	Avg	0	1
user_feedback	60	-0.07	0	28

ermöglicht Monitoring über Zeit

Traces

785 Traces tracked

Scores

Average score per name

Settings
Docs
Support
Feedback

Project
+ New

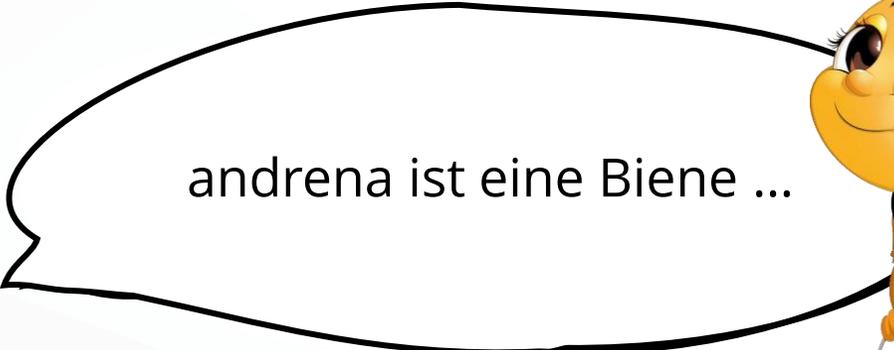
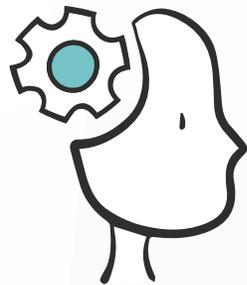
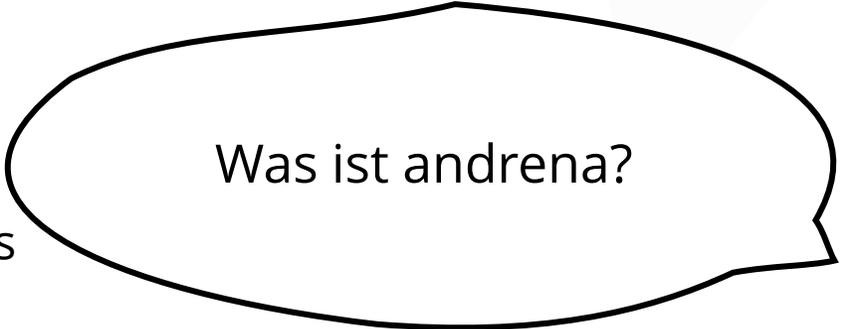
Smartwiki - RAG - ...

A Anja

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..." }		⋮
...010w2kk		● 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 Hesia..." }		⋮
...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?

- **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) ▾

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, ...

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

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

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



 Positive		'Wrong (Halluciation) 'Off-Topic''		<ul style="list-style-type: none">• 'Wrong'• Off-Topic• Not Understood
 Off-topic		'Wrong 'Off-Topic'' 'Off-Turnsin''		<ul style="list-style-type: none">• Not Understank
 Wrong		'Not topic' (Feedbacke)		
 Off-topic				
 Nonment Ndt-not feedback				



generiert mit ChatGPT 4o, DALL-E

Nutzerfeedback

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	Feedback	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)



RAG-Optimierung

generiert mit ChatGPT 4o, DALL·E

RAG-Optimierung

- Ansätze für Optimierung:
 - Ingestion für Tabellen und Bilder
 - Semantic text splitting
 - Advanced retrieval techniques
 - Reranking
 - Guardrails
 - ...
- **Wichtig:** Optimierung **auf Basis der Evaluationsergebnisse!**
 - **Automatische Optimierung** möglich für RAG-Parameter wie Chunk-Size, ...
 - Kreuzvalidierung: Aufspaltung des Goldstandards in Trainings- und Testdatensatz

Ingestion für Tabellen: Wie gehen wir bei Website-Scraping mit Tabellen um?

