



Digital Applications & Data Management

WS25/26

Dr. Jens Kohl



Roadmap Vorlesung

1. Einführung und Übersicht
2. Grundlagen Data Science
3. Vorgehen Data Science Use Case
4. Case Study Data Science
5. Grundlagen unüberwachtes Lernen
6. Grundlagen überwachtes Lernen
(tabellarische Daten)
7. Case Study überwachtes Lernen
(tabellarische Daten)
8. Grundlagen überwachtes Lernen (Bilddaten)
9. Case Study überwachtes Lernen und Transfer Learning (Bilddaten)
10. Grundlagen Generative AI
11. Generative AI mit Texten und Prompt Engineering
12. Agentic AI
13. Ausblick: Machine Learning in der Cloud und Reinforcement Learning



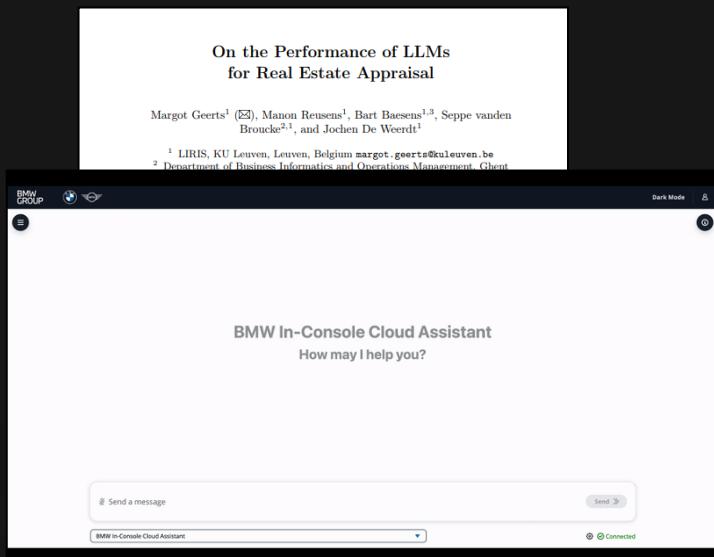
Vorlesung 12: Agentic AI



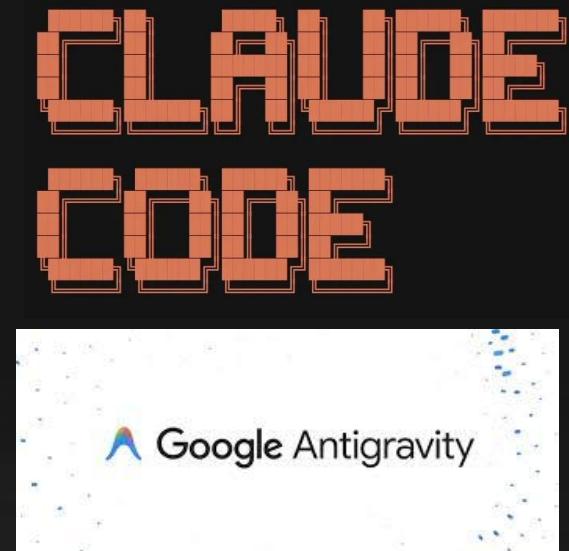
Agents

Motivation

Automatisierung Workflows



Coding assistants



Research assistants

The screenshot shows the Anthropic Research website. At the top, there's a header with the date "February 2, 2025 Release" and a "Try on ChatGPT" button. Below the header, a search bar contains the placeholder "Compile a research report on how the retail industry has cha...". To the right of the search bar, there are navigation links for "Research", "Economic Futures", "Commitments", "Learn", "News", and "Try Claude". The main content area features a section titled "What can I help with?" with a dropdown menu. Further down, there's a section titled "How we built our multi-agent research system" with a sub-section for "Engineering at Anthropic". The footer includes a "Published Jun 13, 2025" timestamp and a note about the research feature using multiple Claude agents.

Source: <https://aws.amazon.com/blogs/industries/bmw-group-develops-a-genai-assistant-to-accelerate-infrastructure-optimization-on-aws/>, <https://arxiv.org/abs/2506.11812>



Agentic AI

Ausgangssituation

wo auf der welt ist es gerade schön? gebe antwort mit 3 zielen und deren normaler temperatur zu dieser jahreszeit.
fasse dich sehr kurz.

Wenn du der Kälte entfliehen möchtest, sind diese drei Ziele im Februar besonders empfehlenswert:

Ziel

Phuket, Thailand

Dubai, VAE

Teneriffa, Spanien

Google Flüge ▾

buch einen flug nach teneriffa

Ich kann keine Flüge direkt für dich buchen, habe dir aber passende Verbindungen von München nach Teneriffa für Mitte März herausgesucht. Hier sind die besten Optionen:

- Günstigste Verbindung: Mit Vueling (1 Stopp in Barcelona) ab 182 € (z. B. 11.03. – 18.03.).
- Direktflug: Mit TUI fly ab 248 € (Flugdauer ca. 5 Std.).
- Komfort-Direktflug: Mit Condor ab 477 €.

