



Staatsbibliothek
zu Berlin
Preußischer Kulturbesitz

Staatsbibliothek zu Berlin

KI & RAG – mit einem Chatbot eigene Dokumentensammlungen befragen

Felix Hagenström

15.07.2026

- Einführung: **Generative KI // LLMs // RAG**
- Vorstellung **Chat AI & Arcana**
 - Allgemeines
 - Einige wichtige Funktionen
- **Anwendungsbeispiele**
- **Hands-on:** Dokumente hochladen und befragen
- **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 werden GenAI/GenKI und LLMs häufig synonym gebraucht

Beispiel KI-Chatbot Chat AI

The screenshot displays the Chat AI interface. At the top, the model is identified as "Meta Llama 3.1 8B Instruct". The user's input is "Was passiert, wenn es regnet?". The AI's response is "Wenn es regnet, werden die Straßen nass und rutschig. Es kann zu Verkehrsstaus kommen, wenn die Straßen unter Wasser stehen oder wenn die Regenwasserabflüsse überlastet sind." Below the response, there are tabs for "Default", "Markdown", "LaTeX", and "Plaintext". The token count "Meta Llama 3.1 8B Instruct • 64 + 49 tokens" is shown. The interface includes a sidebar with icons for adding, editing, and navigating, and a bottom bar with links to "Chat AI v0.9.4", "Datenschutz", and "Impressum", along with an "ISO 27001 Zertifiziert" badge and a language selector.

Meta Llama 3.1 8B Instruct

Was passiert, wenn es regnet?

Wenn es regnet, werden die Straßen nass und rutschig. Es kann zu Verkehrsstaus kommen, wenn die Straßen unter Wasser stehen oder wenn die Regenwasserabflüsse überlastet sind.

Default Markdown LaTeX Plaintext

Meta Llama 3.1 8B Instruct • 64 + 49 tokens

Frag mich

Chat AI v0.9.4 Datenschutz Impressum

ISO 27001 Zertifiziert

Die Funktionsweise von LLMs

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

Die	Straße	ist		
			nass	45 %
			gefährlich	21 %
			breit	11 %
			menschenleer	0.7 %