- Problem: einfaches Wegfiltern von HTML-Tags führt zu Informationsverlust
- Alternativen:
 - Tabellen direkt mit HTML-Tags indizieren
 - ➖ **Contra:** LLM muss HTML-Code verstehen, mehr Kontext
 - Tabellen durch LLM textuell beschreiben
 - ➖ **Contra:** benötigt leistungsstarkes LLM
 - Tabellen in Markdown (z.B. mit pandas) konvertieren
 - ➕ **Pro:** hat bei uns Antwortqualität verbessert



Weitere Entscheidungsprobleme

- Welche Vektordatenbank verwenden wir?
 - ChromaDB, Qdrant, Weaviate, PgVector, ...
- Wie setzt man Chat-Dialog mit Historie um?
 - Historie in Prompt
 - Anfrage an Embedding-Service durch LLM formuliert oder Anfragen konkateniert
- Wie können wir beim Scraping nur geänderte Seiten in Wiki auslesen?
 - RSS-Feed
 - scheduled pipeline job



Weitere Entscheidungsprobleme

- Wie können wir die Quellen/Hyperlinks zu den Wikiseiten angeben?
 - Auflistung von Quellen aus Kontext oder LLM trifft Entscheidung
- Wie bauen wir ein Chat-Frontend?
 - existierendes Frontend (z.B. Streamlit) oder eigene Entwicklung (z.B. TypeScript + React)
- Wie kommuniziert das Frontend mit dem Backend, mit welcher API?
 - WebSockets oder REST
- ...

Fazit: Der Weg zum autarken KI-Chatbot: Ein Praxisbericht

- lokaler Betrieb erfordert **leistungsfähige Hardware**
- **Evaluation essenziell** für produktive KI-Anwendung
 - **Goldstandard** nötig: Fragen und erwartete Antworten für Chatbot
 - 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
 - Antwortqualität stark abhängig von Retrieval (answer F1 selten schlechter als context F1)
- **RAG-Optimierung** auf Basis der Evaluation, sollte bei Retrieval ansetzen
- Chatbot mit RAG **nur ein KI-Anwendungsfall von vielen**



generiert mit ChatGPT 4o



generiert mit Stablediffusion

Anhang

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 ^

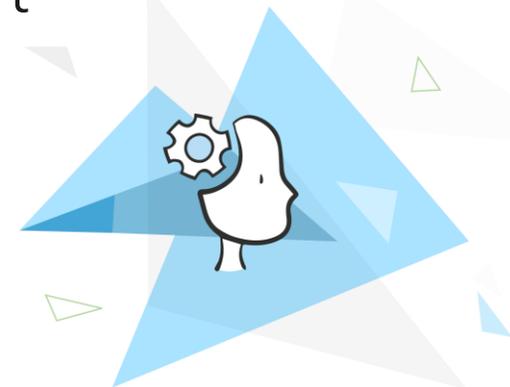
- <https://mail.andrena.de/wiki/Andrena/MERCURYII-SmartWiki>
- <https://mail.andrena.de/wiki/Mannheim/WebLeftBar>

Was ist SmartWiki?

Frage mich etwas!

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

- Chatbot beantwortet Nutzerfragen mit Daten aus unternehmensinternem Wiki
- Antworten werden mithilfe eines **Large Language Models (LLM)** generiert
- Wiki-Daten werden mittels **Retrieval Augmented Generation (RAG)** angereichert
- **Datensicherheit:** alle Komponenten lokal und selbst gehostet
- **Qualitätssicherung:** Antwortqualität wird systematisch evaluiert
- Nutzer kann die Antworten bewerten und **Feedback** geben
- Chatbot liefert **Quellenangaben**
- Nutzer kann Chatbot **konfigurieren**



Wie können Chatverläufe (nicht nur Einzelfragen) getestet werden?

