Staatsbibliothek
zu Berlin
Preußischer Kulturbesitz

Staatsbibliothek zu Berlin

KI & RAG – mit einem Chatbot eigene Dokumentensammlungen befragen

Felix Hagenström

15.10.2026

Überblick

(1) Einführung: **Generative KI // LLMs // RAG**

(2) Vorstellung **Chat AI & Arcana**

- Allgemeines
- Wichtige Funktionen

(3) Anwendungsszenarien

- Befragen von **Brieftranskriptionen** aus der Sammlung der Stabi
- Anwendung einer **soziologischen Theorie** auf ein **Praxisbeispiel**

(4) Hands-on: Dokumente hochladen und befragen

(5) Austausch, Fragen, Abschluss

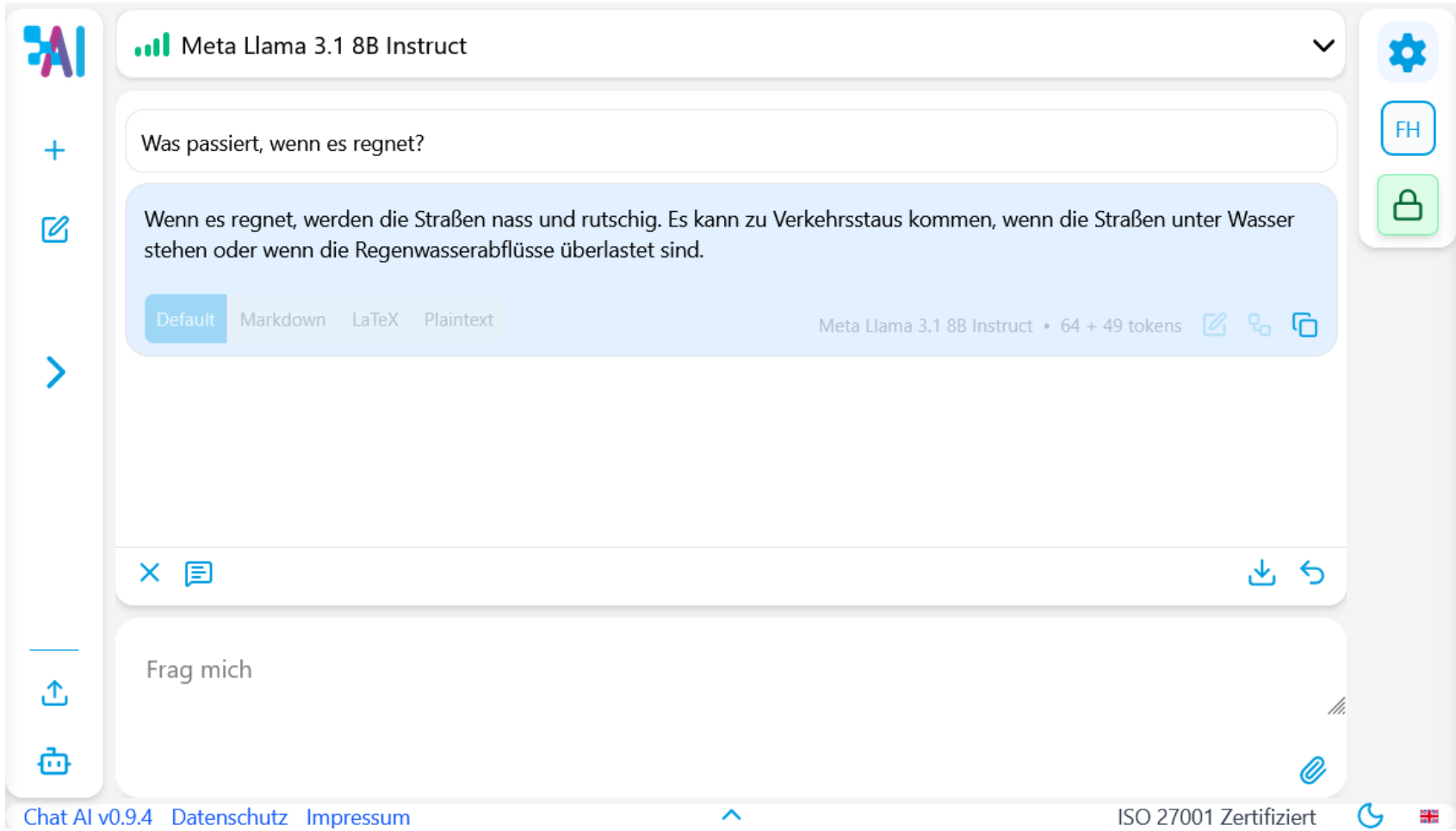
Generative KI: Anwendungen, die aufgrund maschinellen Lernens neue Inhalte erstellen können (z.B. Texte, Bilder, Audio, Video)

Large Language Models (LLMs)/Große Sprachmodelle: auf riesigen Datenmengen trainierte Modelle, die natürlichsprachliche Eingaben verarbeiten und neue Texte generieren können

Beispiele kommerzieller GenAI-Anwendungen: ChatGPT, Claude, Copilot, Gemini, Mistral Vibe, Perplexity etc.