Alle Details und die Links zur Buchung findest du hier:
[Flüge von München nach Teneriffa ansehen](#)

Grenzen klassischer LLM:

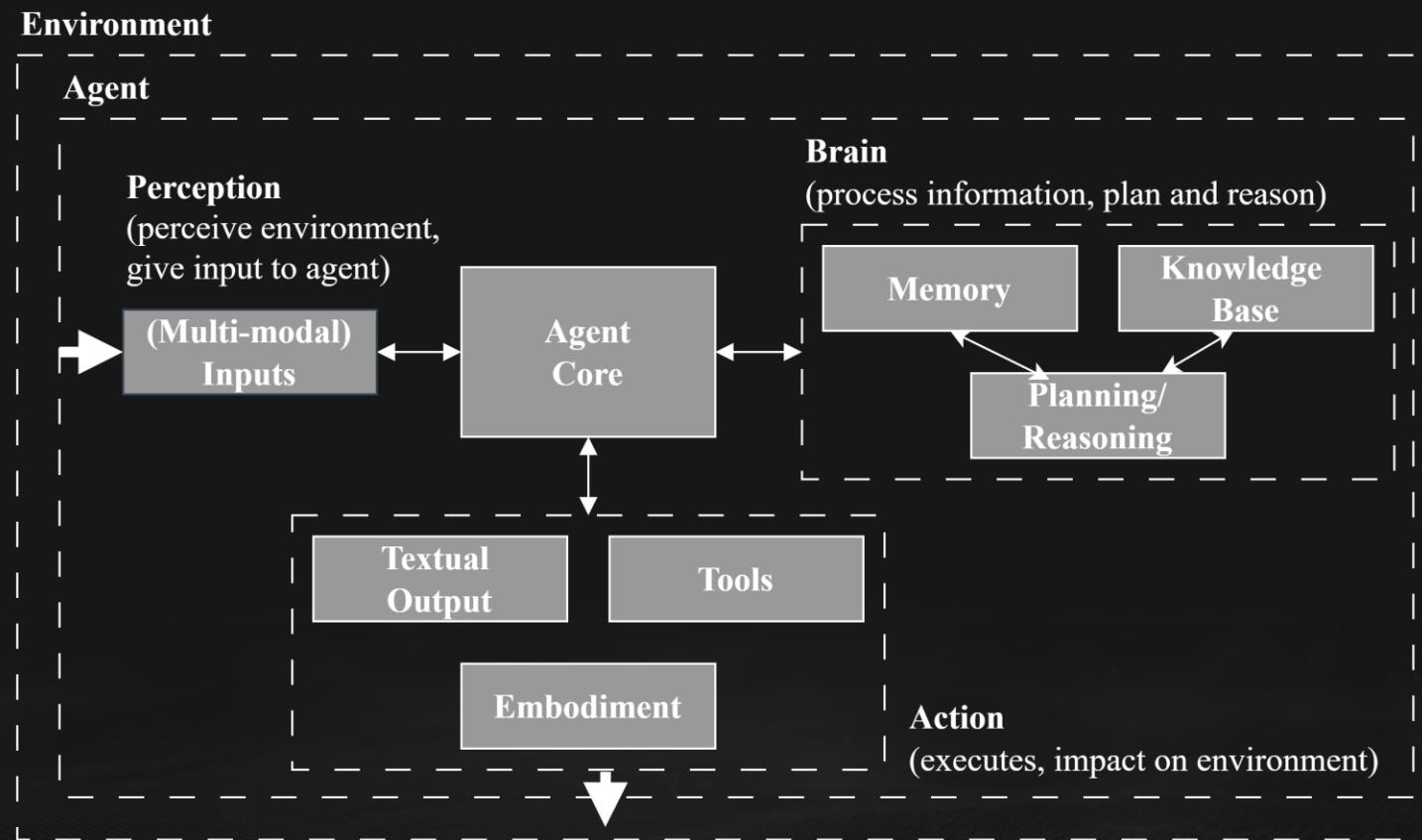
- Kein echtes Handeln/ Interaktion mit Umwelt
- Kein dauerhaftes Gedächtnis
- Keine Feedback-Loop
- Kein Komplexen Plan

Cloud-basierte LLM Produkte werden aktuell so weiter entwickelt, daß sie diese Grenzen weitgehend überwinden können.
Das Thema Interaktion mit Umwelt bleibt jedoch weiter offen....



