~~SiloX-GPT~~

Content Search

Datensilos verbinden mittels Multi-Agenten-Systemen

Data2Day 2024

Moritz Glauner

Head of Data Science @ Bertelsmann

in moritzglauner

Nico Kreiling

Data Scientist @ scieneers

Host of techtiefen.de

in nico-kreiling

Why classical RAG was not an option...

€20.2 Billion
Group revenues

€3,119 Million
Operating EBITDA

€1,326 Million
Group profit

80,418
Employees

Gütersloh
Headquarters

No central
Data Storage

Penguin Random House

BMG

RTL

BERTELSMANN

arvato group

Bertelsmann Marketing Services

Bertelsmann Education Group

BI Bertelsmann Investments

At our starting-point, is was not yet clear, that LLMs can work with Knowledge-Graphs (very well)

Initial Situation:

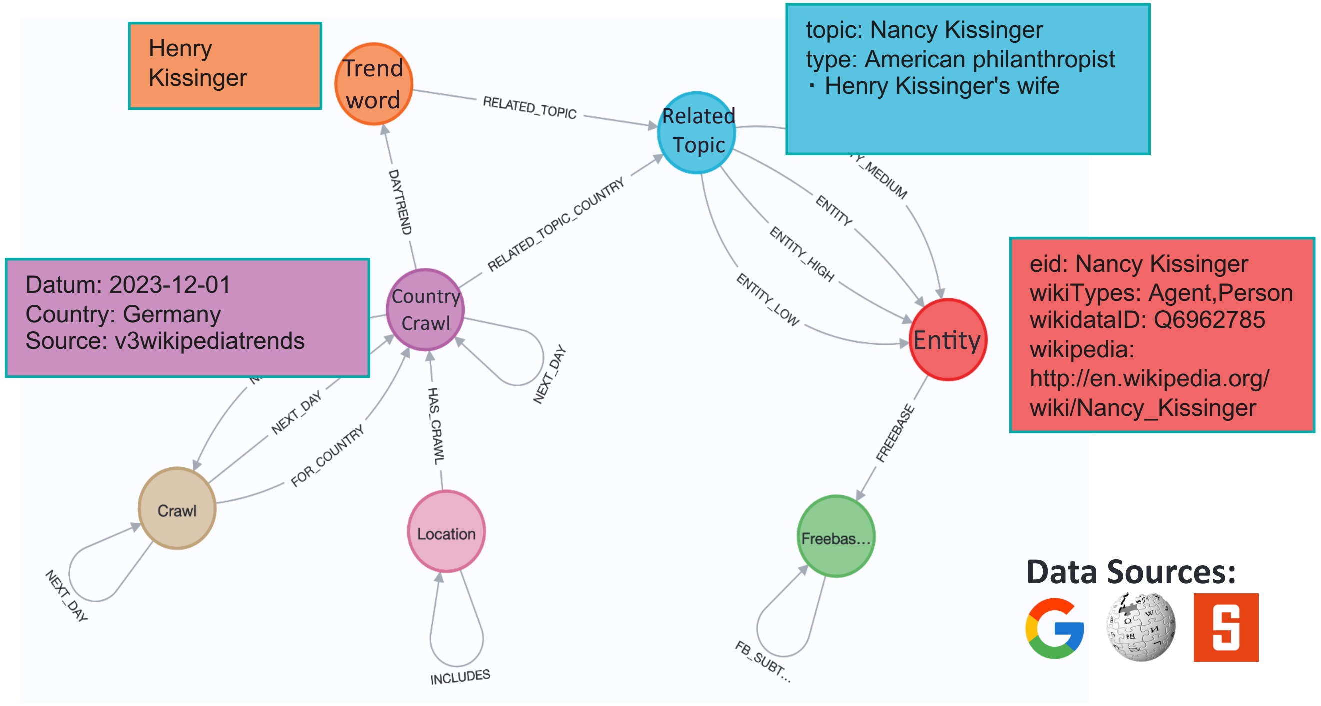
- ➢ Existing knowledge graph from predecessor project in Neo4J

Research-Question:

- ➢ Can natural questions be answered by an LLM and a Knowledge-Graph?

Learnings:

- GPT4 can write decent cypher querries when prompted with schema and some examples
- Simple-intuitive schemas work better then complex-formel
- We want to have a natural-language answer



Henry
Kissinger

topic: Nancy Kissinger
type: American philanthropist
・Henry Kissinger's wife

Datum: 2023-12-01
Country: Germany
Source: v3wikipediatrends

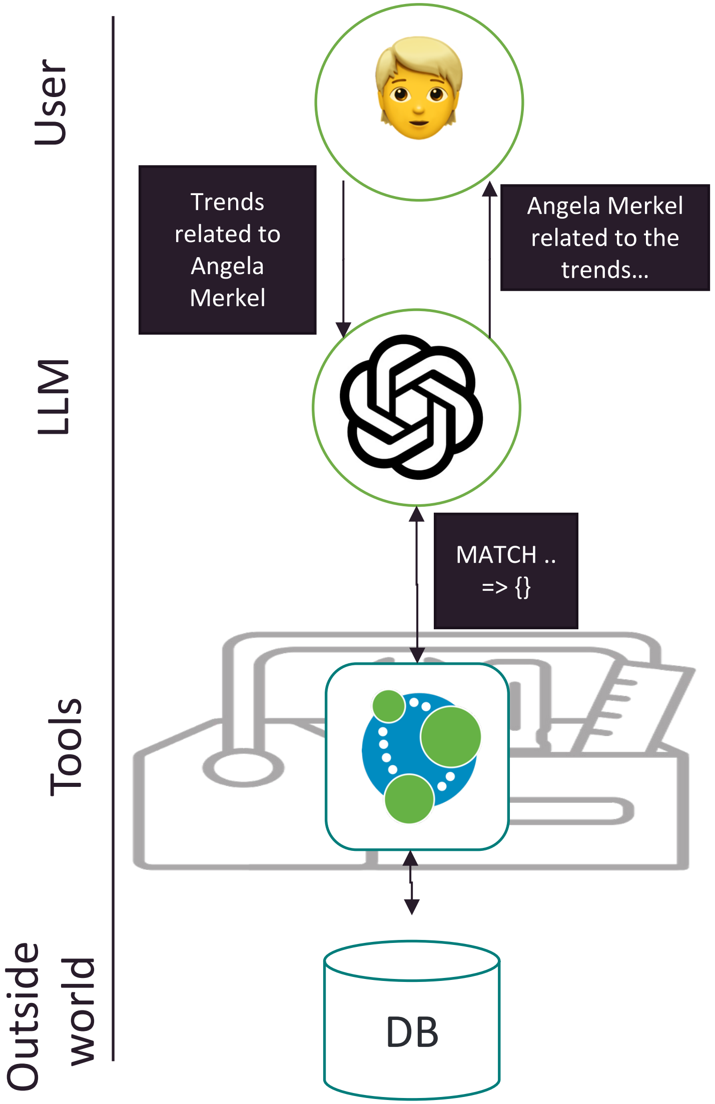
eid: Nancy Kissinger
wikiTypes: Agent,Person
wikidataID: Q6962785
wikipedia:
http://en.wikipedia.org/
wiki/Nancy_Kissinger

Data Sources:

Agents enable LLMs to do real work

Agents give LLMs the ability to interact with the world by executing programs

- Minimal version: Simple function-call with one parameter
- Complex version: Browsing or writing and testing own source code



User

LLM

Tools

Outside world

Trends related to Angela Merkel

Angela Merkel related to the trends...

MATCH .. => {}

DB

User

Welche Trends stehen in Bezug zu Angela Merkel?