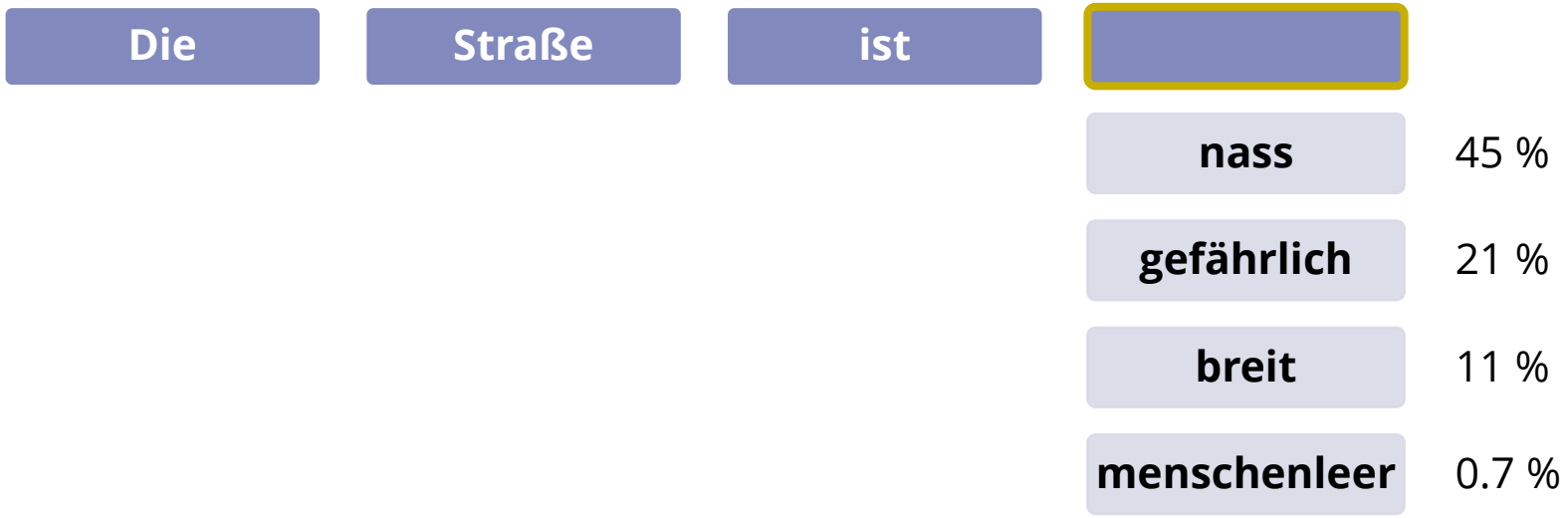
- Diese Anwendungen nutzen verschiedene LLMs
- In der Praxis: GenAI/GenKI und LLM werden häufig synonym gebraucht