Agentic AI

Struktur eines Agenten

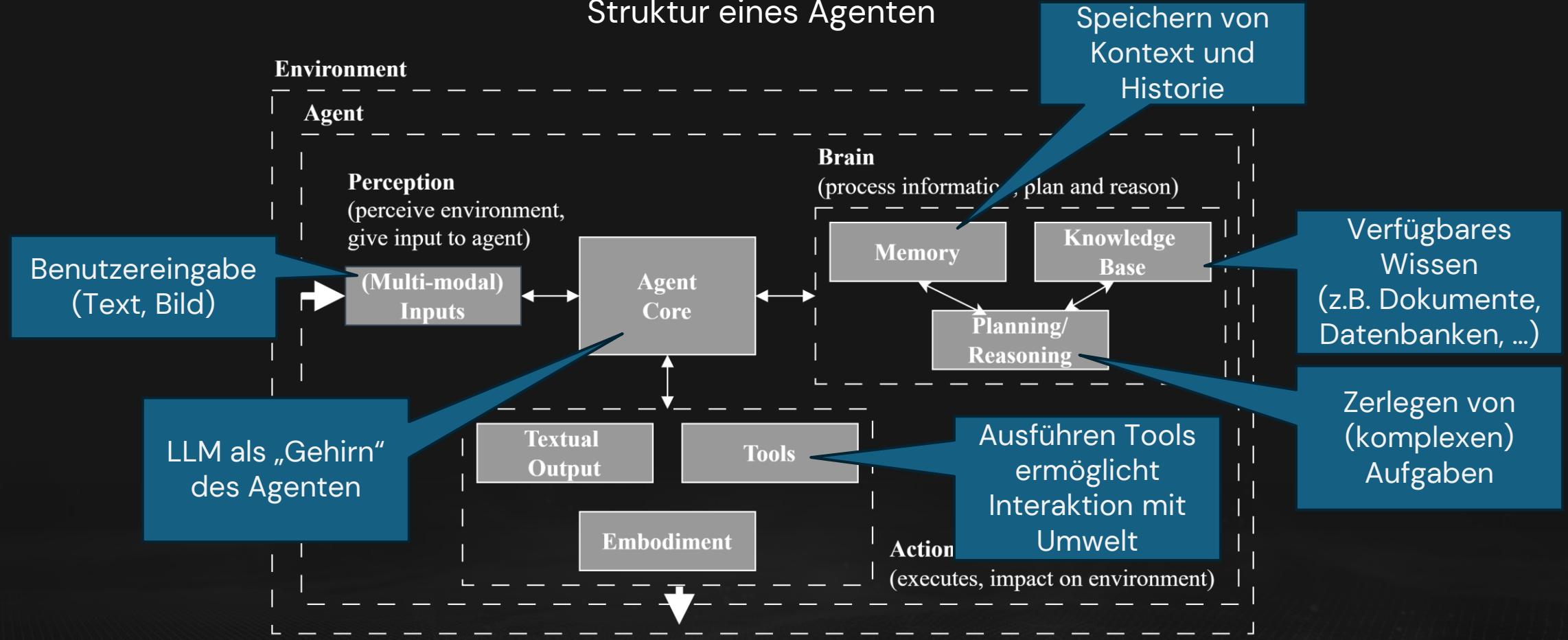


Agenten nutzen LLM als „Gehirn“ und können durch Tools auch Aktionen ausführen



Agentic AI

Struktur eines Agenten



Agenten nutzen LLM als „Gehirn“ und können durch Tools auch Aktionen ausführen



Agentic AI

Detaillierung Inputs

- **Input-Quellen (Modalitäten):**
 - Text, Dokumente, Bilder, Video und Audio
 - Kontext-Variablen wie Zeit, Ort, User-Name.
- **Input Transformation:**
 - **Encoding:** Überführung Input-Daten in einen gemeinsamen Vektorraum (Embedding)
 - **Tokenisierung:** Zerlegung von Text in Token, d.h. numerische Vektoren
- **User Intent & Persona:**
 - **Task:** Definition konkrete Zielsetzung, was der Agent machen soll, in natürlicher Sprache.
 - **Rolle:** Definition des Verhaltensmusters (z. B. „Handle als Forscher“).
- **Constraints (Leitplanken):**
 - Explizite Verbote, einzuhaltende Regeln, Guardrails (Filter problematische Eingabetexte)
 - Zielvorgabe für den Output (z. B. JSON, Tabelle).
- Verknüpfung User-Inputs mit Instruktionen des Agenten (Task, Rolle) um Agentic-Loop zu starten



Agentic AI

Detaillierung Agentic Core

- LLM ist Gehirn des Agenten mit folgenden Aufgaben:
 - **Intent Recognition:** Erkennen, was der User wirklich will.
 - **Planung:** Zerlegen der User Anfrage in Teilaufgaben (chain-of-thought)
 - **Tool selection:** Identifizieren, welches Tool für welche Aufgabe geeignet ist
 - **Stop-Kriterium:** Agenten sind in einer Dauerschleife bis die Aufgabe erfüllt ist.
- **Anforderungen an das Modell:**
 - Großes Kontext-Fenster: Muss Historie, Tool-Beschreibungen, Dokumente, User inputs, ... gleichzeitig verarbeiten.
 - Modell sollte Reasoning Fähigkeiten haben (bspw. Anthropic)
 - Striktes Einhalten von Ausgabeformaten für Interaktion mit Tools (wenn Tool nur ein spezifisches EingabefORMAT hat, muß das genau so bedient werden)
- Temperatur wird oft auf 0 gesetzt, um „Kreativität“ zu unterbinden



Agentic AI

Detaillierung Planning & Reasoning

- Strategische Zerlegung (Decomposition):
 - o Modell generiert Zwischenschritte vor Ausführen einer Aktion (Chain-of-Thought¹).
 - o Dies ermöglicht höhere Genauigkeit bei komplexen Aufgaben.
- ReAct-Framework (Reason + Act²):
 - o Thought: Analyse aktuelle Situation
 - o Action: Gezielter Aufruf eines Tools
 - o Observation: Einbeziehung externes Feedbacks (z. B. Suchergebnis/ Fehlermeldung).
 - o Somit ständiger Abgleich zwischen Planung und Realität/ Feedback möglich.
- Reflexion & Fehlerkorrektur (Self-Correction):
 - o Agent bewertet Zwischenergebnisse auf Plausibilität.
 - o Bei Tool-Fehlern oder Sackgassen wird Plan modifiziert, statt abzubrechen.
- Plan-and-Execute vs. Dynamisches Planning:
 - o Plan-first: Erstellung eines kompletten Ablaufplans vor der ersten Aktion (effizient für Standard-Tasks).
 - o Dynamisch: Schritt-für-Schritt Planung, bei der jeder nächste Schritt von der letzten Beobachtung abhängt.