LLM

Trigger function: `execute_cypher`
With `cypher=”MATCH (c:CountryCrawl)-[d:DAYTREND]->(t:Trendword) WHERE t.trend= 'Angela Merkel'`

Tool

[{year: 2015, topic: Flüchtlingskriese},
{year: 2020, topic: Corona Pandemie}]

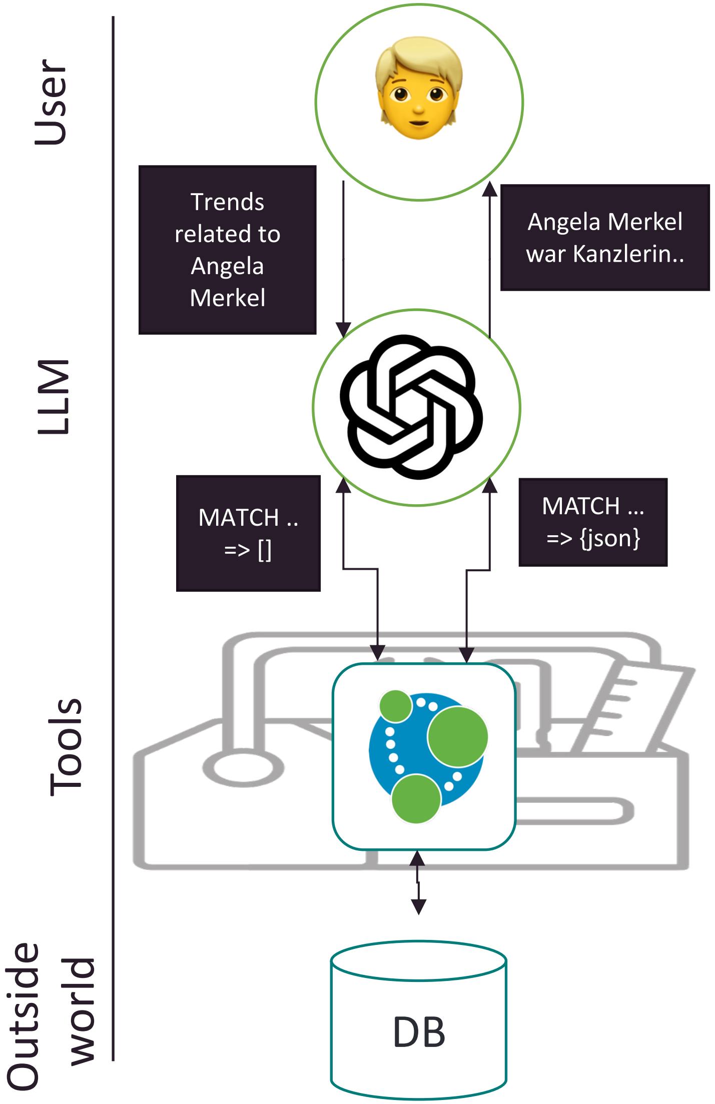
LLM

Relevante Trends in Bezug auf Angela Merkel und ihre Kanzlerschaft waren:
* Flüchtlingskrise: 2015 traf Merkel die kontroverse Entscheidung, Deutschlands Grenzen für Flüchtlinge offen zu halten.
* Corona-Pandemie: In ihrer letzten Amtszeit war die Bewältigung der COVID-19-Pandemie(2020)

Agents enable LLMs to do real work

Agents give LLMs the ability to interact with the world by executing programs

- Minimal version: Simple function-call with one parameter
- Complex version: Browsing or writing and testing own source code


User
LLM
Tools
Outside world
Trends related to Angela Merkel
Angela Merkel war Kanzlerin..
MATCH .. => []
MATCH … => {json}
DB

User

Welche Trends waren besonders relevant für die Kanzlerschaft von Merkel?

LLM

Trigger function: `execute_cypher`
With `cypher="MATCH … WHERE t.trend= 'Kanzlerschaft Merkel'`

Tool

[]

LLM

Trigger function: `execute_cypher`
With `cypher="MATCH … WHERE t.trend= 'Kanzler'`

Tool

[{topic: Angela Merkel},
{topic: Olaf Scholz}]

LLM

Angela Merkel war Kanzlerin der Bundesrepublik Deutschland. Olaf Scholz ist ihr Nachfolger.

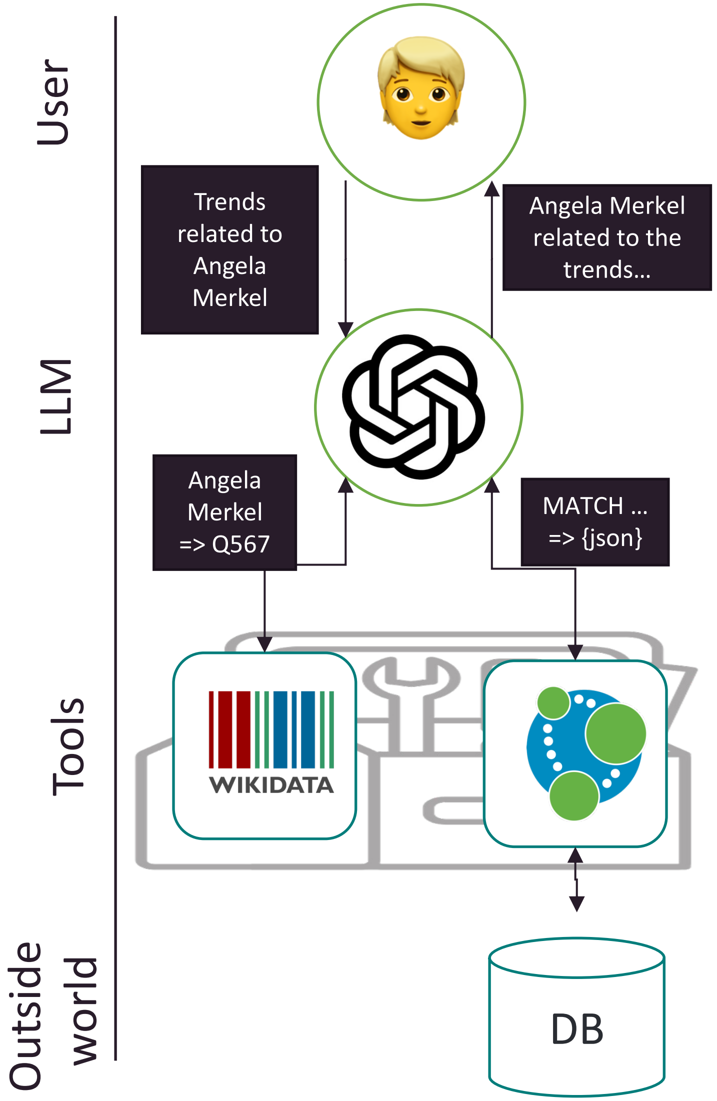
Agents enable LLMs to do real work

Agents give LLMs the ability to interact with the world by executing programs

- Minimal version: Simple function-call with one parameter
- Complex version: Browsing or writing and testing own source code

Learnings:

- Agents are much more powerfull then simple LLMs
- Important to keep the focus of an agent



Trends related to Angela Merkel

Angela Merkel related to the trends...

Angela Merkel => Q567

MATCH ... => {json}

DB

User

Welche Trends waren besonders relevant für die Kanzlerschaft von Merkel?

LLM

Trigger function: `lookup_wikidata_id`
With `name="Angela Merkel"`

Tool

Q567

LLM

Trigger function: `execute_cypher`
With `cypher=”MATCH (c:CountryCrawl)-[d:DAYTREND]->(t:Trendword) WHERE t.wikidata_id = 'Q567'`

Tool

[{year: 2015, topic: Flüchtlingskriese},
{year: 2020, topic: Corona Pandemie}]

LLM

Relevante Trends in Bezug auf Angela Merkel und ihre Kanzlerschaft waren:
* Flüchtlingskrise: 2015 traf Merkel die kontroverse Entscheidung, Deutschlands Grenzen für Flüchtlinge offen zu halten.
* Corona-Pandemie: In ihrer letzten Amtszeit war die Bewältigung der COVID-19-Pandemie(2020)

Agents are similiar to humans: Specialication and Collaboration increases the scope of possibilities – and the complexity

Using multiple agents increases flexibility of the system

- Use different LLMs for different tasks
- Have clear and focused system-prompts
- Control the length of the chat-history
- Gain more