Beispiel KI-Chatbot Chat AI



Die Funktionsweise von LLMs

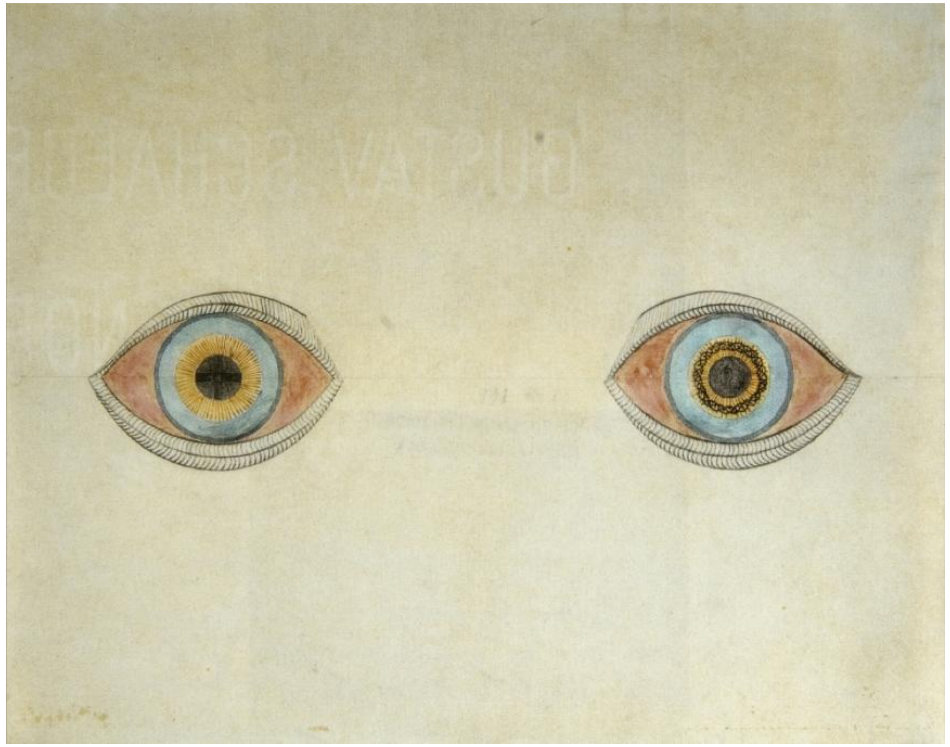
- LLMs sind *keine Wissensmodelle*
  - Outputs von LLMs beruhen auf *Wahrscheinlichkeiten*
  - Text wird nach statistischen Kriterien und bestimmten Parametern erstellt



Die Funktionsweise von LLMs

- LLMs sind *keine Wissensmodelle*
  - Outputs von LLMs beruhen auf *Wahrscheinlichkeiten*
  - Text wird nach statistischen Kriterien und bestimmten Parametern erstellt

Halluzinationen (1)



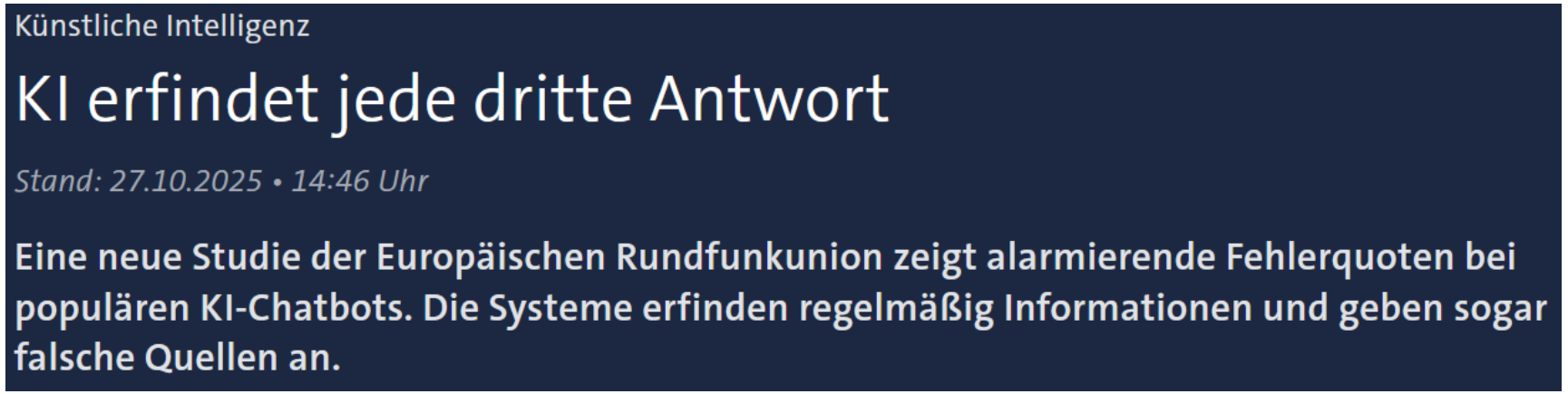
By August Natterer, Public Domain, https://commons.wikimedia.org/w/index.php?curid=12756650

- Mensch: Wahrnehmung von etwas, aber *ohne externe Stimuli*
- LLMs: Überzeugend formulierter, aber *falscher Output*

Alternativer Begriff zu Halluzination:
„Bullshit“ (Harry Frankfurt) https://stabikat.de/Record/1616072253

- Indifferenz gegenüber Wahrheit
- Kein Anspruch der Wirklichkeitsrepräsentation

Halluzinationen (2)



Künstliche Intelligenz

KI erfindet jede dritte Antwort

Stand: 27.10.2025 • 14:46 Uhr

Eine neue Studie der Europäischen Rundfunkunion zeigt alarmierende Fehlerquoten bei populären KI-Chatbots. Die Systeme erfinden regelmäßig Informationen und geben sogar falsche Quellen an.

https://www.tagesschau.de/wissen/technologie/kuenstliche-intelligenz-fakten-100.html

STUDIE

Zahl falscher KI-Referenzen in Forschungsarbeiten nimmt zu

Falsche Zitate und Quellen in wissenschaftlichen Publikationen nehmen zu. Das deutet darauf hin, dass Prüfmechanismen unzureichend funktionieren.