Quelle:

¹ Wei et al.: "Chain-of-Thought Prompting Elicits Reasoning in Large Language Models", 2022
² Yao et al.: "ReAct: Synergizing Reasoning and Acting in Language Models", 2022.



Agentic AI

ReAct Loop am Beispiel Wetter Suche

```
from langchain_openai import ChatOpenAI
from langchain_community.tools import DuckDuckGoSearchRun
from langchain_core.messages import HumanMessage, ToolMessage

# 1. Setup Tool and LLM
search = DuckDuckGoSearchRun() # we use DuckDuckGo as search engine
tools = [search] # define tools, in our case just 1
llm = ChatOpenAI(model="gpt-4o").bind_tools(tools) # register tool

def run_manual_react(query: str):
    messages = [HumanMessage(content=query)] # Initialize message history

    while True:
        # --- STEP 1: REASONING (The "Thought") ---
        response = llm.invoke(messages)
        messages.append(response)
        # If the LLM doesn't want to use a tool, it's done reasoning
        if not response.tool_calls:
            break

        # --- STEP 2: ACTING (The "Action") ---
        for tool_call in response.tool_calls:
            print(f"Action: Searching for '{tool_call['args']['query']}...'")
            result = search.invoke(tool_call["args"]) # Execute the tool call

            # --- STEP 3: OBSERVATION (The "Observation") ---
            messages.append(ToolMessage(tool_call_id=tool_call["id"], content=result)) # Feed tool result back into agent conversation
            print(f"Observation: Found data. Summarizing...")

    return messages[-1].content

# we now want to call the agent for a specific city
city = "Munich"
answer = run_manual_react(f"What is the current weather and a top news story in {city}?")
print(f"\nFinal Answer:\n{answer}")
```



Agentic AI

Detaillierung Memory

Short-term Memory:



- **Inhalt:** Aktueller Dialogverlauf, Tool-Feedback und Denkprozesse.
- **In-Context Learning:** Modell „lernt“ aus Infos aus Prompt, ohne internen Gewichte zu verändern
- Memory wird durch Länge Context Window LLM begrenzt (z. B. 1M Token bei Claude)
- **Context Optimization (Management-Strategien)**
 - o **Sliding Window:** Fokus auf N aktuellsten Token, Rest wird gelöscht
 - o **Summarization:** bisherige Verlauf wird komprimiert zusammengefasst

Long-term Memory:



- Inhalt: Langfristiges Speichern von Wissen, Historie und Nutzerpräferenzen über verschiedene Sessions hinweg.
- Ablage: Daten werden langfristig gespeichert in
 - Klassische Datenbanken: für exaktes Suchen
 - Knowledge Graphen: Speichern von Beziehungen von Daten untereinander
 - Vektor-Datenbanken: ermöglichen semantische Suche (d.h. Ähnlichkeit)

Ziel ist das Sicherstellen der Kohärenz über lange, komplexe Workflows sowie Vermeiden redundanter Tool-Aufrufe und wiederholter Fehler.