But they also increase the complexity

- Requires communication accross agents
- Generel technical complexity increases

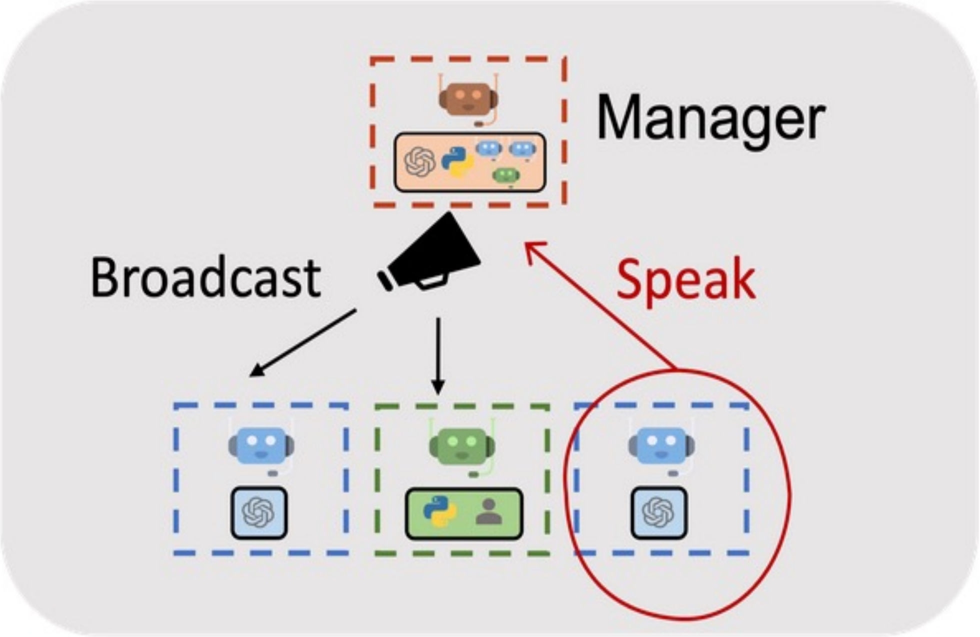
AutoGen uses an open-discussion-style communication

AutoGen in a nutshell:

- AutoGen was the first project that specificly targeted cooperation among multiple agents to solve tasks
- Message flow looks like an open discussion between agents
- Has dedicated agents for user, that is also responsible for tool execution

Learnings:

- Open message flow limits system relieabilty
- No support for parallel execution of agents
- Customization requires deep interventions in code



AutoGen

Image taken from: https://github.com/langchain-ai/langgraph/blob/main/examples/visualization.ipynb

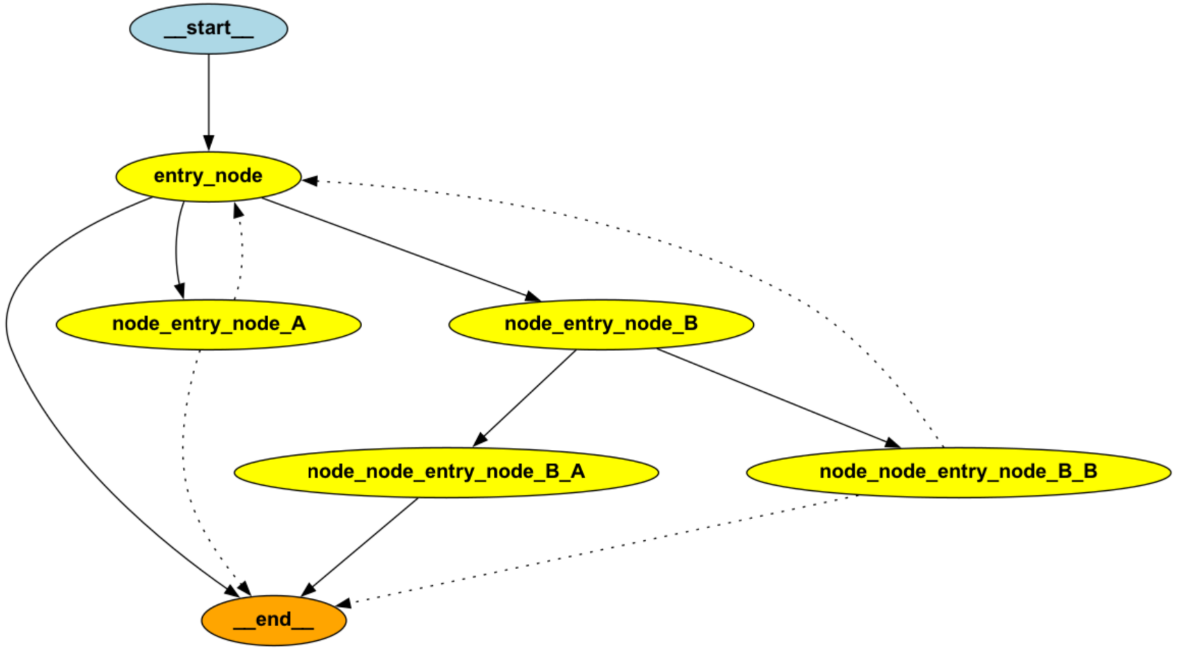
LangGraph uses a graph to design the message flow

LangGraph in a nutshell:

- Released with LangChain v0.1
- Graph design inspired by networkx
- Provides a nice way to realize very complex message flows in a deterministic way
- LangChain ecosystem brings great tool support
- Much better then early LangChain versions, but still not always intuitive

Learnings:

- Graph-Design suits much better to our use-case
- LangChain ecosystem seams more production-ready



LangGraph

Image taken from: https://github.com/langchain-ai/langgraph/blob/main/examples/visualization.ipynb

Each division has so much individual knowledge! But it's sometimes hard to make sense of it

What music do we have the rights to?

Content Intelligence

BMG

RTL

Penguin Random House

What video material do we have about the elections?

What books do we have about football?

Who is watching dating shows on RTL+?

Audience Intelligence

RTL

dpv Deutscher Pressevertrieb

AdAlliance

RIVERTY

What age is the average viewer of a show?

Trend Intelligence

BeTrend

RTL

Which international shows are receiving the most attention?

What topics dominated the news recently?

How can we query this data?

What music do we have the rights to?

What video material do we have about the elections?

What books do we have about football?

Which international shows are receiving the most attention?

Who is watching dating shows on RTL+?

What products do people buy?

What topics dominated the news recently?

Just use a public LLM

Fine-Tune your own LLM

RAG System

At this, Winky howled even harder, her squashed-tomato of a nose dribbling all down her front, though she made no effort to stem the flow.

"Dobby has traveled the country for two whole years, sir, trying to find work!" Dobby squeaked. "But Dobby hasn't found work, sir, because Dobby wants paying now!"

The house-elves all around the kitchen, who had been listening and watching with interest, all looked away at these words, as though Dobby had said something rude and embarrassing. Hermione, however, said, "Good for you, Dobby!"

"Thank you, miss!" said Dobby, grinning toothily at her. "But most wizards doesn't want a house-elf who wants paying, miss. 'That's not the point of a house-elf,' they says, and they slammed the door in Dobby's face! Dobby likes work, but he wants to wear clothes and he wants to be paid, Harry Potter. . . . Dobby likes being free!"

The Hogwarts house-elves had now started edging away from Dobby, as though he were carrying something contagious. Winky, however, remained where she was, though there was a definite increase in the volume of her crying.

"And then, Harry Potter, Dobby goes to visit Winky, and finds out Winky has been freed too, sir!" said Dobby delightedly.

At this, Winky flung herself forward off her stool and lay facedown on the flagged stone floor, beating her tiny fists upon it and positively screaming with misery. Hermione hastily dropped dow to her knees beside her and tried to comfort her, but nothing s said made the slightest difference. Dobby continued with his sto shouting shrilly over Winky's screeches.

"And then Dobby had the idea, Harry Potter, sir! 'Why doesn'

Dobby and Winky find work together?' Dobby says. 'Where is there enough work for two house-elves?' says Winky. And Dobby thinks, and it comes to him, sir! *Hogwarts!* So Dobby and Winky came to see Professor Dumbledore, sir, and Professor Dumbledore took us on!"

Dobby beamed very brightly, and happy tears welled in his eyes again.

"And Professor Dumbledore says he will pay Dobby, sir, if Dobby wants paying! And so Dobby is a free elf, sir, and Dobby gets a Galleon a week and one day off a month!"