24.06.2026
https://www.forschung-und-lehre.de/forschung/zahl-falscher-ki-referenzen-in-forschungsarbeiten-nimmt-zu-7760

- RAG kann Halluzinationen nicht verhindern, das Risiko jedoch abmildern
  - ➢*Aber wie?*

RAG steht für **Retrieval-Augmented Generation** (etwa: *per Abfrage erweiterte Generierung*)

- RAG-Systeme stellen LLMs **externe Quellen zur Verfügung** (z.B. lokale Dokumente/Daten oder Zugriff aufs Web)
- **Daten** werden in **Chunks zerlegt, indexiert** und in einer **Vektordatenbank verknüpft**

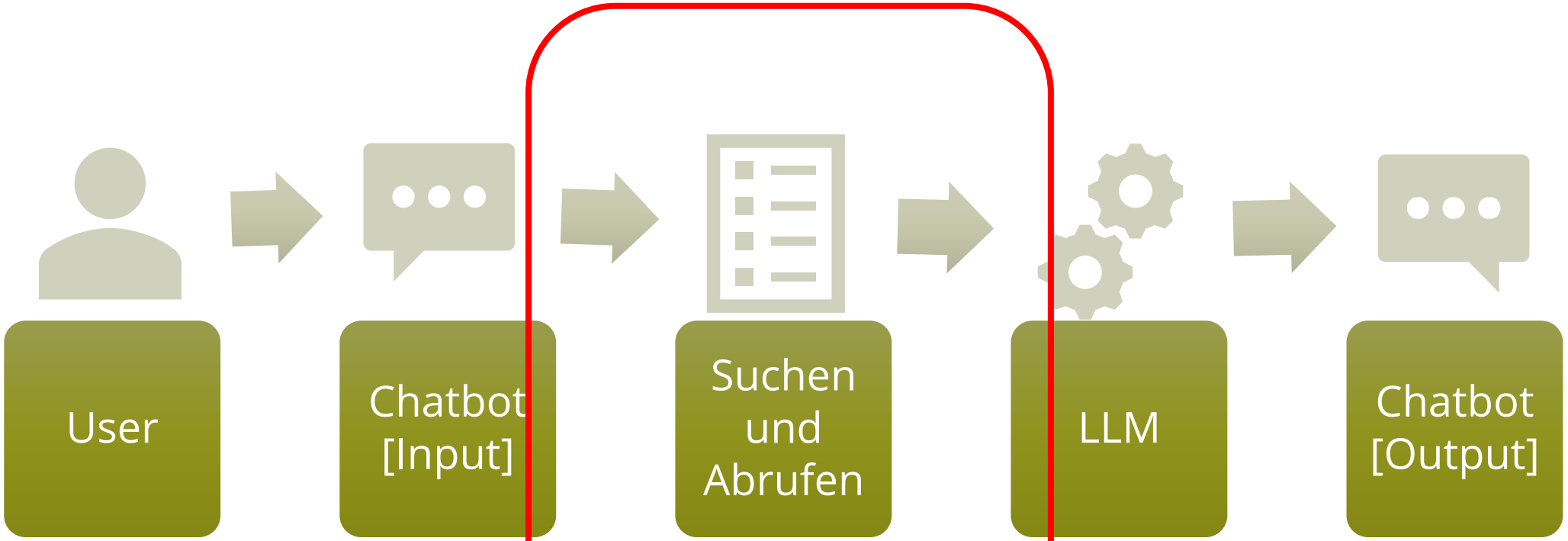
Warum nutzt man RAG?

