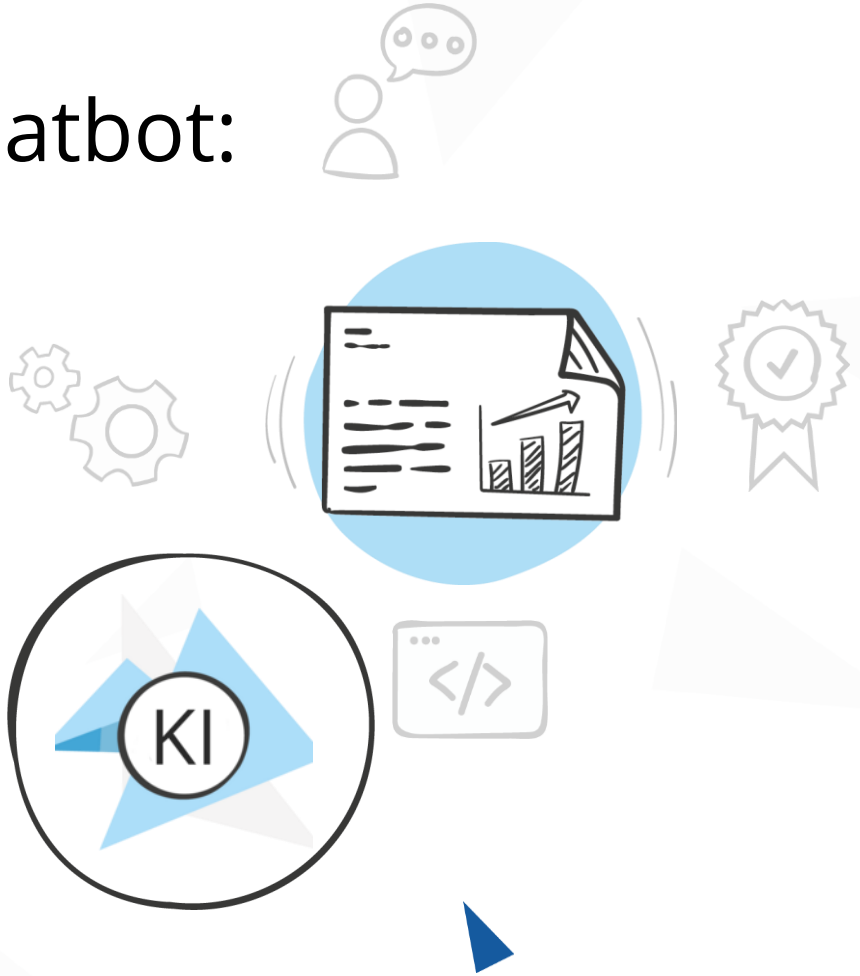
andrena
OBJECTS

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

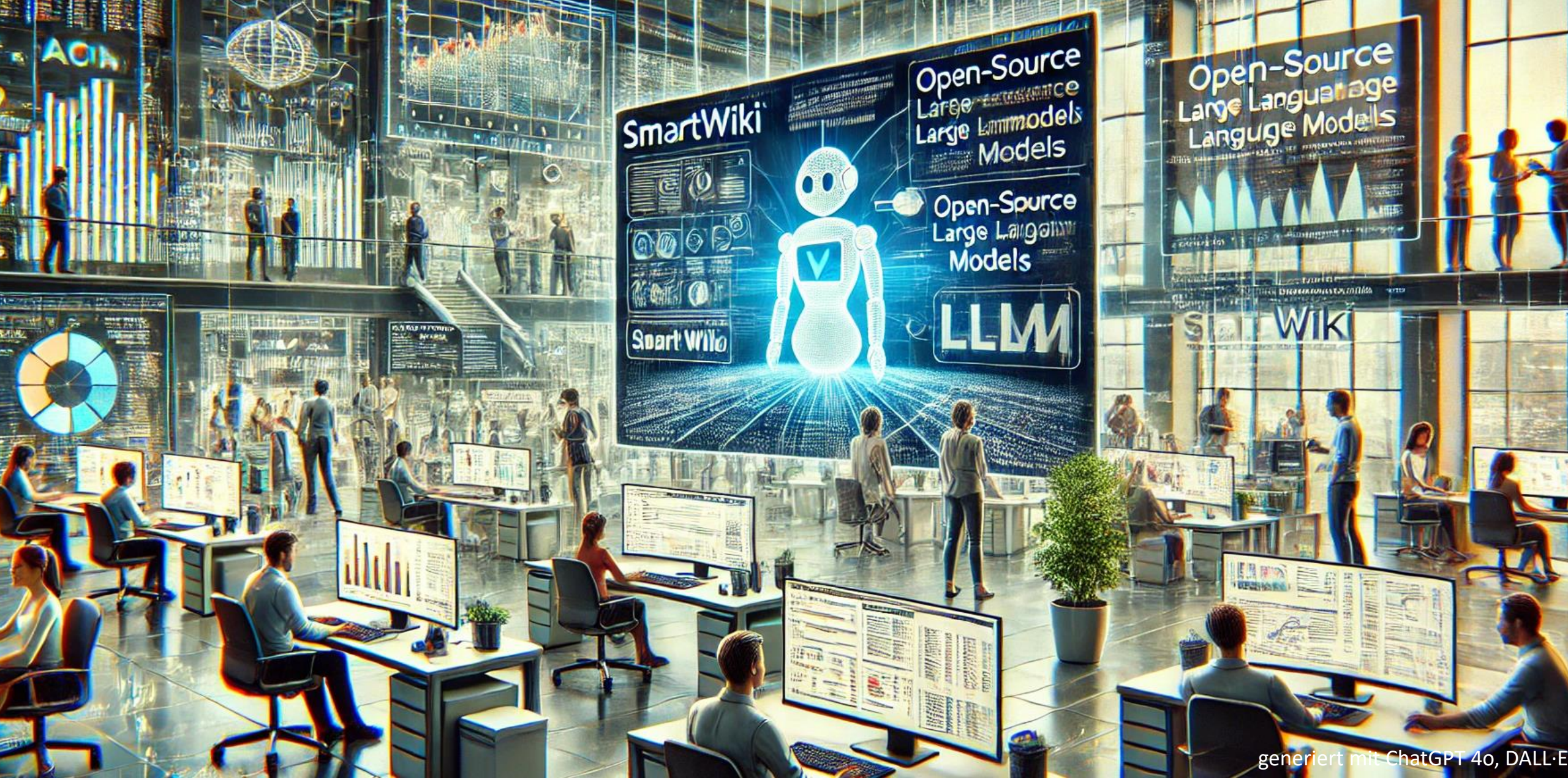
Wer sind wir?