Agentic AI

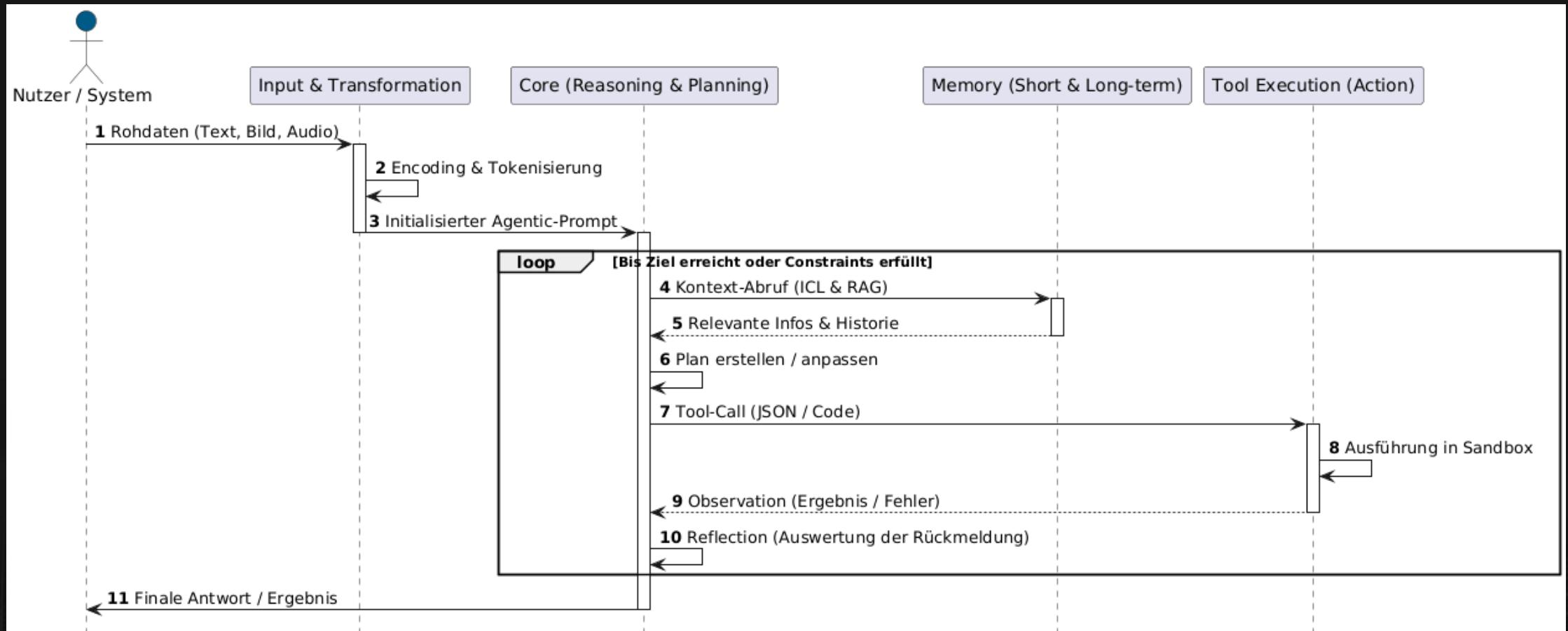
Detaillierung Tools

- Registrierung Tools für den Agenten:
 - o Tools werden LLM über präzise Meta-Daten (Name, Parameter, Beschreibung) „vorgestellt“.
 - o Tools werden über standardisierte Schnittstellen oder über generierten Code aufgerufen.
- Auswahl Tool durch Agenten:
 - o Modell wählt basierend auf dem aktuellen Plan das passende Werkzeug.
 - o Tool erhält Input gemäß seines gewünschten Formats
- Ausführung Tool durch Agenten:
 - o Ausführung von Code in isolierten Umgebungen zur Sicherheit (Sandbox)
 - o Ergebnis Aktion (Erfolg/Fehler) wird erfaßt.
- Auswertung Rückmeldung Tools durch Agenten
 - o „Observation“ wird als neuer Kontext dem Reasoning-Prozess des LLM zurückgeführt.
 - o Bei Fehlern (z. B. Timeout) entscheidet Agent autonom über einen Retry oder Strategiewechsel.



Agentic AI

Wie alles zusammenspielt





Agentic AI

Herausforderungen

- **Zuverlässigkeit & Vorhersagbarkeit:**
 - o Agenten können fiktive Tools erfinden oder falsche Rückmeldungen fehlinterpretieren (**Halluzinationen**)
 - o Kleine Änderungen im Prompt oder Input können zu unterschiedlichen Handlungsplänen führen.
- **Effizienz & Kosten:**
 - o Loop (Reasoning -> Action -> Observation) benötigt Zeit und ist langsamer als einzelnes LLM
 - o Iterationen und das Anwachsen Historie führen zu hohen Nutzungskosten (Verrechnung auf Token-Basis!)
- Sicherheit (Security):
 - o **Prompt Injection:** internen "Leitplanken" (Constraints) des Agenten umgehen.
 - o Missbrauch bei Schreibzugriffen auf Datenbanken oder Dateisysteme
- **Komplexität der Evaluierung:**
 - o Schwierigkeit, Agenten systematisch zu testen, da der Pfad zur Lösung jedes Mal variieren kann.
 - o Risiko, dass der Agent in einer Endlosschleife aus Fehlern und Korrekturversuchen hängen bleibt.

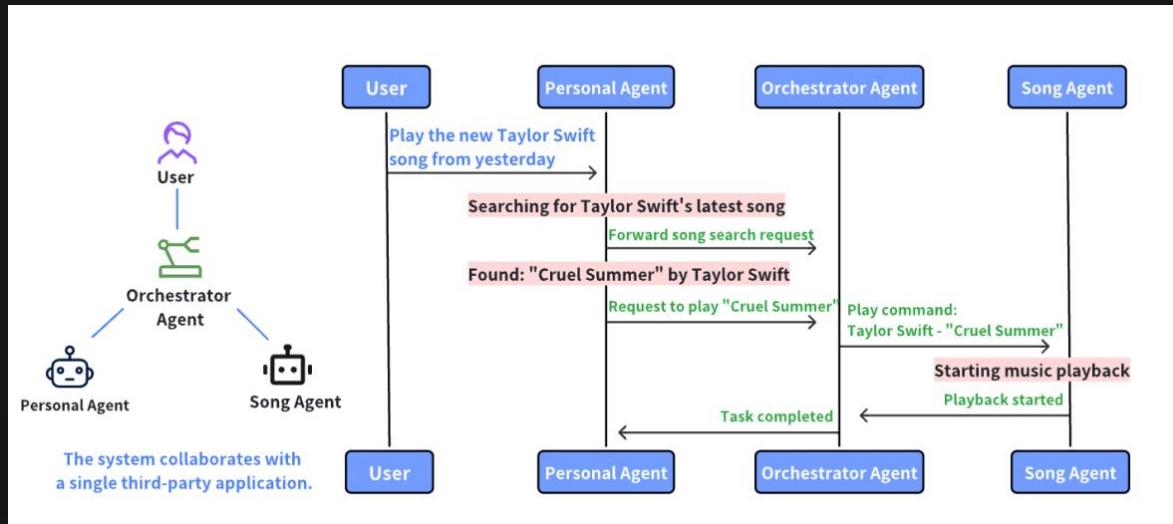
Auf absehbare Zeit wird man aufgrund dieser Herausforderungen immer noch einen Menschen als letzte Instanz für kritische Umfänge benötigen („Human-in-the-Loop“)