LLMs haben große Schwachstellen:

- **Halluzinationen:** Mangel an Faktentreue lässt ein LLM plausibel klingende, aber faktisch falsche Antwort erfinden
- **Wissensstopp (Knowledge Cut-off):** Die Trainingsdaten reichen nur bis zu einem bestimmten Datum; aktuelle Ereignisse sind unbekannt

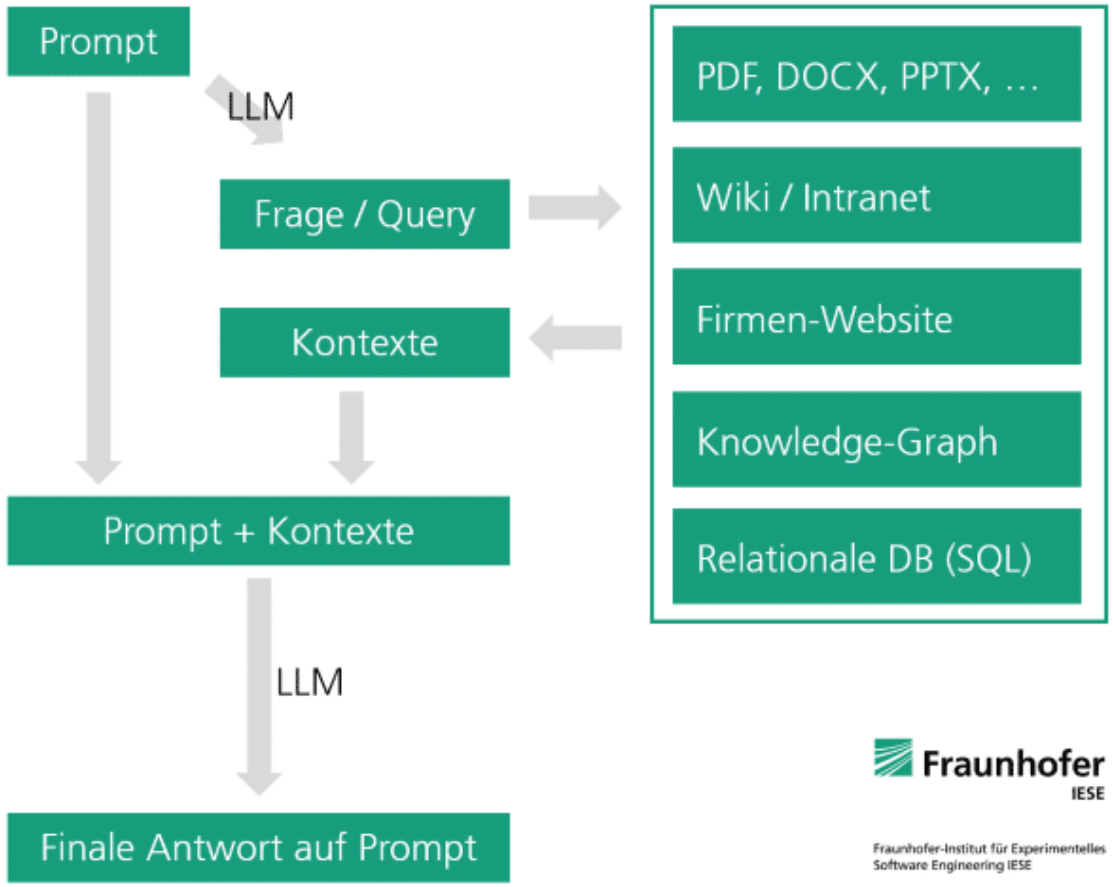
LLM & RAG: Funktionsweise

LLM & RAG: Funktionsweise



Vorteile von RAG
Aktualität
Faktentreue
Transparenz/Zitierfähigkeit
Datenschutz

LLM & RAG: Funktionsweise



https://www.iese.fraunhofer.de/blog/wp-content/uploads/2024/05/retrieval_augmented_generation_RAG_600px.png

Chat AI – Infrastruktur & Anbieter

- Service von KISSKI/GWDG (Göttingen)
- KISSKI: Eines von vier KI-Servicezentren in Deutschland (Förderung seit 2022 vom Bund)