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

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

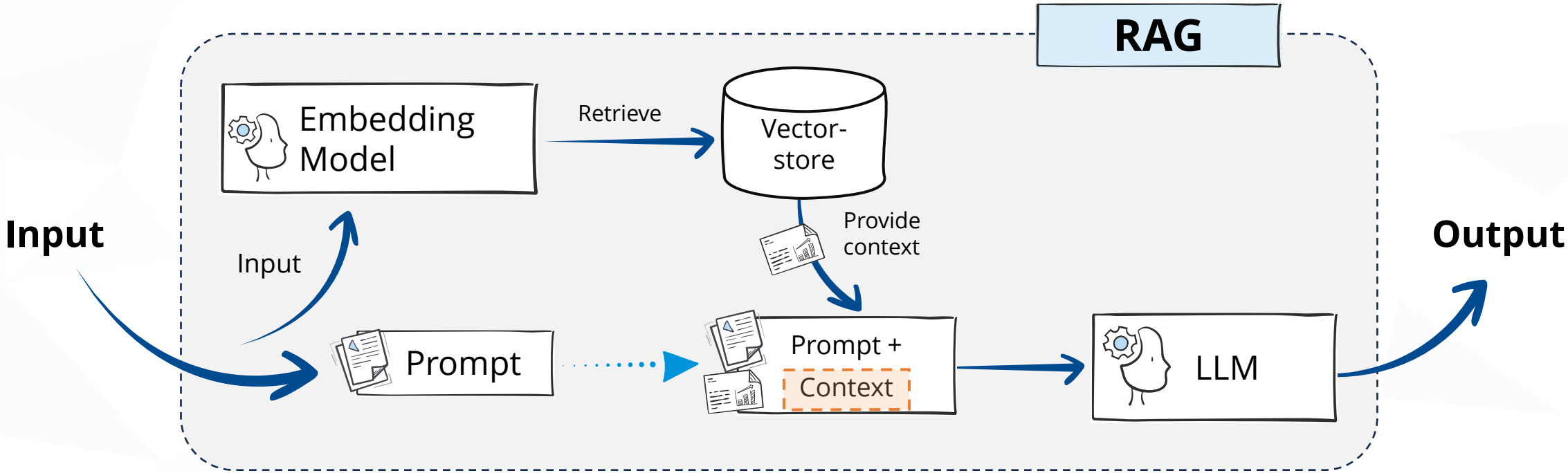


generiert mit ChatGPT 4o, DALL·E

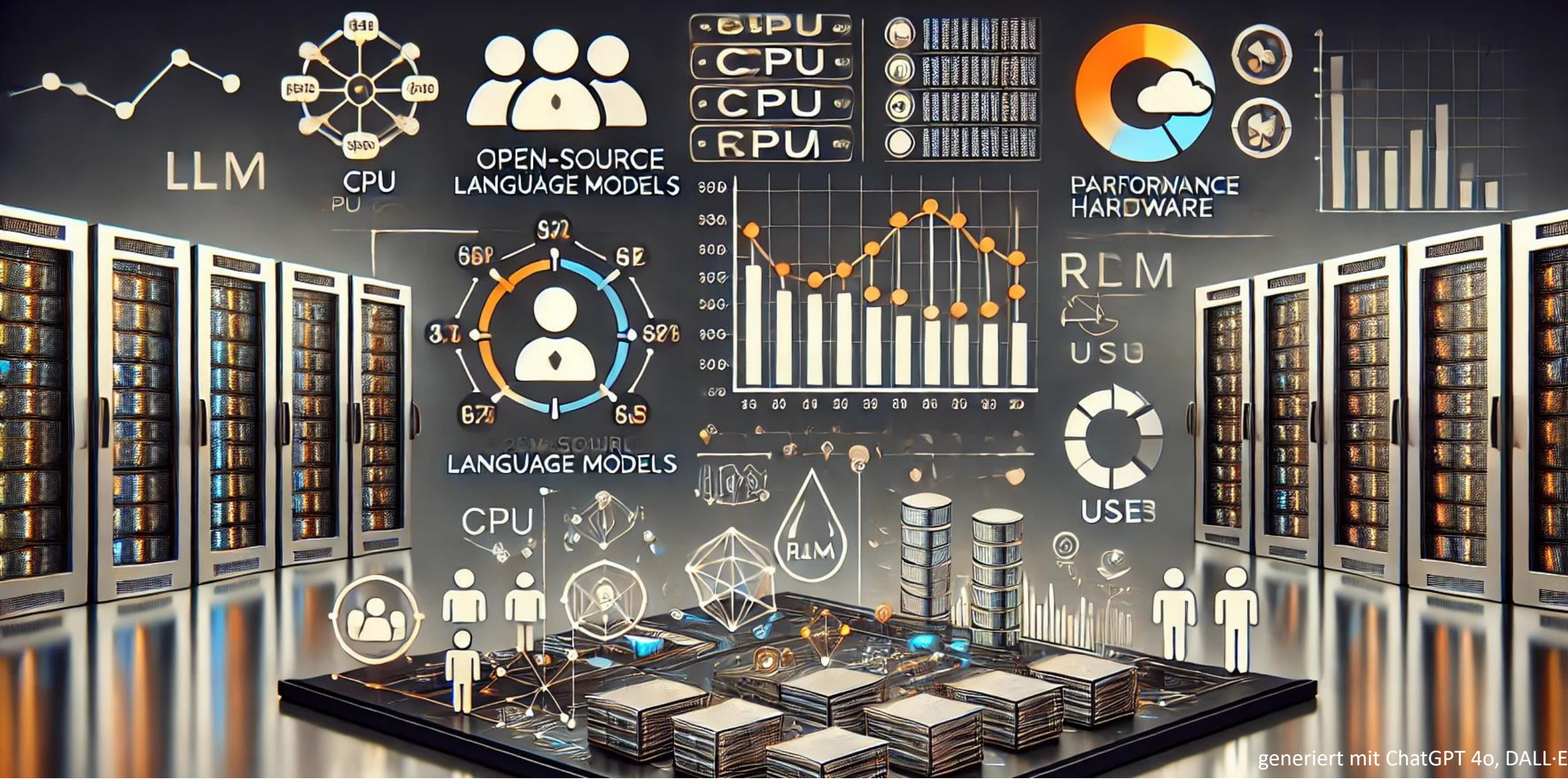
Vorstellung lokale RAG-Anwendung: KI-Chatbot „SmartWiki“

andrena OBJECTS

Retrieval Augmented Generation (RAG)