"That's not very much!" Hermione shouted indignantly from the floor, over Winky's continued screaming and fist-beating.

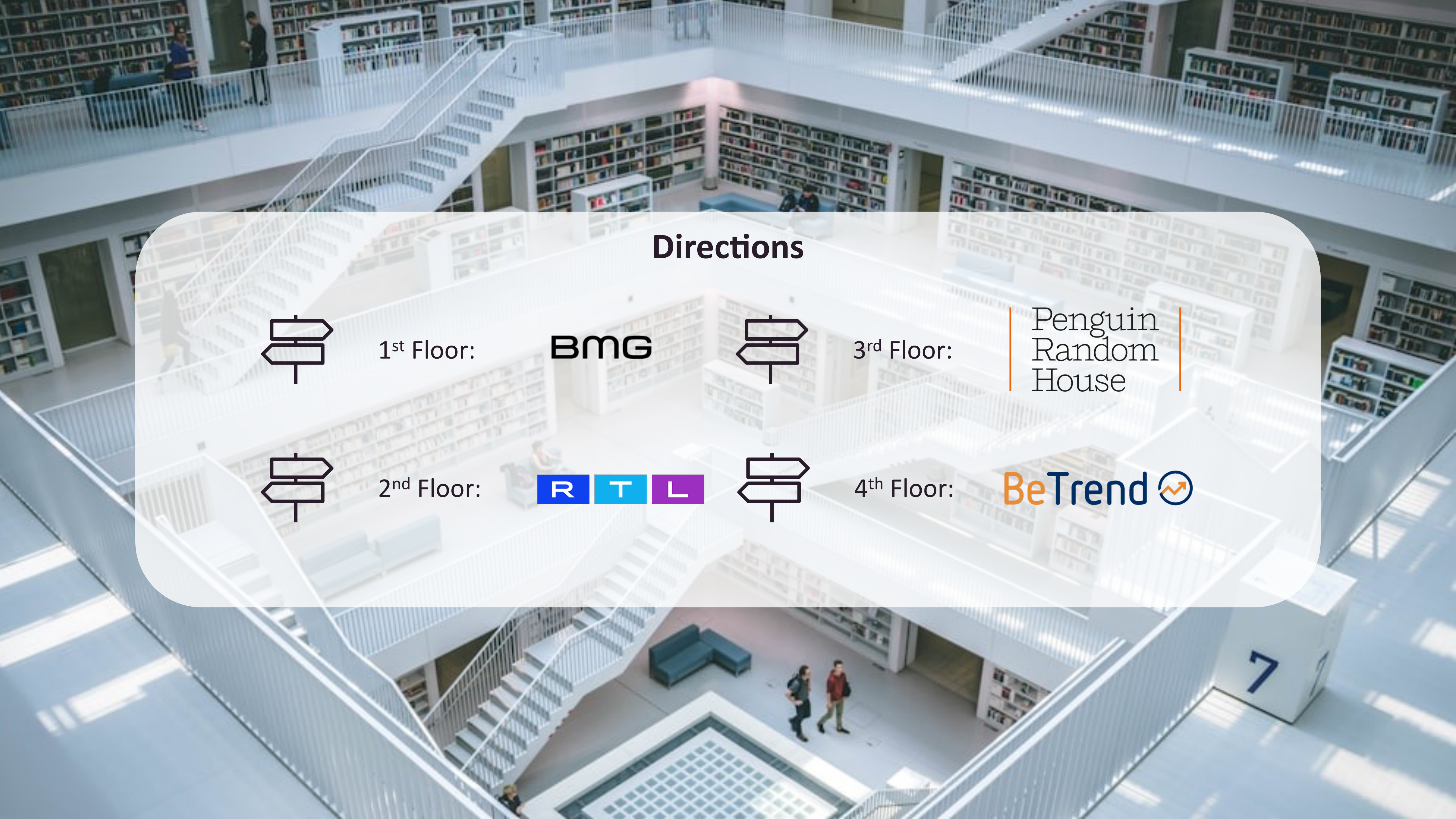
"Professor Dumbledore offered Dobby ten Galleons a week, and weekends off," said Dobby, suddenly giving a little shiver, as though the prospect of so much leisure and riches were frightening, "but Dobby beat him down, miss. . . . Dobby likes freedom, miss, but he isn't wanting too much, miss, he likes work better."

"And how much is Professor Dumbledore paying *you*, Winky?" Hermione asked kindly.

If she had thought this would cheer up Winky, she was wildly mistaken. Winky did stop crying, but when she sat up she was glaring at Hermione through her massive brown eyes, her whole face sopping wet and suddenly furious.

"Winky is a disgraced elf, but Winky is not yet getting paid!" she squeaked. "Winky is not sunk so low as that! Winky is properly ashamed of being freed!"

"Ashamed?" said Hermione blankly. "But — Winky, come on! It's Mr. Crouch who should be ashamed, not you! You didn't do anything wrong, he was really horrible to you —"



Directions

1st Floor: BMG

2nd Floor: RTL

3rd Floor: Penguin Random House

4th Floor: BeTrend

Different agent-categories are used to access different types of data

RTL+
Fremantle
BENET

Agents using vector search:
Similar to classical RAG, but with the possibility to add filters and to use tools multiple times

Tools: qdrant

IPRNOW
Penguin Random House
Penguin Random House Verlagsgruppe

API-based Agents
Using tools that wrap the relevant parts of the specific APIs

Only custom Tools

BeTrend

Graph-Based Agents:
Executes dynamically filled Cypher-Template-Querries

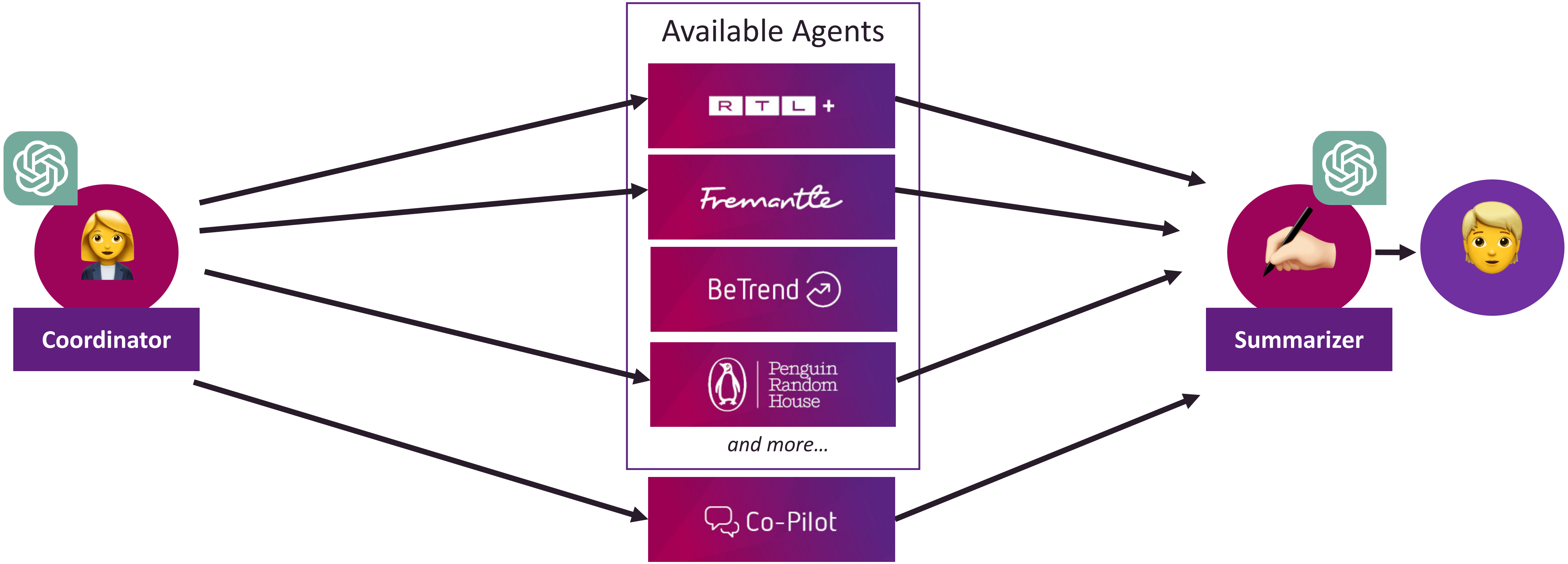
Tools: neo4j
WIKIPEDIA

Websearch
Co-Pilot

Other Agents:
Perform web-search and add relevancy information for the selected agents