Zugang über Academic Cloud

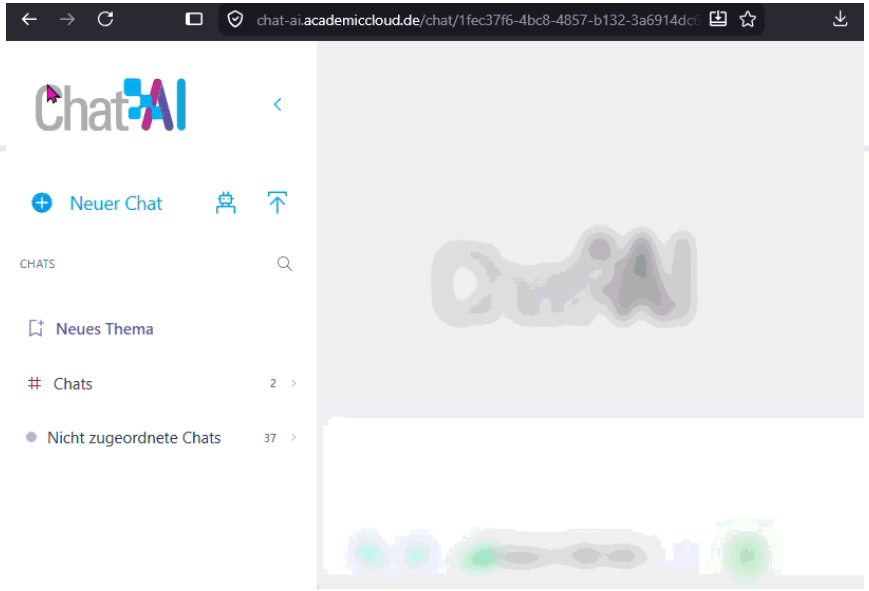
Bildquelle: https://kisski.gwdg.de/ueber-uns/konsortium/

Chat AI – Allgemeine Charakteristika

- Teil der **KI-Infrastruktur für Wissenschaft und Forschung** in Deutschland
- Zugang über **Academic Cloud**
- **Datenschutz** bei Sprachmodellen gewährleistet (auf lokalen Servern gehostet); Ausnahme: eigene Institution bietet Zugriff auf weitere Modelle
- **Daten werden nur im Browser (lokal) gespeichert**
- **Auswahl verschiedener Modellen** (https://docs.hpc.gwdg.de/services/chat-ai/models/index.html)
- Größtmögliche **Transparenz** und **Steuerungsmöglichkeit**
- Nutzung im **Webbrowser** oder via **API**
- Zugriff über die **eigene Institution** oder einen **kostenfreien Einzelaccount**

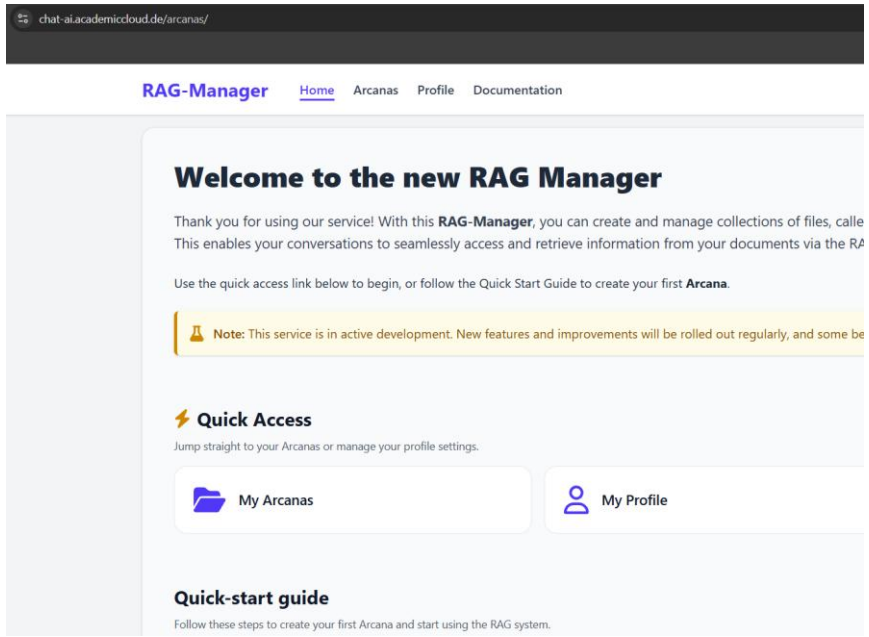
Chat AI – RAG (Arcanas)

- Chat AI bietet einen **RAG-Manager**
- Dokumentensammlungen heißen bei Chat AI: **Arcanas**



Kurzanleitung

- Auf der Chat-AI-Benutzeroberfläche (Browser) oben links auf das Logo klicken und ‚RAG-Manager' auswählen
- ‚Arcanas' oder ‚My Arcanas' anklicken – durch den Prozess leiten lassen + indexieren
- Dateien
  - Unterschiedliche Formate möglich (PDF, DOCX, CSV, TXT, XML, MD)
  - Max. Größe 300 MB