Agentic AI

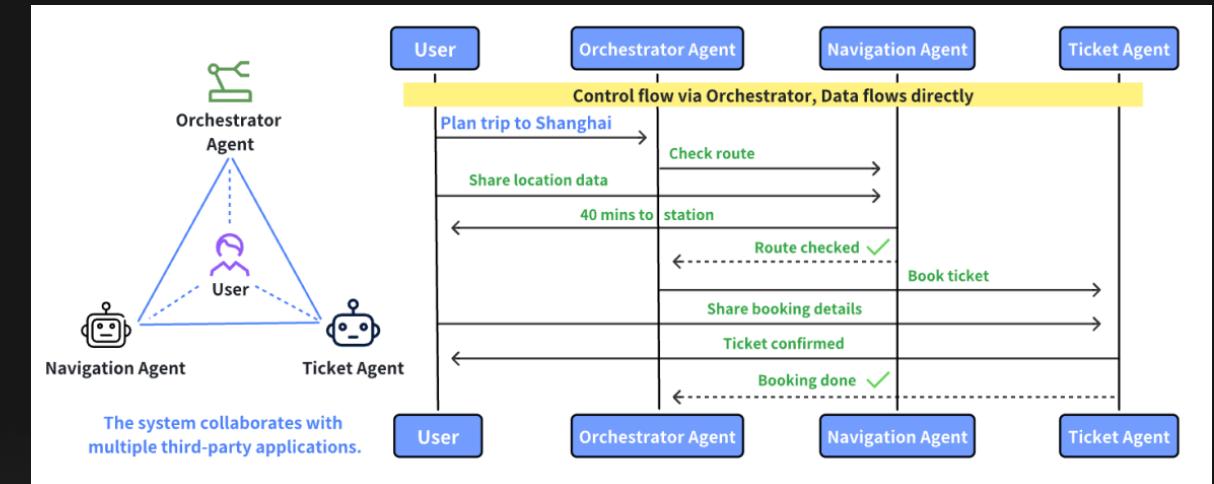
Multi-agenten System

Hierarchische Kollaboration:



1 zentraler Agent interagiert mit User, delegiert und führt Ergebnisse dann zusammen.

Sequentielle Kollaboration:



Agent A liefert Output, der für Agent B Input ist....

Falls Aufgaben für einen Agenten zu komplex oder fehleranfällig werden, kann man mehrere einzelne Agenten mit festen Rollen verwenden, die gemeinsam kollaborieren.



Case studies



Anwendungsbeispiel: Notebook LLM

Das Gesamtwerk von William Shakespeare öffentlicht

Quellen

- Alle Quellen auswählen
- A Lover's Complaint
- A Midsummer Night's Dream
- All's Well That Ends Well
- Antony and Cleopatra
- As You Like It
- Coriolanus
- Cymbeline, King of Britain
- History of Henry IV, Part I
- History of Henry IV, Part II
- History of Henry V
- History of Henry VI, Part I.md
- History of Henry VI, Part II
- History of Henry VI, Part III
- History of Henry VIII
- History of King John
- History of Richard II

Chat

Kunst und Kultur
Das Gesamtwerk von William Shakespeare
45 Quellen · 26.04.2025

In diesem Notebook für Schüler, Studenten, Wissenschaftler und Theaterliebhaber finden Sie **das Gesamtwerk Shakespeares**. Sehen Sie sich die Mindmap der wichtigsten Konzepte an, um einen Überblick über das Werk des „Barden“ zu erhalten. Sie können sich auch die Audio-Zusammenfassungen anhören, in denen die Handlung von Shakespeares berühmtesten Stücken, darunter *Othello* und *König Lear*, besprochen wird. Leser können um Erklärungen zu bestimmten Passagen oder Szenen bitten oder sich sogar kreativere Interpretationen anhören, wie etwa *Hamlet* als *Zeitungsartikelserie*. Oder Sie lesen die Stücke einfach im Original und nutzen einen KI-Leitfaden, wenn Sie Hilfe bei der Übersetzung aus dem elisabethanischen Englisch benötigen.