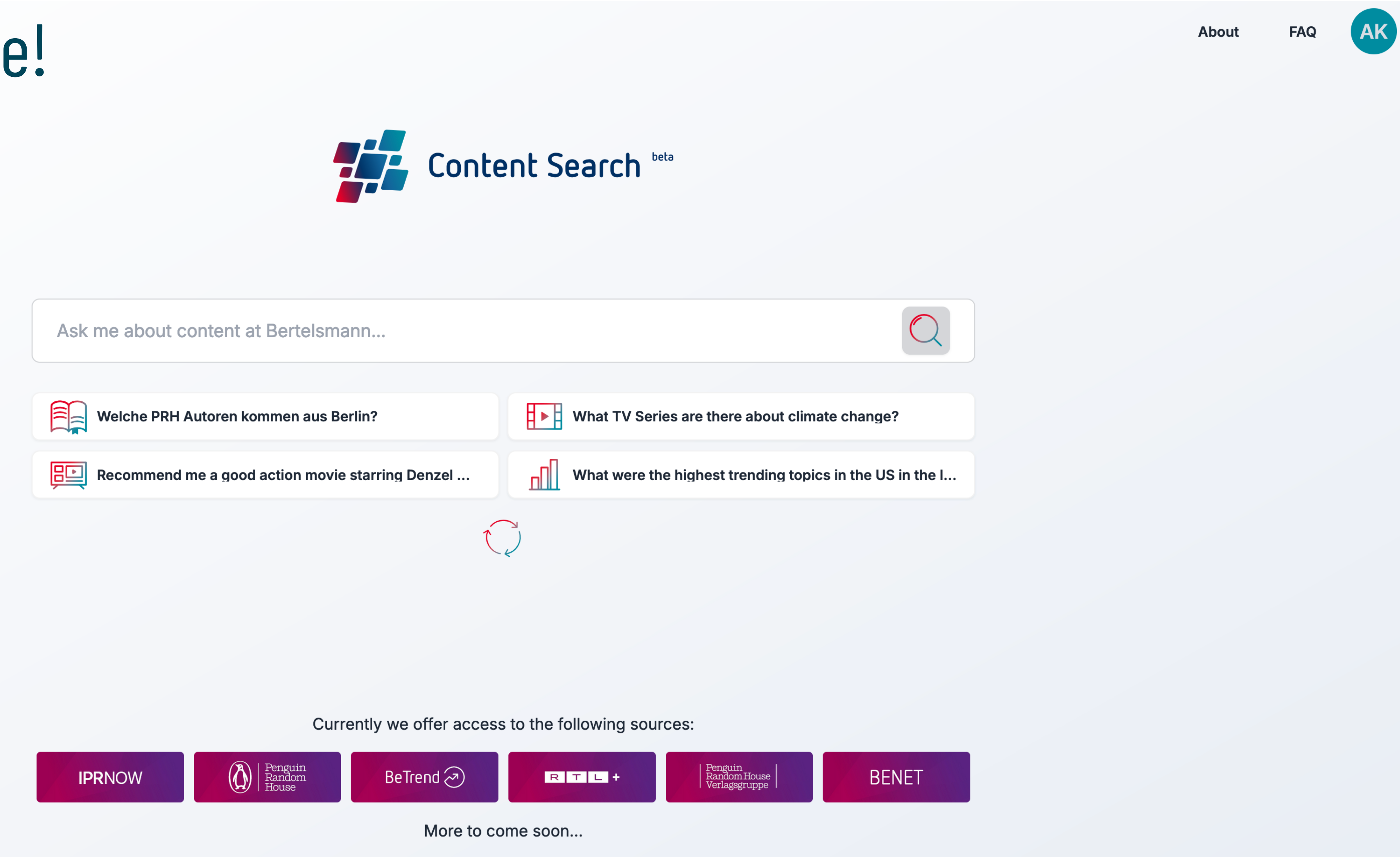
Tools: Tavily

Based on the user request the relevant agents get triggered to combine them in a single answer



BERTELSMANN scieneers DRIVEN BY DATA

It's demo time!



About FAQ AK

Content Search beta

Ask me about content at Bertelsmann...

- Welche PRH Autoren kommen aus Berlin?
- What TV Series are there about climate change?
- Recommend me a good action movie starring Denzel ...
- What were the highest trending topics in the US in the l...

Currently we offer access to the following sources:

IPRNOW | Penguin Random House | BeTrend | RTL+ | Penguin Random House Verlagsgruppe | BENET

More to come soon...

The LangChain Ecosystem Offers Much

Most Popular High-Level LLM Building Framework
⭐ 92k Github, Python + JS,

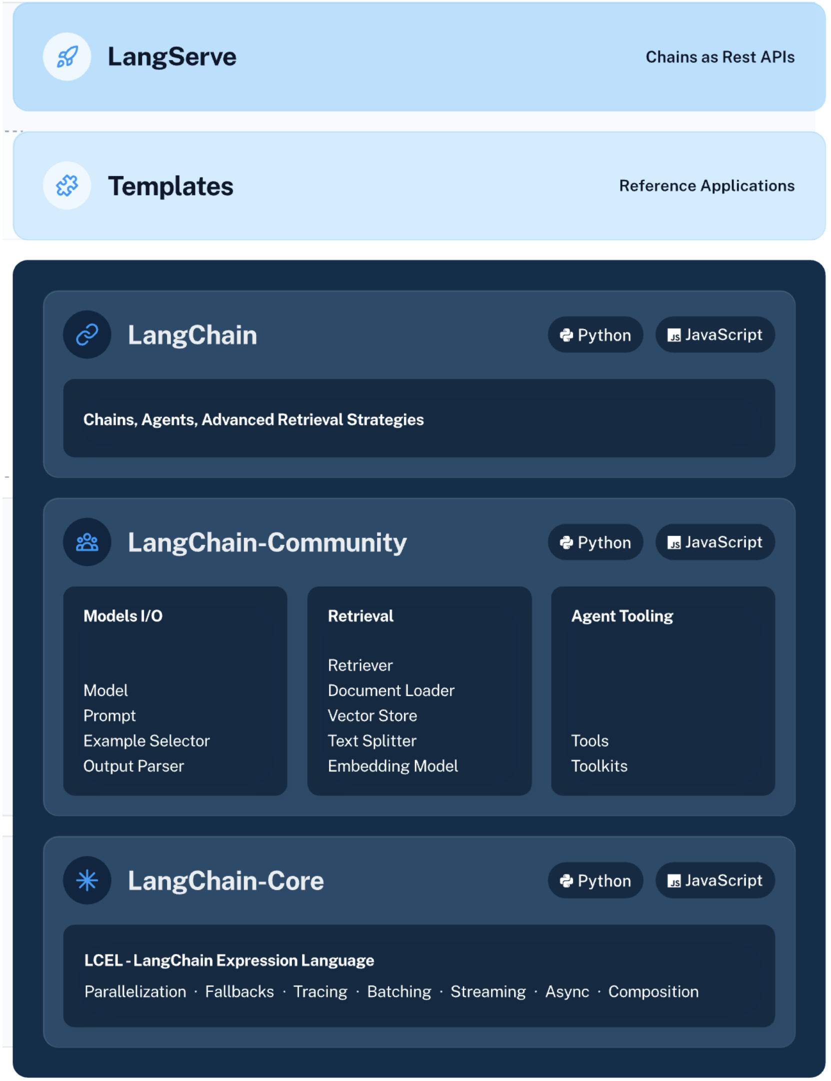
Getting (a bit) more production ready

- (Mostly) Backwards compatible with major versions now
- v0.3 now with Pydantic 2 support
- $10M Seed funding