Schauen wir mal rein…

https://chat-ai.academiccloud.de/

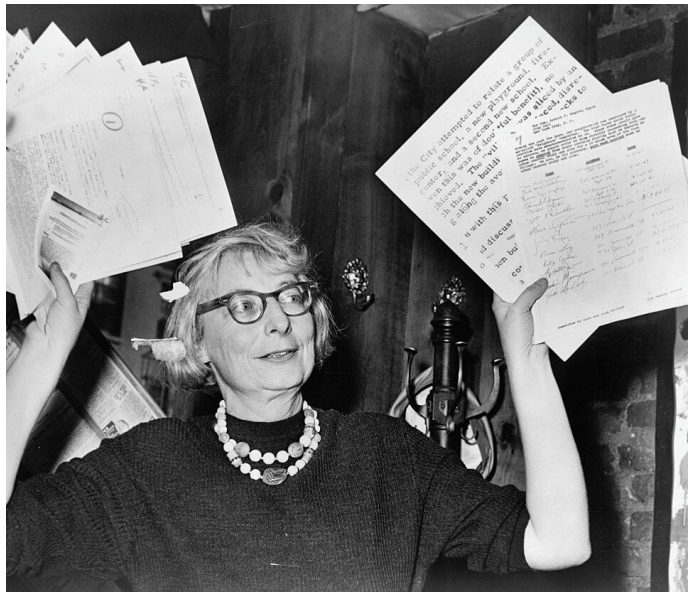
Anwendungsbeispiele

- **Beispiel 1: Digitalisierte Briefe von und an Julie Elias**
  Szenario: Wir arbeiten zum Werk von Julie Elias und wollen ihre Briefwechsel für unsere Studie auswerten. Dazu wollen wir ausgewählte Transkriptionen aus der Sammlung der Stabi befragen.

❖ *Dokumente*

  – Briefe von Julie Elias (PDF) (liegen auch in anderen Formaten vor)

https://commons.wikimedia.org/w/index.php?curid=63504233



https://upload.wikimedia.org/wikipedia/commons/1/14/Jane_Jacobs.jpg

- **Beispiel 2: Jane Jacobs' Theorie der ‚Eyes on the Street'**
  Szenario: Wir forschen im Bereich der Stadtsoziologie und wollen Jacobs' Theorie auf empirische Fälle in Berlin anwenden. Es geht um die Frage, welche Straßen als sicher gelten.

❖ *Dokumente:*

  - Zusammenfassung der Theorie: *Urban-Vitality_eyes-on-the-street_Natual-surveillance_Jane-Jacobs.md*
  - Liste mit Berliner Straßen für die Anwendung (.txt)

Einige Grenzen von RAG

- **Datenqualität** beeinflusst Zuverlässigkeit
  - Benennung und Strukturierung der Dokumente wichtig
- **Keine Garantie** für völlige Halluzinationsfreiheit
- **Keine übergreifenden Erkenntnisse** aus komplexer Dokumentensammlung

➢ RAG erfordert einige Arbeit

➢ Kritische Überprüfung immer notwendig

Probieren Sie es aus! Dokumentensammlung erstellen

Staatsbibliothek zu Berlin
Preußischer Kulturbesitz

- Chat AI aufrufen und **neues Arcana anlegen**
- **Eigene Dokumente** oder **bereitgestellte Dateien** hochladen
  - Weitere **Beispieldokumente**:
    - Briefe von Julie Elias (PDF)
    - Auszug aus Till Eulenspiegel (DOCX)
    - *Die fröhliche Wissenschaft* (F. Nietzsche) (TXT)
      - ➢ Für die Beispieldateien bei Webex unten rechts auf die **drei Punkte** und dann auf **„Inhalt“/„Content“** klicken
- **Indexieren** Sie die Dokumente
- Kopieren Sie die **Arcana-ID** in die Chat-AI-Einstellungen
- Evtl. temp & top_p und/oder Systemprompt anpassen

Mögliche Fragen für die Beispieldokumente

Julie Elias:

- An wen schreibt Julie Elias?
- Auf wann sind die Briefe Julie Elias' datiert?
- Geht es in einem der Briefe um Alkohol?

Jane Jacobs:

- Auf Grundlage von Jane Jacobs' Theorie der 'Eyes on the Street' zieh eine Schlussfolgerung: Welche der aufgelisteten Straßen ist höchstwahrscheinlich die sicherste und warum?

Till Eulenspiegel:

- In welche Orte/Städte/Dörfer reist Till Eulenspiegel?