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



generiert mit ChatGPT 4o, DALL·E

Betrieb von Open-Source-LLMs

andrena OBJECTS

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

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

andrena OBJECTS

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

andrena OBJECTS

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

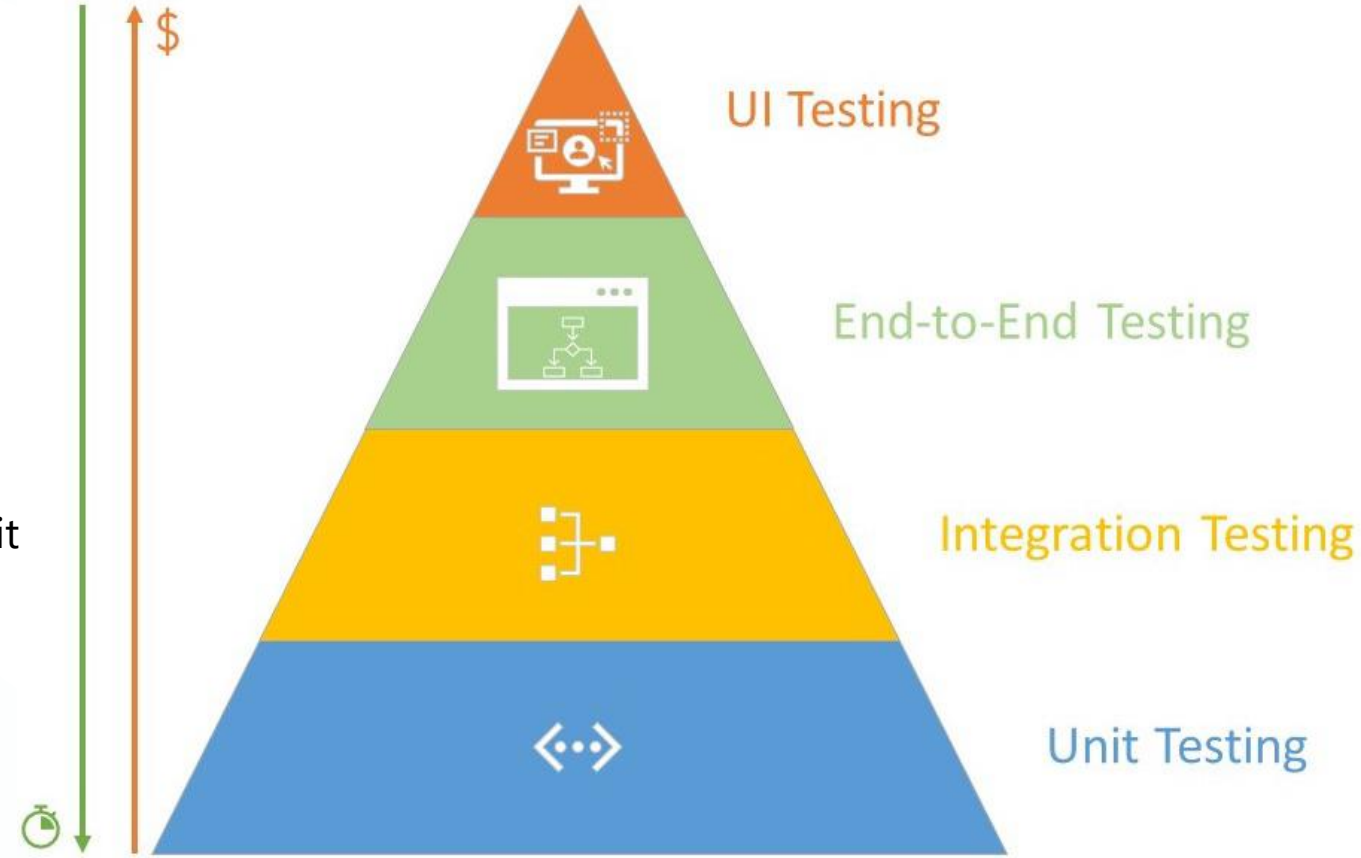
generiert mit ChatGPT 4o, DALL·E

Testing und Evaluation

andrena OBJECTS

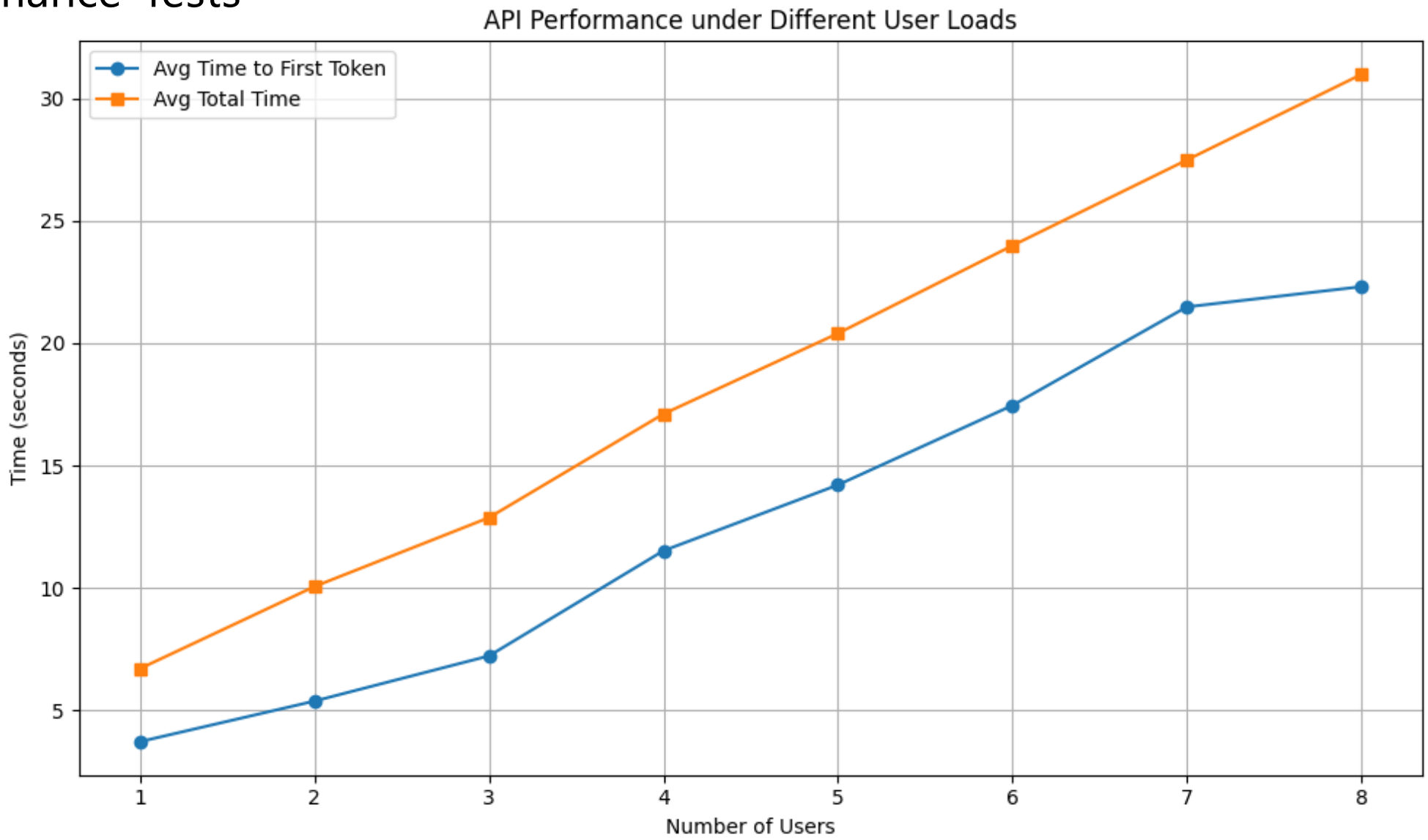
Übersicht Testarten für KI-Anwendungen

- 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



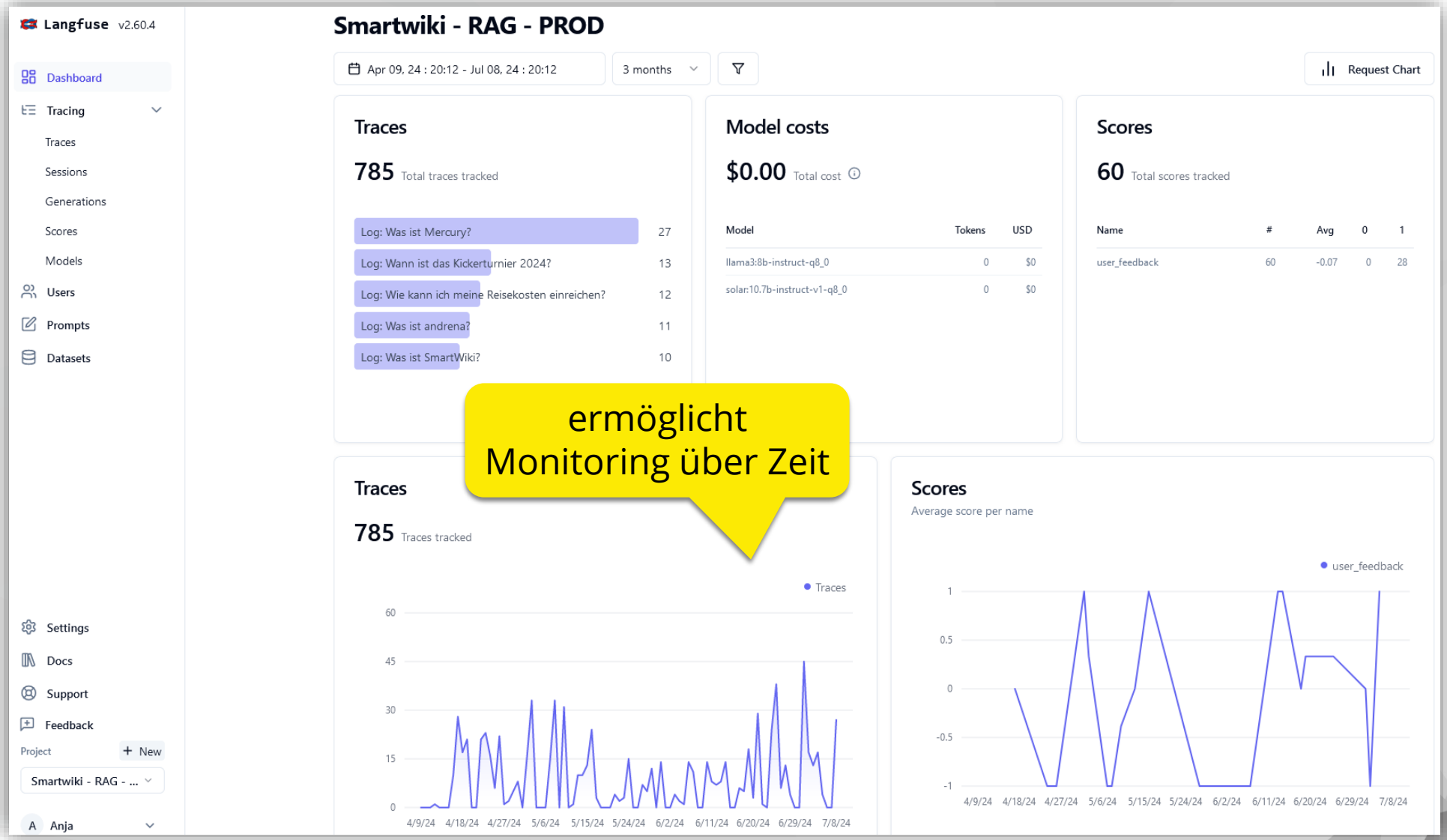
andrena OBJECTS

Performance-Tests



Welche Evaluationsplattformen gibt es?