Chains as Core Concept

- Chain = Input + Prompt + LLM
- Plus, optional components

LangChain Overview – the core package to create LLM chains



LangServe Chains as Rest APIs

Templates Reference Applications

LangChain Python JavaScript

Chains, Agents, Advanced Retrieval Strategies

LangChain-Community Python JavaScript

Models I/O	Retrieval	Agent Tooling
Model Prompt Example Selector Output Parser	Retriever Document Loader Vector Store Text Splitter Embedding Model	Tools Toolkits

LangChain-Core Python JavaScript

LCEL - LangChain Expression Language

Parallelization · Fallbacks · Tracing · Batching · Streaming · Async · Composition

Image taken from: https://python.langchain.com/v0.1/docs/get_started/introduction

The LangChain Ecosystem goes beyond the LangChain package

LangChain

Build individual components by combining LLM, storage and prompts

LangGraph

Orchestrate complex flows through one or more components

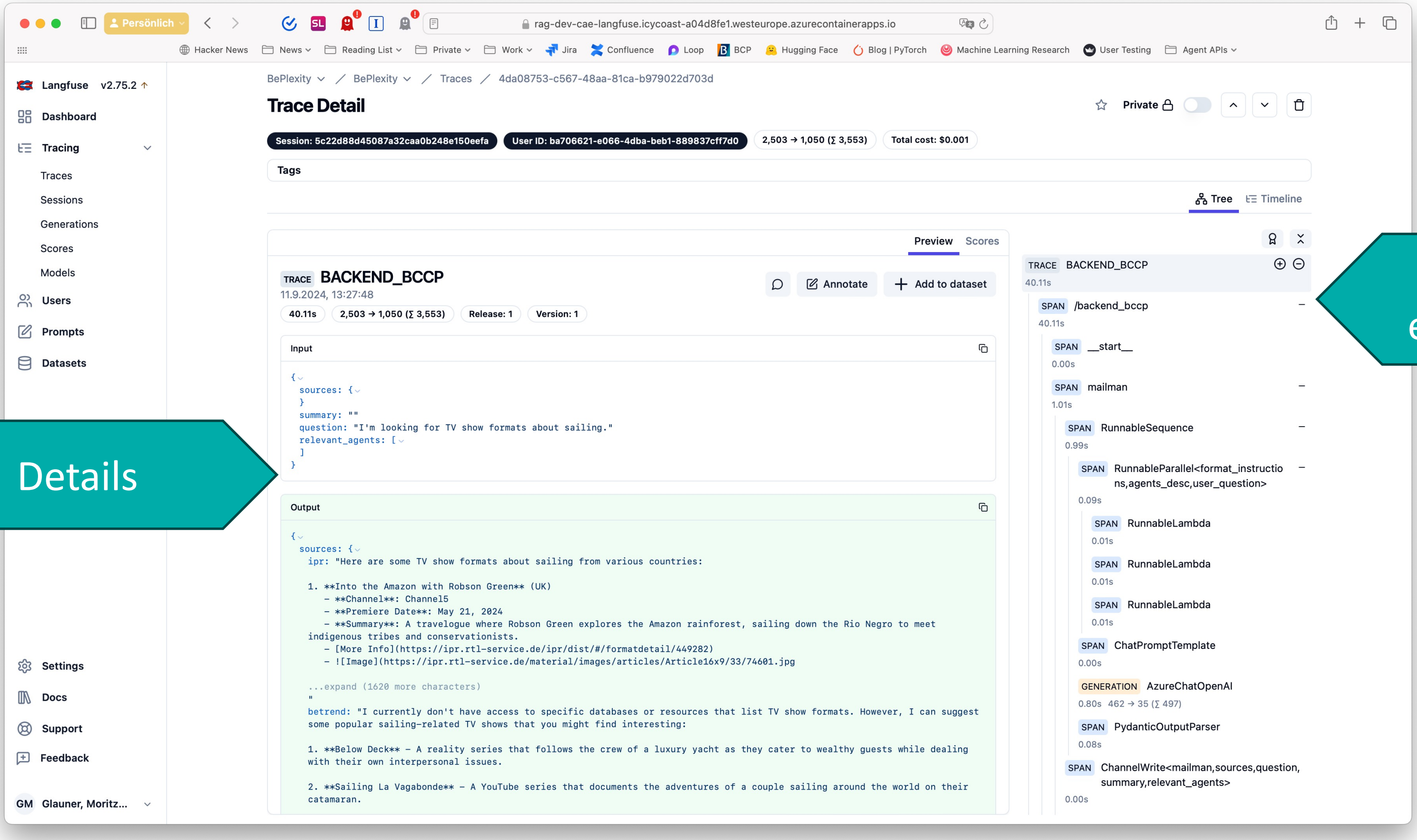
LangServe

Wrap LangChain chains available as FastAPI endpoints

Langfuse

Monitor and continuously evaluate applications

LangFuse is the LLMOps jack-of-all-traits – Tracking and Tracing



Trace Detail

Session: 5c22d88d45087a32caa0b248e150eefa
User ID: ba706621-e066-4dba-beb1-889837cff7d0
2,503 → 1,050 (∑ 3,553)
Total cost: $0.001

Tags

TRACE **BACKEND_BCCP**
11.9.2024, 13:27:48
40.11s
2,503 → 1,050 (∑ 3,553)
Release: 1
Version: 1

Input

```
{
  sources: {
  }
  summary: ""
  question: "I'm looking for TV show formats about sailing."
  relevant_agents: [
  ]
}
```