Text eingeben... 45 Quellen →

+ Notebook erstellen Einstellungen PRO

Studio

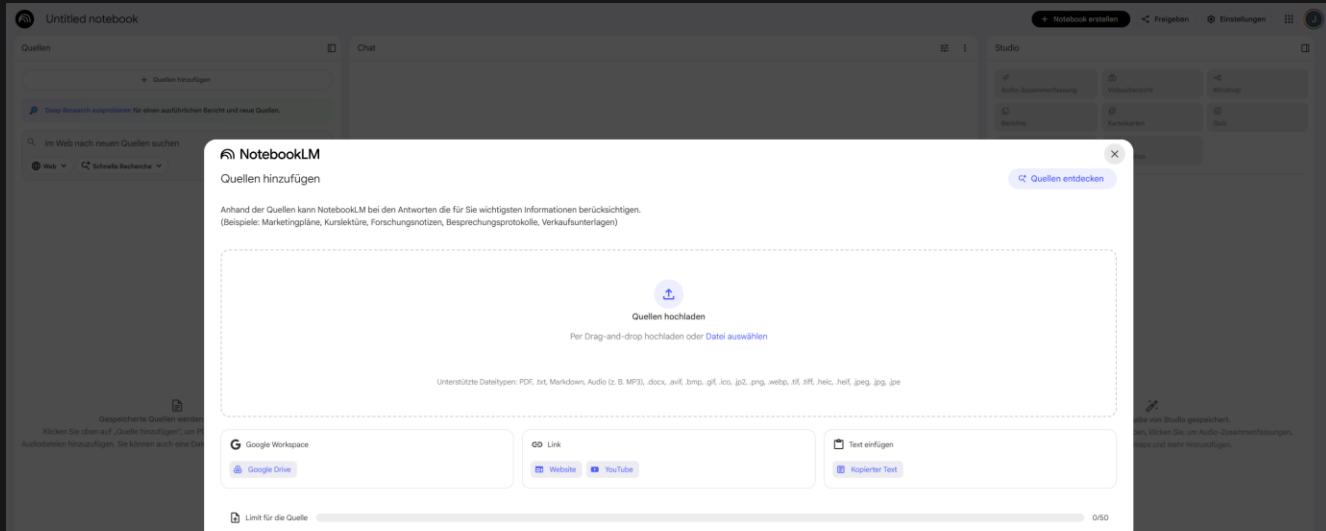
Diese Studioausgaben beinhalten eine ausführliche visuelle Übersicht und Audio-Zusammenfassung zum Thema des Notebooks

- The Burden of Kingship 45 Quellen · Vor 5 Tagen
- Shakespearean Death Causes Infographic 45 Quellen · Vor 5 Tagen
- How To Get The Most Out Of This Notebook Vor 108 Tagen
- Hamlet: Explained 2 Quellen · Vor 121 Tagen
- Study Guide: A Midsummer Night's Dream Vor 137 Tagen
- Study Guide: Romeo And Juliet Vor 137 Tagen
- Shakespeare's Tragedies: Unpacking Lear,... 3 Quellen · Vor 143 Tagen
- Macbeth: A Comprehensive Study Guide Lernplan · 2 Quellen · Vor 143 Tagen
- A Midsummer Night's Dream: Study Guide Lernplan · 1 Quelle · Vor 143 Tagen
- How Is Shakespeare's World Different from Moder... Vor 143 Tagen



Notebook-LLM: Use Case Analyse größerer Dokumente

Laden der Dokumente



Hier die Quellen, bspw. PDF-Datei, hochladen.
Sie können hier (oder im nächsten Schritt) auch andere Quellen
hochladen oder Google suchen lassen



Notebook-LLM: Use Case Analyse größerer Dokumente

Mögliche Prompts für effizientes Auswerten

1/ Get an overview

"Give me an overview of this chapter. What is the main idea? What are the key concepts?"

"Read this PDF like a teacher, not a summarizer. Explain the core idea first, then build up slowly. Assume I'm smart but new to this topic."

"I uploaded this PDF. Give me a high-level overview of the entire document, broken into key themes and concepts, as if you're introducing it to someone seeing it for the first time."

"Turn the main ideas of this document into structured outlines or mental models that help me see how everything connects."

2/ Turn Chapters Into Classes

"Turn this chapter into a 30-minute lesson plan. Include sections, examples, and quick checks for understanding."

"Turn each section or chapter of this document into a short lesson with a clear explanation, examples, and a quick summary at the end."

"Teach the content of this document step by step, starting from the basics and gradually increasing difficulty. Assume I'm learning this subject for the first time."

3/ Active Learning Mode

"Pause after every major concept and ask me a question. If I answer wrong, explain again using a different analogy."

"Create practice questions from this document that test my understanding. Ask me questions first, then explain the correct answers after."

"Extract all important concepts, terms, and definitions from this document and explain each one clearly, as if I need to remember them for an exam."

4/ Cut the Academic Noise

"Extract the 10 ideas from this PDF that actually matter in real life. Ignore filler, theory padding, and academic fluff."

"Identify the most complex or confusing parts of this document and explain them in simple language using analogies or real-world examples."

"Rewrite this document into concise study notes optimized for quick revision, highlighting only what actually matters."

5/ Teach for Application, Not Exams

"Teach this document using real-world examples from [your field]. Assume I want to apply it, not memorize it."

6/ Examiner Mode

"Act like an examiner and test me on this document. Then explain what my answers reveal about my understanding gaps."

"Pretend I just studied this document. Ask me questions and check my answers. If I'm wrong or unclear, correct me and explain what I missed."



Notebook-LLM: Use Case Analyse größerer Dokumente

Visualisieren Auswertungen

The screenshot shows a web-based application interface for document analysis and visualization. On the left, there's a sidebar titled "Grundlagen von Diffusion Models" with sections for "Quellen" (Sources) and "Chat". The "Quellen" section includes a search bar, a button to "Deep Research ausprobieren", and a "Schnelle Recherche" button. Below these are filters for "Web" and "Alle Quellen auswählen", and a link to "https://www.arxiv.org/pdf/2510.21890". The "Chat" section contains a large block of text about diffusion models, with numbered footnotes at the bottom. A callout box in the center says "Sie können Ergebnisse als Notiz speichern". On the right, there's a "Studio" panel with tabs for "Audio-Zusammenfassung", "Videoübersicht", "Mindmap", "Berichte", "Karteikarten", "Quiz", "Infografik", and "Präsentation". A blue callout box on the right side says "Oder per Klick Standard-Visualisierungen durchführen". At the bottom, there are buttons for "In Notiz speichern", "Text eingeben...", "Notiz hinzufügen", and "Notebook erstellen".