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

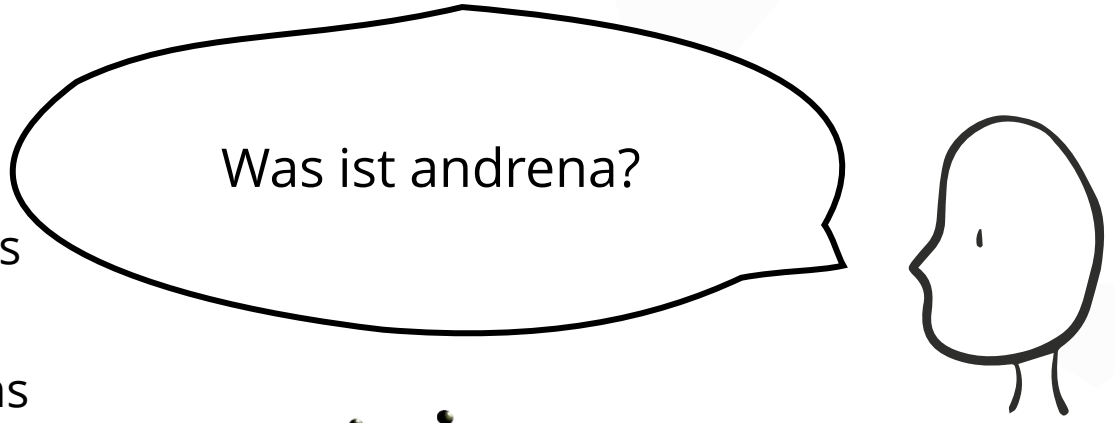


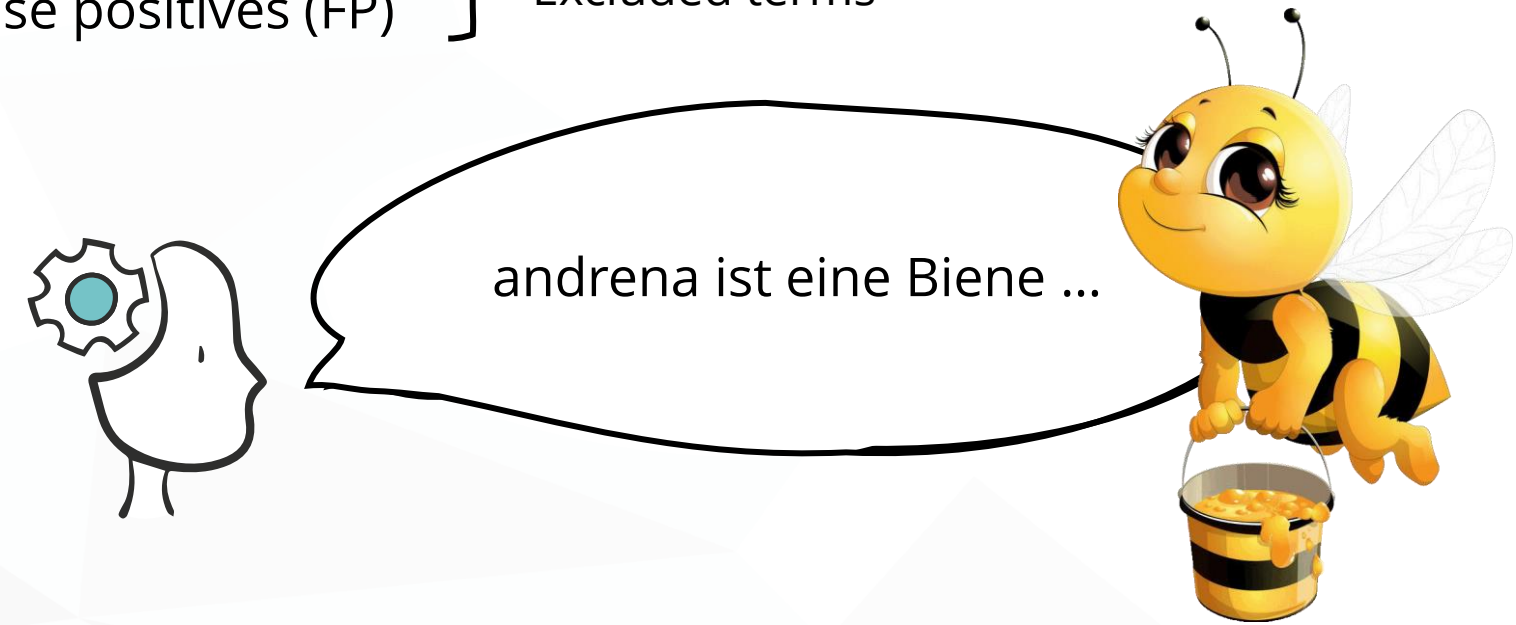
Wie baut man einen Evaluationsdatensatz (Goldstandard/ground truth) auf?

- Wichtige Begriffe
  - True positives (TP)
  - False negatives (FN)
  - False positives (FP)

Included terms

Excluded terms





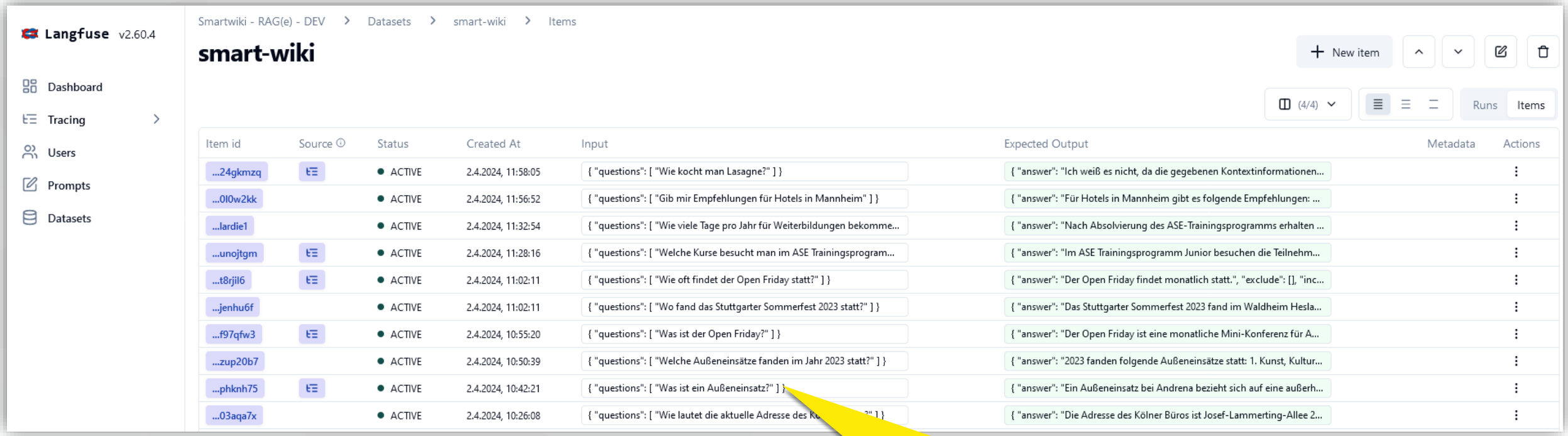
Wie baut man einen Evaluationsdatensatz (Goldstandard/ground truth) auf?