Output

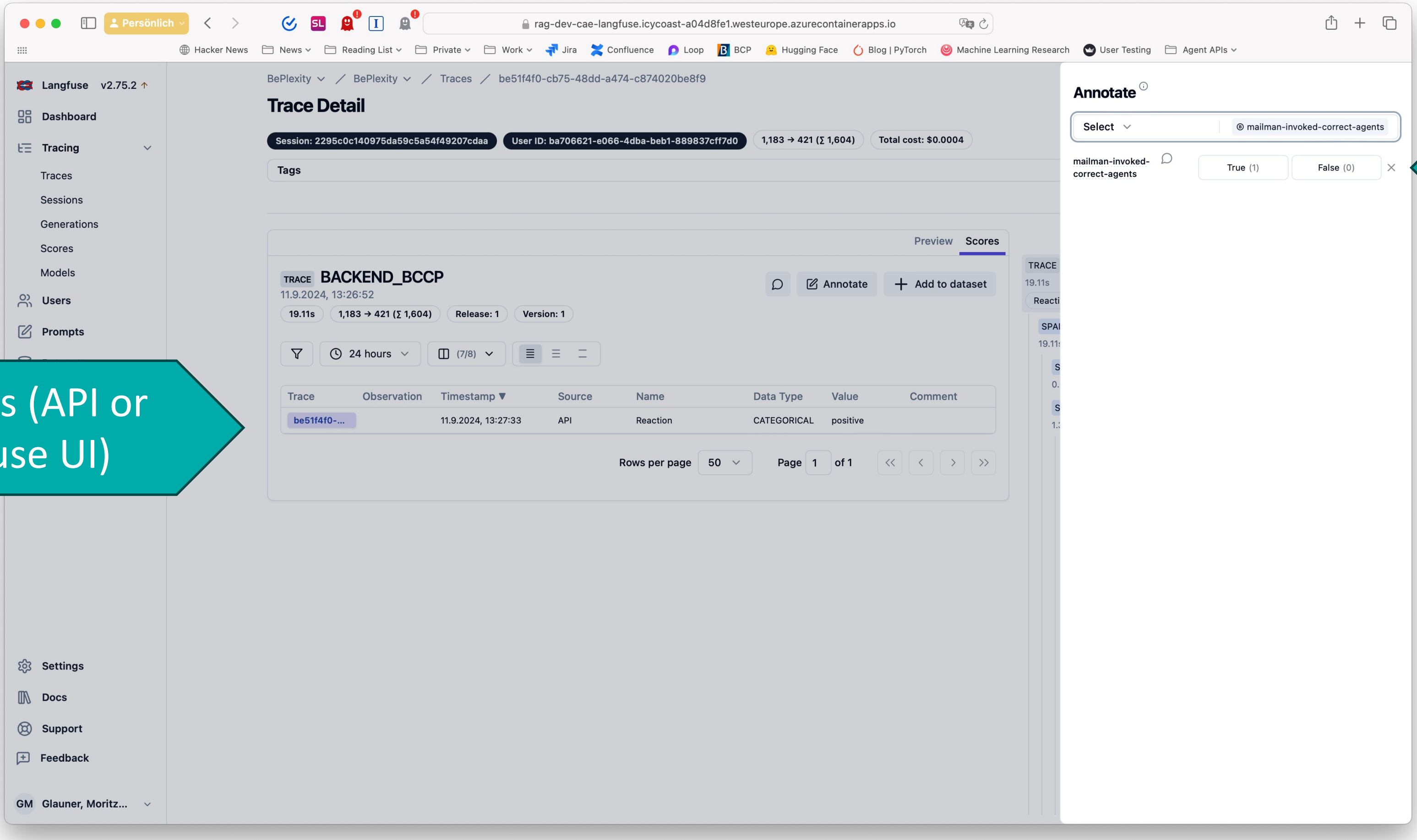
```
{
  sources: {
    ipr: "Here are some TV show formats about sailing from various countries:

    1. **Into the Amazon with Robson Green** (UK)
       - **Channel**: Channel5
       - **Premiere Date**: May 21, 2024
       - **Summary**: A travelogue where Robson Green explores the Amazon rainforest, sailing down the Rio Negro to meet
    indigenous tribes and conservationists.
       - [More Info](https://ipr.rtl-service.de/ipr/dist/#/formatdetail/449282)
       - ![Image](https://ipr.rtl-service.de/material/images/articles/Article16x9/33/74601.jpg

    ...expand (1620 more characters)
    "
    betrend: "I currently don't have access to specific databases or resources that list TV show formats. However, I can suggest
    some popular sailing-related TV shows that you might find interesting:

    1. **Below Deck** - A reality series that follows the crew of a luxury yacht as they cater to wealthy guests while dealing
    with their own interpersonal issues.

    2. **Sailing La Vagabonde** - A YouTube series that documents the adventures of a couple sailing around the world on their
    catamaran.
```

```
TRACE BACKEND_BCCP
40.11s
  SPAN /backend_bccp
  40.11s
    SPAN __start__
    0.00s
    SPAN mailman
    1.01s
      SPAN RunnableSequence
      0.99s
        SPAN RunnableParallel<format_instructions,agents_desc,user_question>
        0.09s
          SPAN RunnableLambda
          0.01s
          SPAN RunnableLambda
          0.01s
          SPAN RunnableLambda
          0.01s
        SPAN ChatPromptTemplate
        0.00s
        GENERATION AzureChatOpenAI
        0.80s 462 → 35 (∑ 497)
        SPAN PydanticOutputParser
        0.08s
      SPAN ChannelWrite<mailman,sources,question,summary,relevant_agents>
      0.00s
```

Details

Overview of executed steps

LangFuse is the LLMOps jack-of-all-traits – Evaluation (Scoring)



BePlexity / BePlexity / Traces / be51f4f0-cb75-48dd-a474-c874020be8f9

Trace Detail

Session: 2295c0c140975da59c5a54f49207cdaa
User ID: ba706621-e066-4dba-beb1-889837cff7d0
1,183 → 421 (∑ 1,604)
Total cost: $0.0004

Tags

TRACE **BACKEND_BCCP**
11.9.2024, 13:26:52
19.11s
1,183 → 421 (∑ 1,604)
Release: 1
Version: 1

Trace	Observation	Timestamp ▼	Source	Name	Data Type	Value	Comment
be51f4f0-...		11.9.2024, 13:27:33	API	Reaction	CATEGORICAL	positive	

Rows per page 50 Page 1 of 1

Annotate

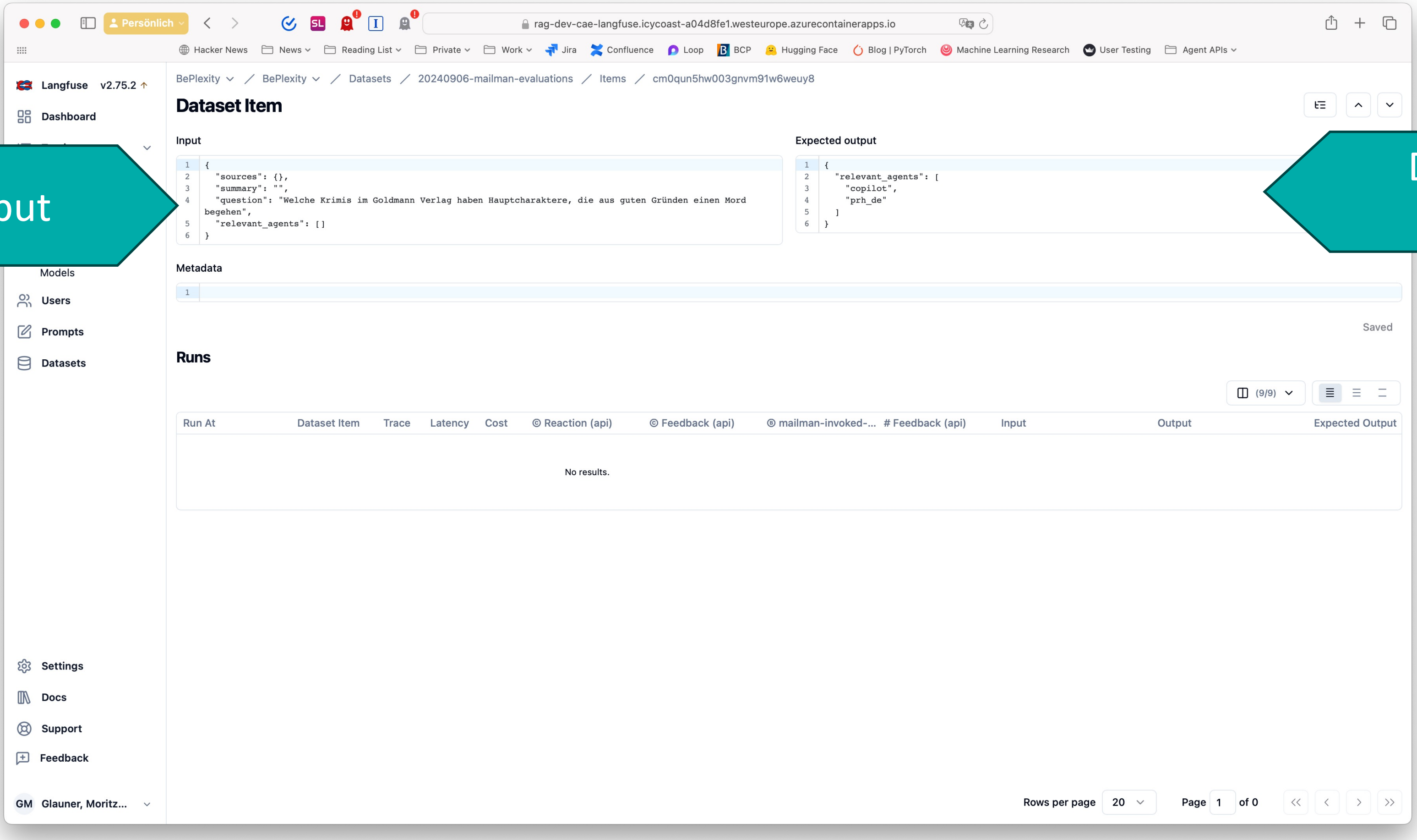
Select mailman-invoked-correct-agents

mailman-invoked-correct-agents True (1) False (0)

All Scores (API or LangFuse UI)

Manual annotation

LangFuse is the LLMOps jack-of-all-traits – Evaluation (Annotation & datasets)



Dataset Item

User input

Input

```
{
  "sources": {},
  "summary": "",
  "question": "Welche Krimis im Goldmann Verlag haben Hauptcharaktere, die aus guten Gründen einen Mord begehen",
  "relevant_agents": []
}
```

Desired system output

Expected output

```
{
  "relevant_agents": [
    "copilot",
    "prh_de"
  ]
}
```