Die Funktionsweise von LLMs

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

Die

Straße

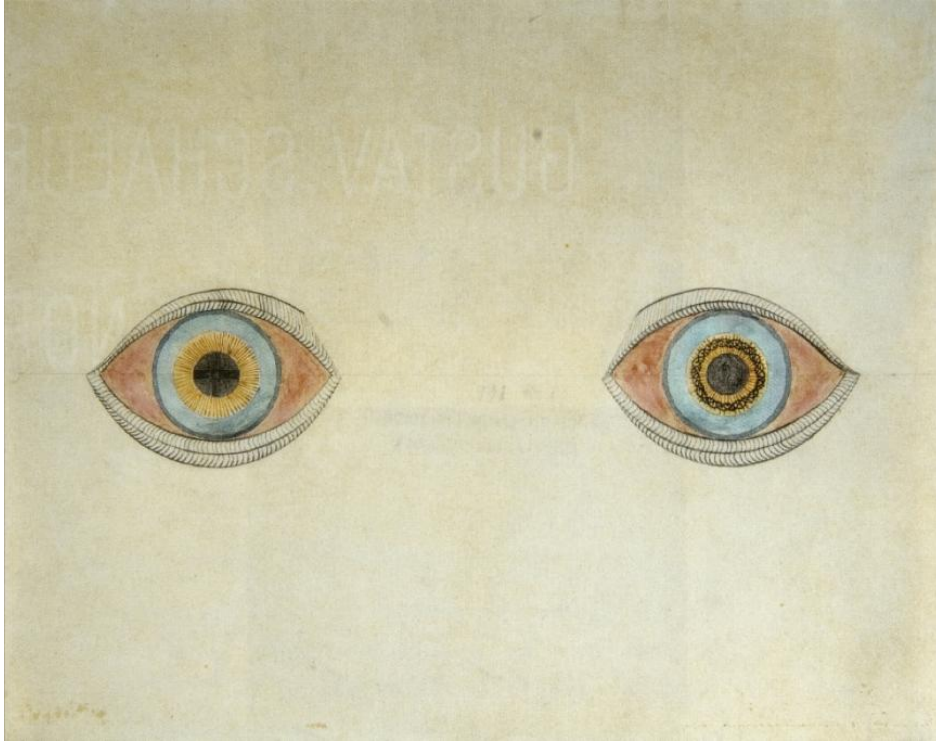
ist

nass

gefährlich

breit

menschenleer



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)

- 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, aber das Risiko abmildern
➤ *Aber wie?*

Retrieval-Augmented Generation

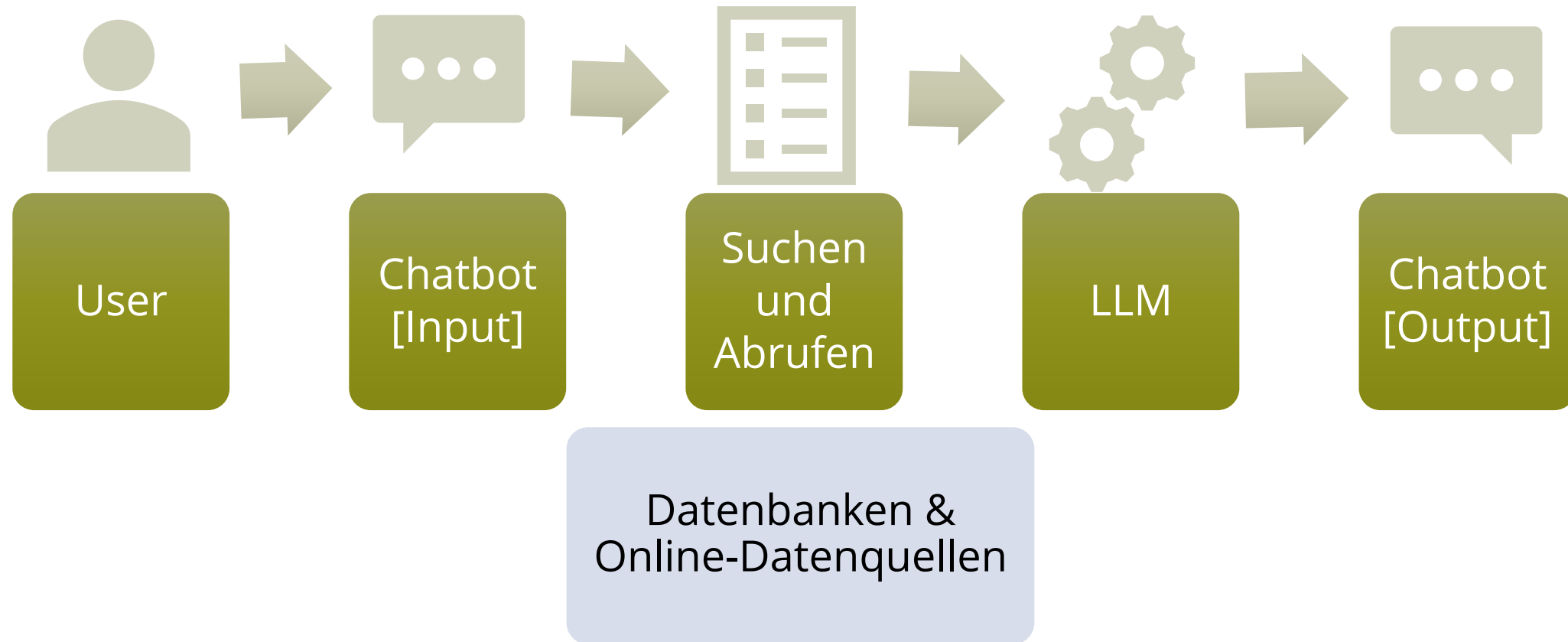
RAG steht für **Retrieval-Augmented Generation** (etwa: *per Abfrage erweiterte Generierung*)