- **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

andrena OBJECTS

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

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



schauen wir uns genauer an...

andrena OBJECTS

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

Smartwiki - RAG(e) - DEV › Datasets › smart-wiki › Items › clui4s7f000ls6m367phknh75

Dataset Item

Input

```
{
  "questions": [
    "Was ist ein Außeneinsatz?"
  ]
}
```

Expected output

```
{
  "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.",
  "exclude": [
    {
      "variants": [
        "Girls Day"
      ]
    }
  ],
  "include": [
    {
      "variants": [
        "Veranstaltung",
        "Aktivität"
      ]
    },
    {
      "variants": [
        "Wandern",
        "Paddeln",
        "Malen",
        "Mountainbiken",
        "Survival-Training",
        "kulturelle Ausflüge"
      ]
    }
  ],
  "sources": [
    "https://mail.andrena.de/wiki/Andrena/ArchivBisherigerAusseneinsaetze"
```

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

Dataset Item

Input

```
{
  "questions": [
    "Was ist ein Außeneinsatz?"
  ]
}
```

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

Expected output

```
{
  "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?

```
  "include": [
    {
      "variants": [
        "Veranstaltung",
        "Aktivität"
      ]
    },
    {
      "variants": [
        "Wandern",
        "Paddeln",
        "Malen",
        "Mountainbiken",
        "Survival-Training",
        "kulturelle Ausflüge"
      ]
    }
  ],
```

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

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

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

andrena OBJECTS

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)

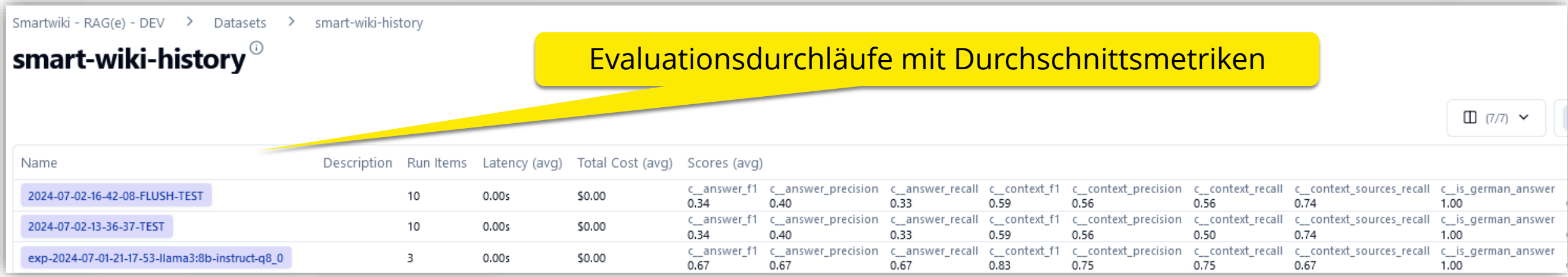
generiert mit ChatGPT 4o

andrena OBJECTS

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



- 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, ...

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

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

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

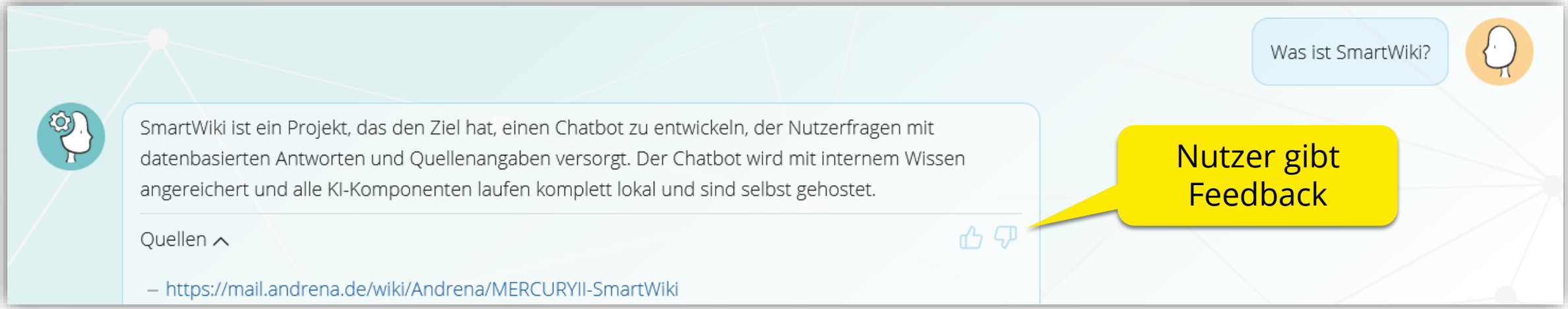
- **Metriken für Generation:**
  - answer recall, answer precision, answer F1
  - RAGAS: answer similarity, answer correctness, faithfulness, answer relevancy
  - friendliness, harmfulness, conciseness, ...

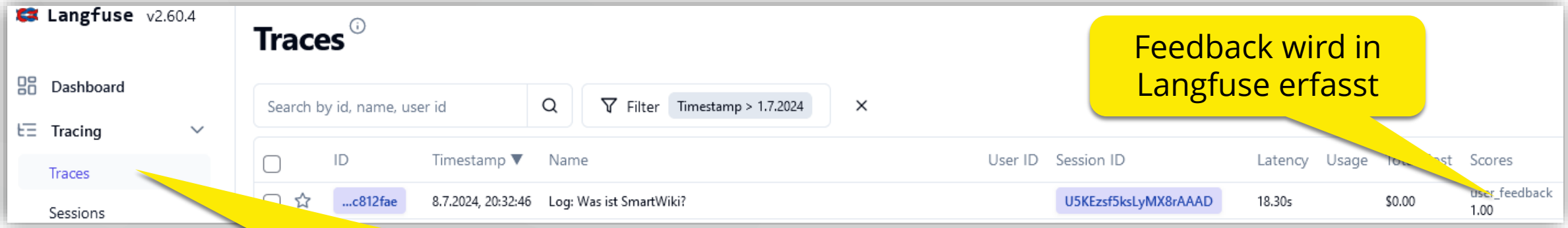
generiert mit ChatGPT 4o, DALL·E