Research multi-agent

Details

```
root_agent = Agent([
    name="ResearchCoordinator",
    model=Gemini(
        model="gemini-2.5-flash-lite",
        api_key=GEMINI_API_KEY,
        retry_options=retry_config
    ),
    # instruction tells root agent HOW to use its tools (which are the other agents).
    instruction="""You are a research coordinator. Your goal is to answer the user's query by orchestrating a workflow.
1. First, you MUST call the 'ResearchAgent' tool to find relevant information on the topic provided by the user.
2. Next, after receiving the research findings, you MUST call the 'SummarizerAgent' tool to create a concise summary.
3. Finally, present the final summary clearly to the user as your response.""",
    # wrap sub-agents in 'AgentTool' to make them callable tools for the root agent.
    tools=[AgentTool(research_agent), AgentTool(summarizer_agent)],
)
```

```
research_agent = Agent(
    name="ResearchAgent",
    model=Gemini(
        model="gemini-2.5-flash-lite",
        api_key=GEMINI_API_KEY,
        retry_options=retry_config
    ),
    instruction="""You are a world class specialized research agent.
Your only job is to use the Google_search tool to find 2-3 pieces of
relevant information on the given topic and present the findings with
citations."""
    tools=[google_search],
    # Results of this agent will be stored in the session state with this name.
    output_key="research_findings",
)
```

```
summarizer_agent = Agent(
    name="SummarizerAgent",
    model=Gemini(
        model="gemini-2.5-flash-lite",
        api_key=GEMINI_API_KEY,
        retry_options=retry_config
    ),
    # instruction modified to request a bulleted list for a clear output format.
    instruction="""Read the provided research findings: {research_findings}.
Create a concise summary as a bulleted list with 3-5 key points."""
    output_key="final_summary",
)
```

```
print(response)

[Event(model_version='gemini-2.5-flash-lite', content=Content(
    parts=[
        Part(
            function_call=functionCall(
                args={
                    'request': 'latest advancements in Diffusion models and their implications for AI'
                },
                id='adk-f4564cd5-53ec-42bc-bc54-bf1128a6f0a8',
                name='ResearchAgent'
            )
        ),
        role='model'
    ), grounding_metadata=None, partial=None, turn_complete=None, finish_reason=<FinishReason.STOP: 'STOP'>, error_code=None, error_message=None, interrupted=None, custom_metadata=None, usage_metadata=None
), grounding_token_count=23, prompt_token_count=190, prompt_tokens_detail=[ModalityTokenCount(modality=MediaModality.TEXT, token_count=190)], total_token_count=213], live_session_resumption_update=None, input_transcription=None, output_transcription=None, avg_logprobs=None, logprobs_result=None, cache_metadata=None, citation_metadata=None, invocation_id=id'adk-f4564cd5-53ec-42bc-bc54-bf1128a6f0a8', name='ResearchAgent', response={

'result': """The latest advancements in diffusion models are significantly enhancing their capabilities in generating high-quality and realistic content, including images, videos, and audio. These models are revolutionizing generative AI development and enabling more personalized and intelligent automation across various industries. The implications for AI are profound, as diffusion models are surpassing traditional models like Generative Adversarial Networks (GANs) in terms of quality and efficiency. Researchers are focusing on improving the computational efficiency of diffusion models to reduce processing time and energy consumption. Techniques such as parallel processing and memory optimization are being explored to handle larger datasets and more complex models. Beyond image and video generation, diffusion models are finding applications in diverse fields such as drug discovery (generating molecular structures), scientific simulations (predicting chemical reactions), and even artistic applications (generating music and visual art). The integration of diffusion models with other AI technologies, such as transformers for language processing, is also a active area of research. Innovations like the Poisson Flow Generative Model+ (PFGM+), which integrates diffusion with Poisson Flow principles, are pushing the boundaries of what is possible in generative AI. The future of AI development looks promising, with diffusion models playing a central role in shaping the landscape of generative AI applications."}'})
    ],
    role='user'
), grounding_metadata=None, partial=None, turn_complete=None, finish_reason=None, error_code=None, error_message=None, interrupted=None, custom_metadata=None, usage_metadata=None, live_session_resumption_update=None]
```

Source Code verfügbar unter:

https://github.com/JensKohl/Lecture_Materials_Technical_Applications_Data_Management/blob/main/Notebooks/V12_Research_Agent.ipynb



Literatur und weitere Quellen

Veröffentlichungen:

- Z. Xi et al.: "The rise and potential of large language model based agents: a survey," *Science China Information Sciences*, vol. 68, no. 2, p.121101, 2025.
- Wei et al.: "Chain-of-Thought Prompting Elicits Reasoning in Large Language Models", 2022.
- Yao et al.: "ReAct: Synergizing Reasoning and Acting in Language Models", 2023
- Google: "Introduction to agents", 2025. [Link](#)
- Weng: "LLM Powered Autonomous Agents", 2023. [Link](#)

Frameworks:

- Anthropic Model Context Protocol (Quasi Standard für Verknüpfung Agenten mit externen Systemen): [Link](#)
- Google Agent to Agent Protokoll. [Link](#)
- LangGraph: Framework für Bau von Agenten. [Link](#)
- Generative AI Toolkit: Framework für Bau und Testen von Agenten. [Link](#)
- CrewAI: Framework für den Bau von Multi-Agenten. [Link](#)



Backup