- 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**

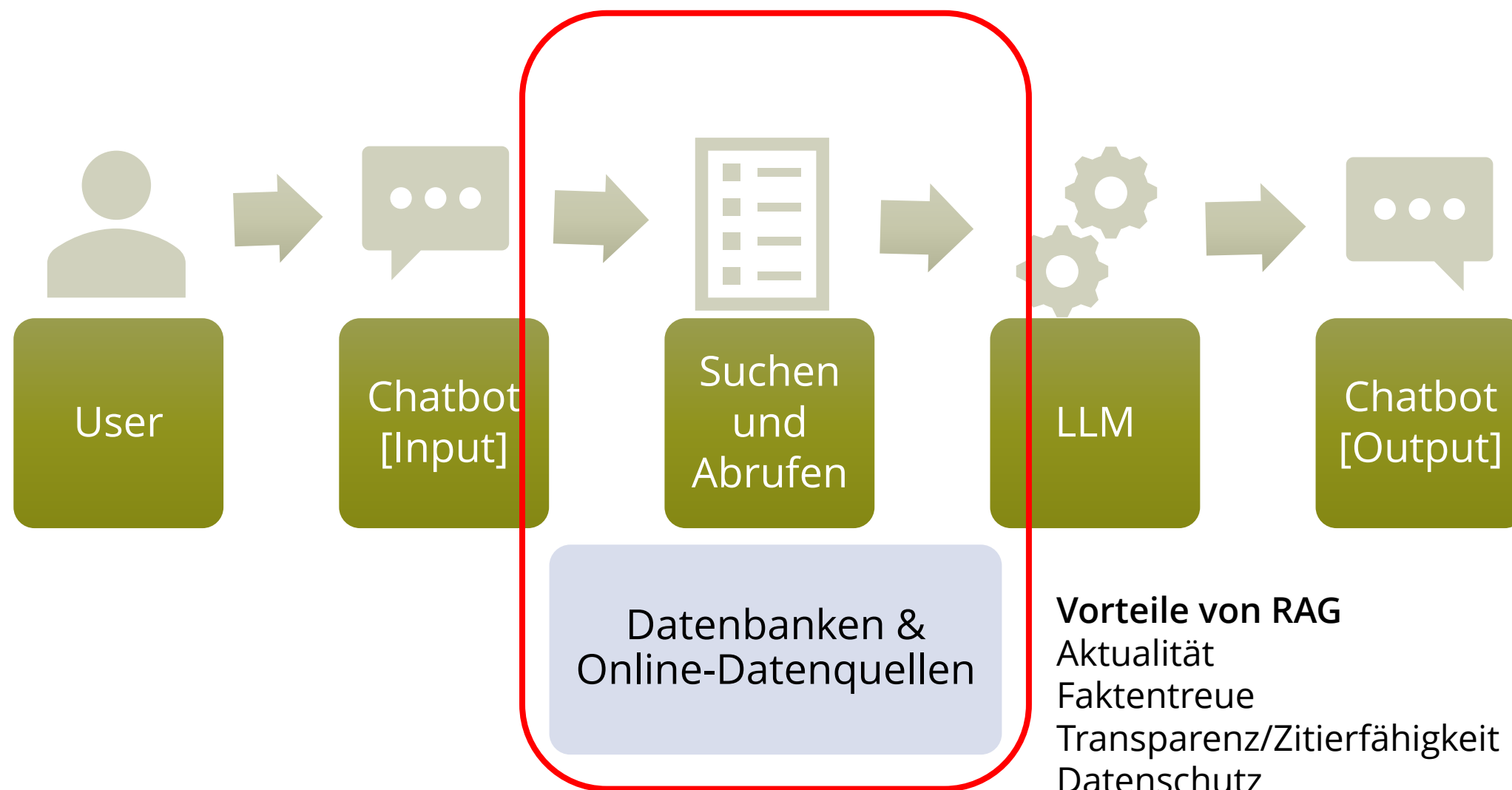
Warum braucht man RAG?