Nutzerfeedback

andrena OBJECTS

Wie kann man Nutzerfeedback sammeln und zur Verbesserung verwenden?





Langfuse zeigt Fragen und Antworten des Chatbots (anonymisiert)

generiert mit ChatGPT 4o, DALL·E

RAG-Optimierung

andrena OBJECTS

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

andrena OBJECTS

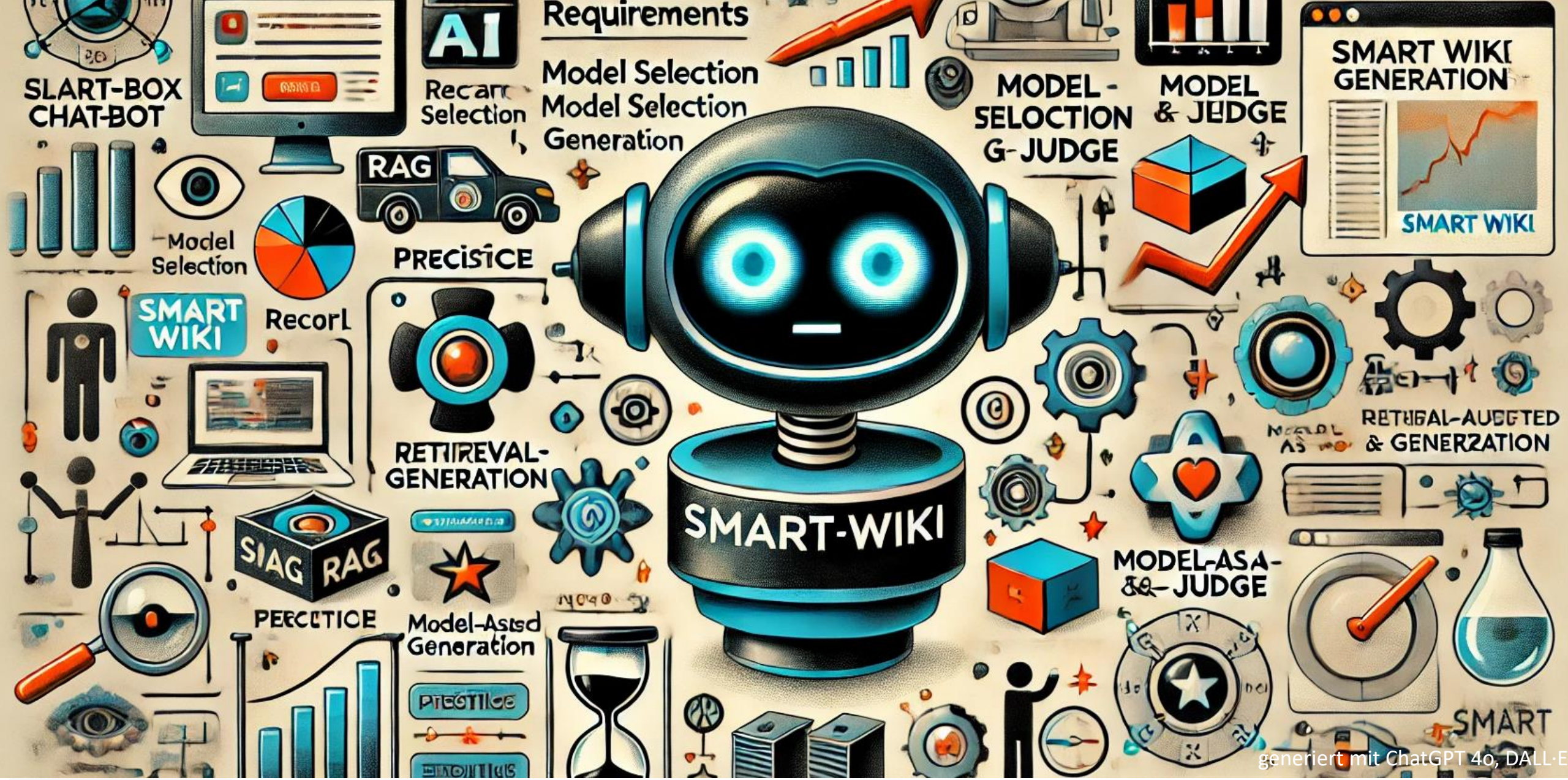
Weitere Entscheidungsprobleme

- 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

andrena OBJECTS

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
- ...

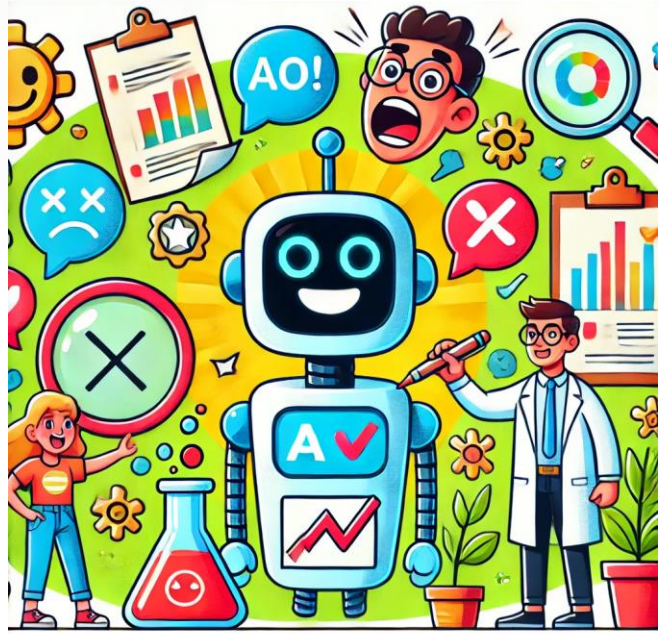


generiert mit ChatGPT 4o, DALL·E

Fazit

andrena OBJECTS

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

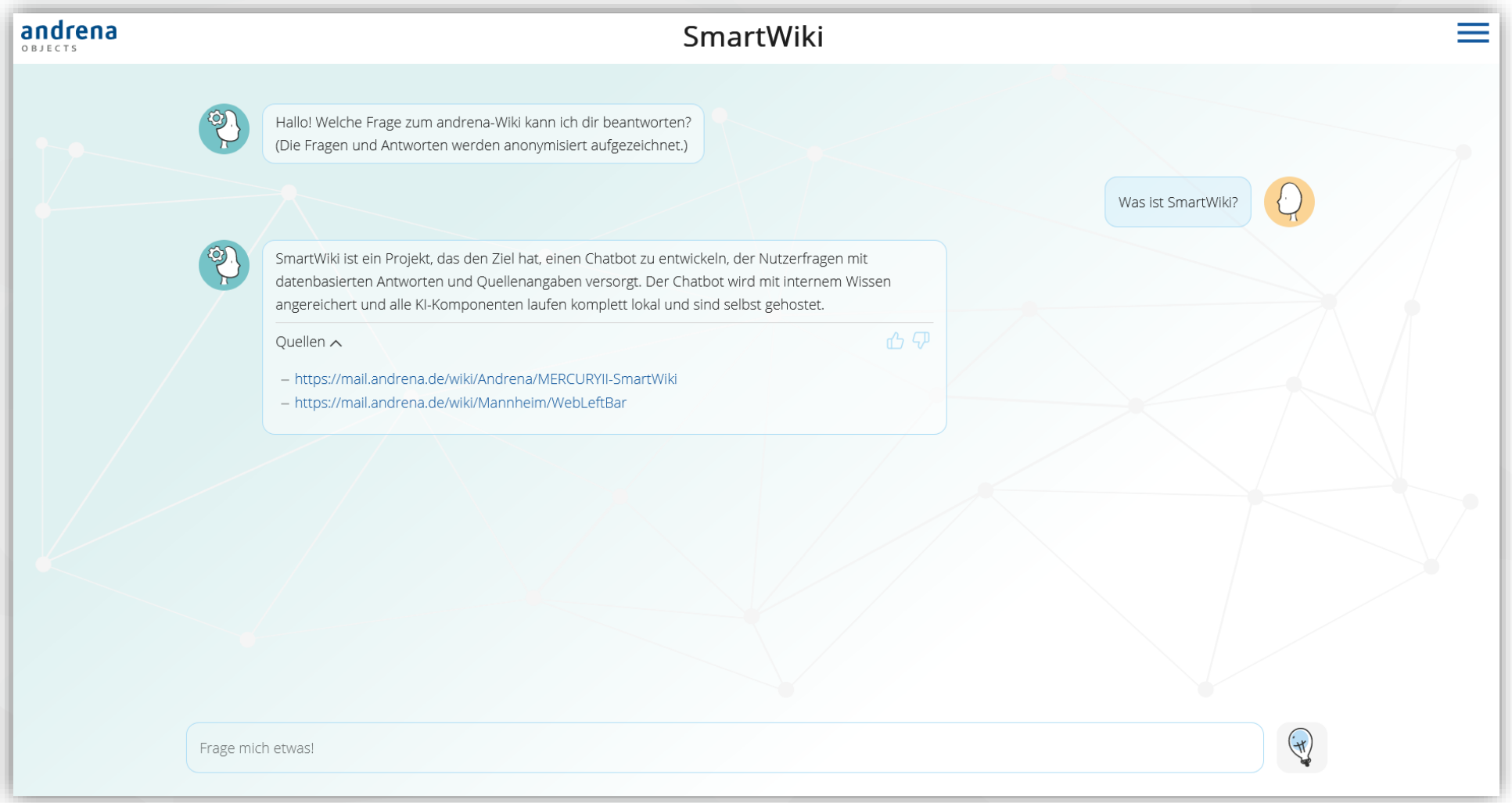


generiert mit ChatGPT 4o

- 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**

Anhang

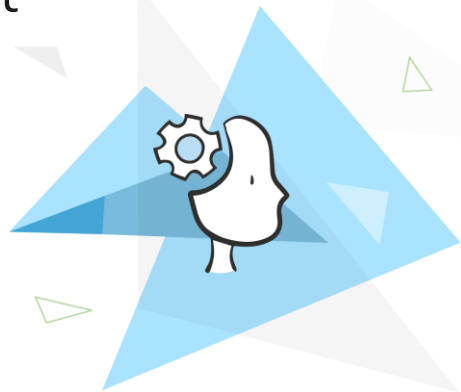
Unsere lokale RAG-Anwendung: KI-Chatbot „SmartWiki“



andrena OBJECTS

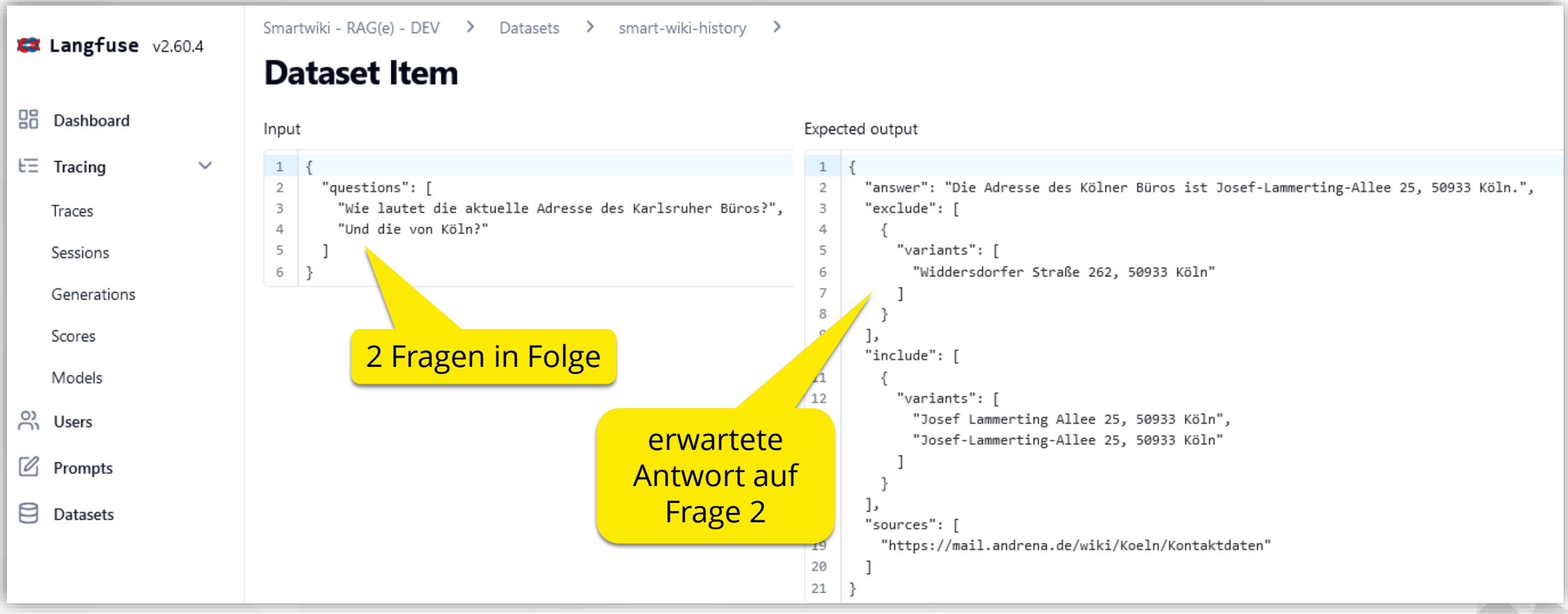
Unsere lokale RAG-Anwendung: KI-Chatbot „SmartWiki“

- 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**



andrena OBJECTS

Wie können Chatverläufe (nicht nur Einzelfragen) getestet werden?



andrena OBJECTS

Advanced Retrieval Techniques

- Basic Search
- Query Expansion
- HyDE
- HyQE