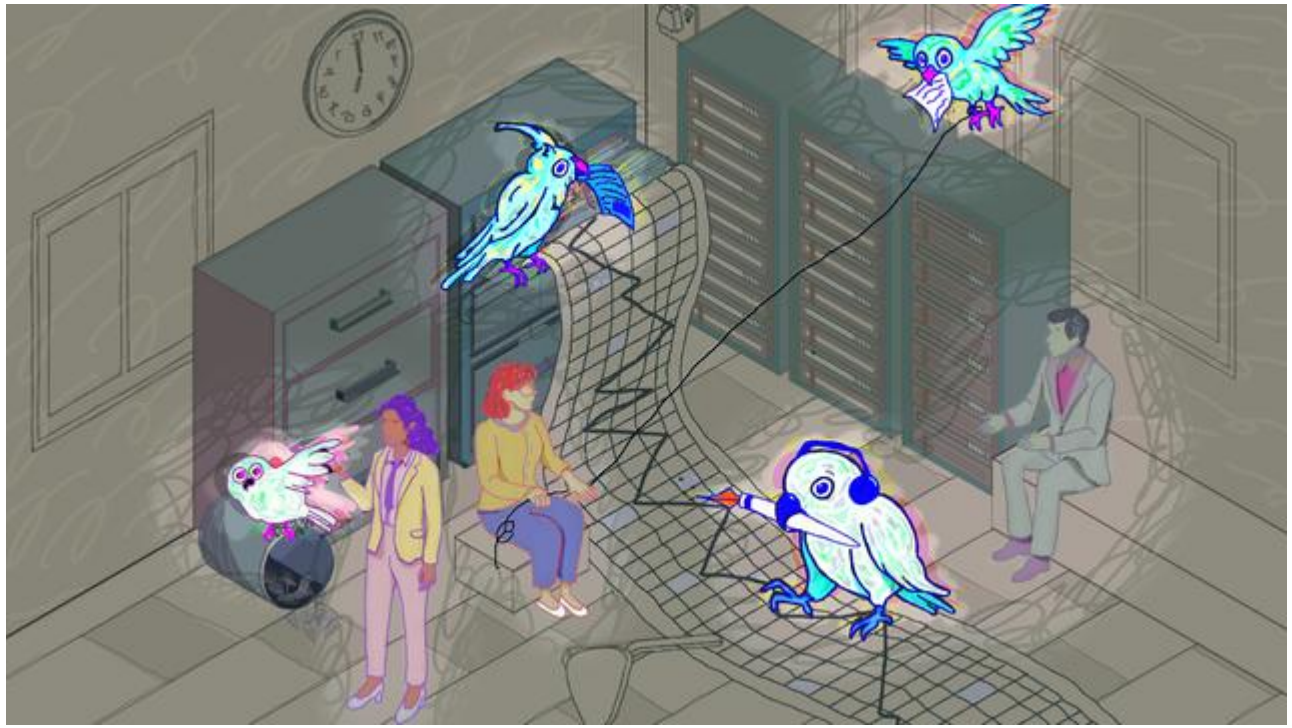
Friedrich Nietzsche:

- Was ist Nietzsches erklärte Absicht hinter der Schrift „Die fröhliche Wissenschaft“?

Praxisphase – Eigene Erfahrungen

1. Was haben Sie ausprobiert?
2. Was gut geklappt?
3. Was war frustrierend?
4. Was würden Sie in Zukunft weiter nutzen?

Dankeschön



IceMing & Digit / https://betterimagesofai.org / https://creativecommons.org/licenses/by/4.0/

Fragen & Austausch

Vorschau – Stabi-Veranstaltungen zu KI:

- 4.11., 14:30 Uhr, Online-Workshop: *KI und Literaturrecherche*
- 4.11., 14 Uhr, Online-Workshop Stabi Lab: *Cards to Catalogue – Sharing Experiences in LLM-Based Card Catalogue Conversion*
- 18.11., 10 Uhr, Hands-On-Workshop: *Plausibel, aber wahr? Quellenkritik im Zeitalter von KI*

Alle Stabi-Veranstaltungen:
https://blog.sbb.berlin/termine/#veranstaltungen

felix.hagenstroem@sbb.spk-berlin.de

Links

Halluzinationen

https://www.iese.fraunhofer.de/blog/halluzinationen-generative-ki-llm/

RAG

https://www.iese.fraunhofer.de/blog/retrieval-augmented-generation-rag/

Chat AI | Academic Cloud | Arcana

https://chat-ai.academiccloud.de/ | https://academiccloud.de/ | https://chat-ai.academiccloud.de/arcanas/ | **Anleitung Arcana:** https://docs.hpc.gwdg.de/services/ai-services/arcana/how-to-use/index.html

Stabi Lab

https://lab.sbb.berlin/ | https://lab.sbb.berlin/datenset-transkriptionen-julie-elias/