LLMs haben zwei große Schwachstellen:

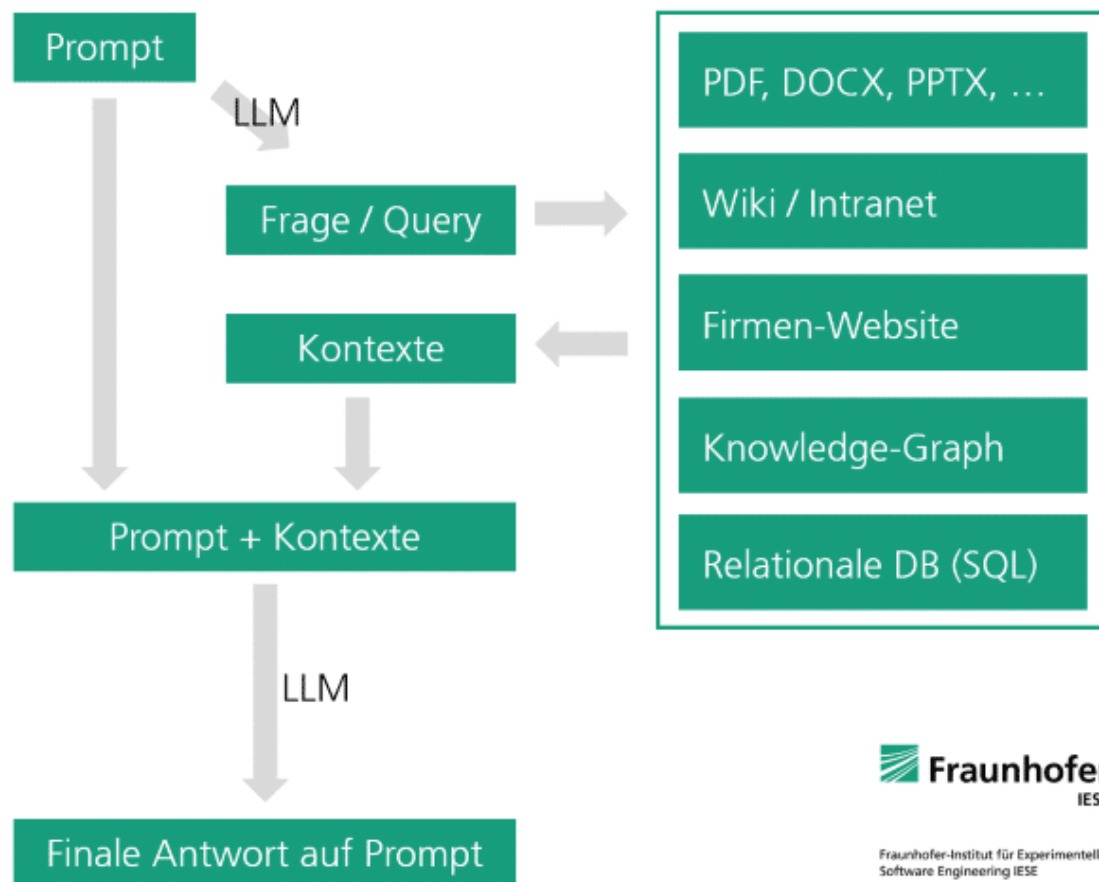
- **Halluzinationen:** Mangel an Faktentreue lässt ein LLM plausibel klingende, aber faktisch falsche Antwort erfinden
- **Wissensstopp (Knowledge Cut-off):** Die Trainingsdaten reichen nur bis zu einem bestimmten Datum; aktuelle Ereignisse sind unbekannt