Langfuse v2.60.4

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

Dataset Item

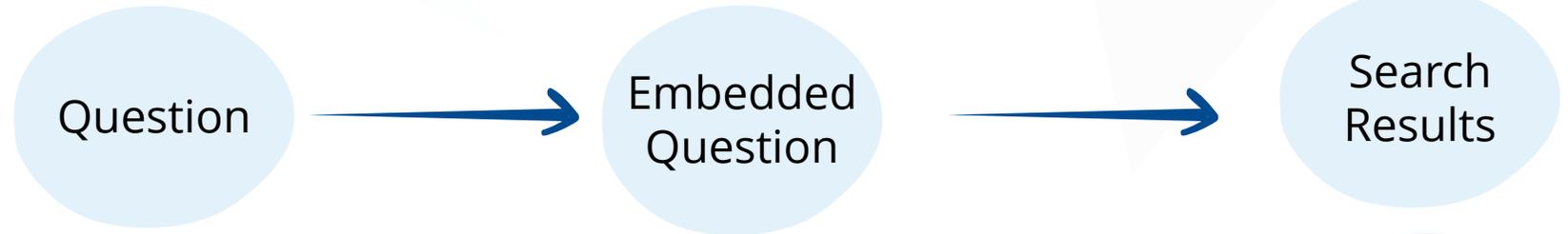
Input	Expected output
<pre> 1 { 2 "questions": [3 "Wie lautet die aktuelle Adresse des Karlsruher Büros?", 4 "Und die von Köln?" 5] 6 } </pre>	<pre> 1 { 2 "answer": "Die Adresse des Kölner Büros ist Josef-Lammerting-Allee 25, 50933 Köln.", 3 "exclude": [4 { 5 "variants": [6 "Widdersdorfer Straße 262, 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

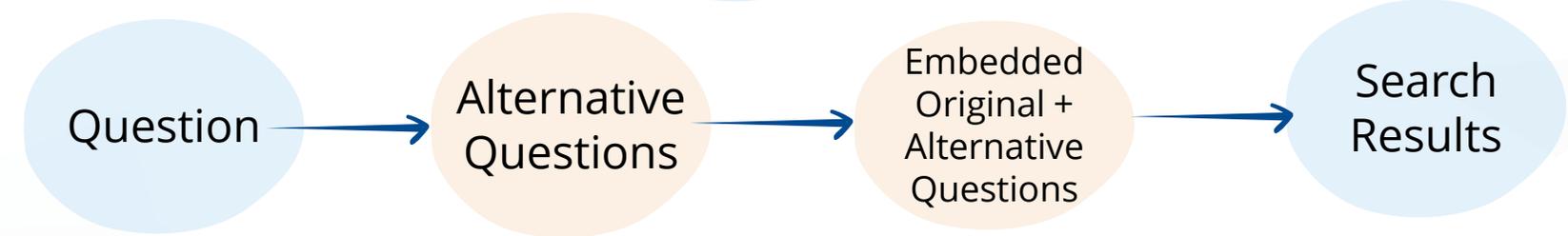
erwartete Antwort auf Frage 2

Advanced Retrieval Techniques

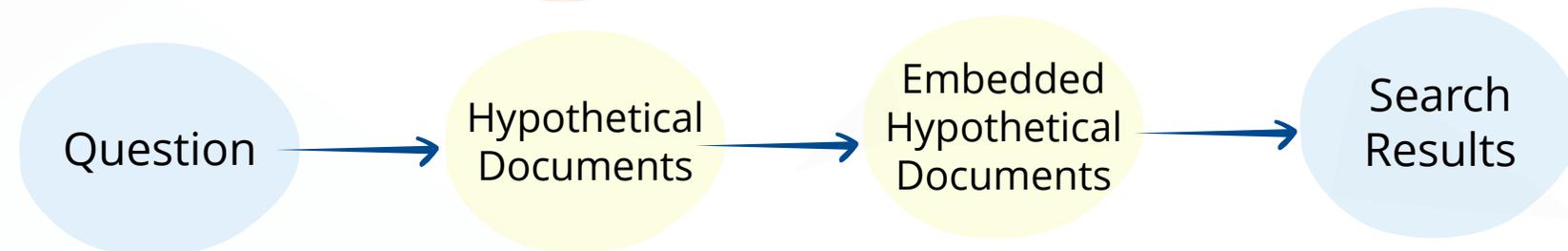
Basic Search



Query Expansion



HyDE



HyQE