LLM & RAG: Funktionsweise



LLM & RAG: Funktionsweise



Chat AI – Infrastruktur & Anbieter

ChatAI

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

Zugang über Academic Cloud



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



- Datenschutz gewährleistet (in Deutschland gehostet)
- Vielzahl von Modellen (<https://docs.hpc.gwdg.de/services/chat-ai/models/index.html>)
- Hohe Transparenz
- Teil der KI-Infrastruktur für Wissenschaft und Forschung in Deutschland
- Nutzung im Webbrowser oder via API
- Zugriff über die eigene Institution oder einen kostenfreien Einzelaccount



- 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)
 - Max. Größe 300 MB



Schauen wir uns das mal an...

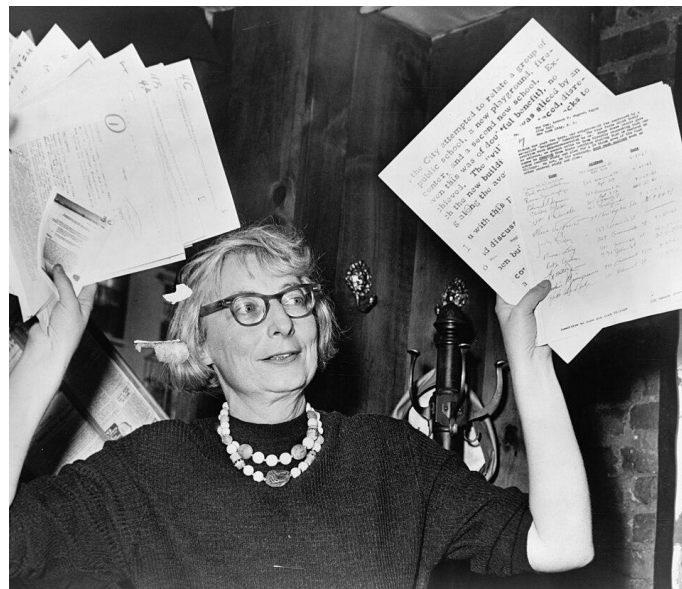
<https://chat-ai.academiccloud.de/>

Anwendungsbeispiele

▪ Beispiel 1: Einsatz von KI an der HU Berlin

Szenario: Wir sind Lehrbeauftragte an der HU Berlin und wollen wissen, ob wir den Studierenden automatisiert von einer KI Noten für Ihre Arbeiten geben dürfen

- ❖ Empfehlungspapier: KI in Studienleistungen und Prüfungen (.pdf)
 - Offizielles Dokument beschreibt Empfehlungen der Universität



▪ 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_Natural-surveillance_Jane-Jacobs.md
 - Liste mit Berliner Straßen für die Anwendung (txt-Format)

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

Einige Grenzen von RAG

- Datenqualität beeinflusst Zuverlässigkeit
 - Benennung und Strukturierung wichtig
 - Keine Garantie für völlige Halluzinationsfreiheit
 - Keine übergreifenden Erkenntnisse aus komplexer Dokumentensammlung
-
- RAG erfordert einige Arbeit
 - Kritische Überprüfung immer notwendig

Hands-on: Eigene Dokumentensammlung erstellen

Probieren Sie es aus!

- 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

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?

Till Eulenspiegel:

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

Nietzsche:

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

Dankeschön



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

Fragen & Austausch

Vorschau – Stabi-Veranstaltungen zu KI:

- **9.9., 14 Uhr, Online-Workshop: Chat AI – ein datenschutzkonformer KI-Chatbot für die Wissenschaft**
- **16.9., 14:30 Uhr, Online-Workshop: KI und Literaturrecherche**

Für ein Teilnahmezertifikat wenden Sie sich bitte an: fachinfo@sbb.spk-berlin.de

felix.hagenstroem@sbb.spk-berlin.de

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/>

Stabi Lab

<https://lab.sbb.berlin/>