Metadata

Saved

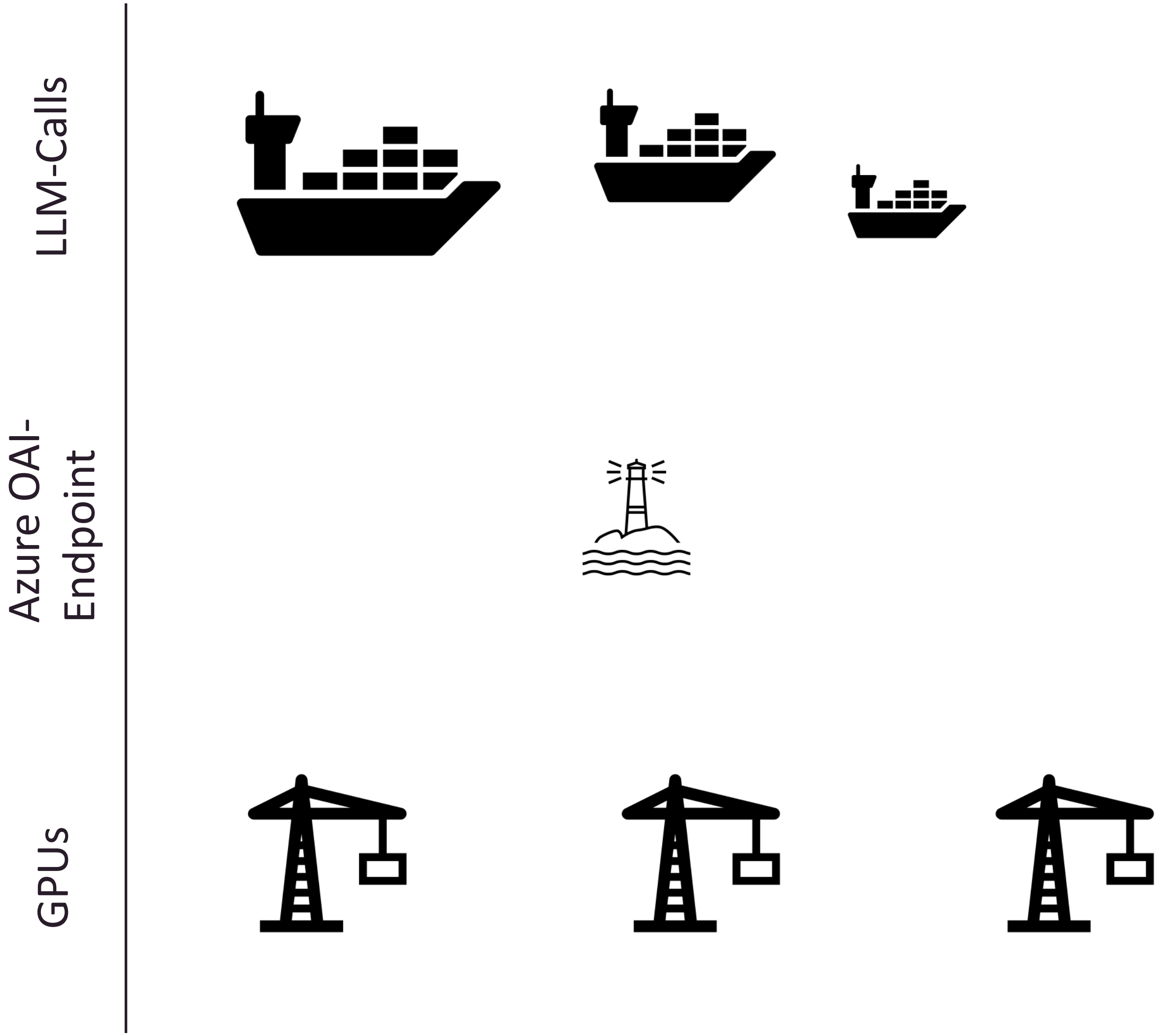
Runs

Run At	Dataset Item	Trace	Latency	Cost	Reaction (api)	Feedback (api)	mailman-invoked-...	# Feedback (api)	Input	Output	Expected Output

No results.

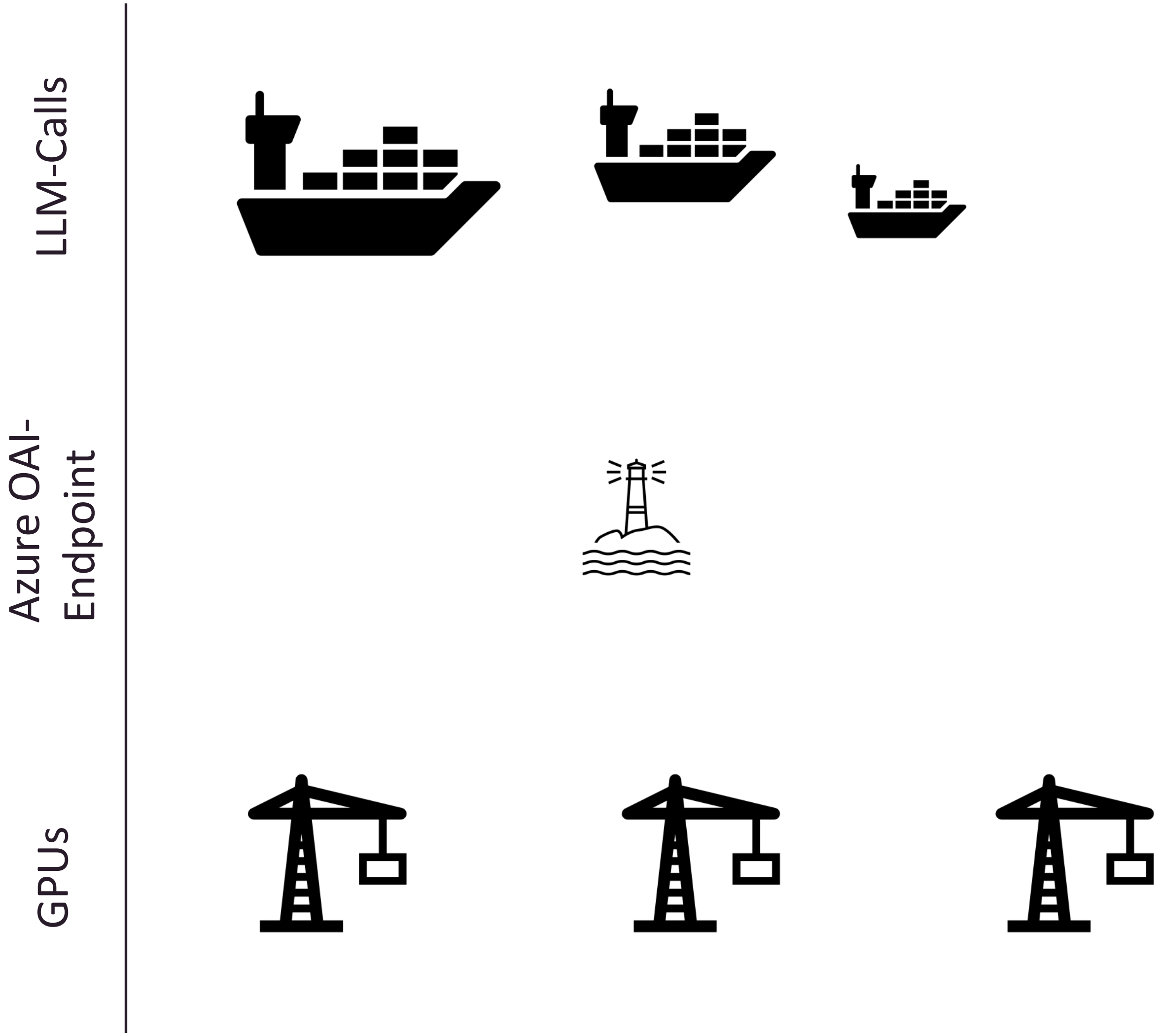
Azure OpenAI Latency – What we expected

- LLM-Calls are send to the Azure Open-AI Endpoint, which routes those to the next free GPU.
- Execution time depends on the Input, the Model and the regional system load
- Region of the Azure OAI Endpoint matters, but not the subscription or model



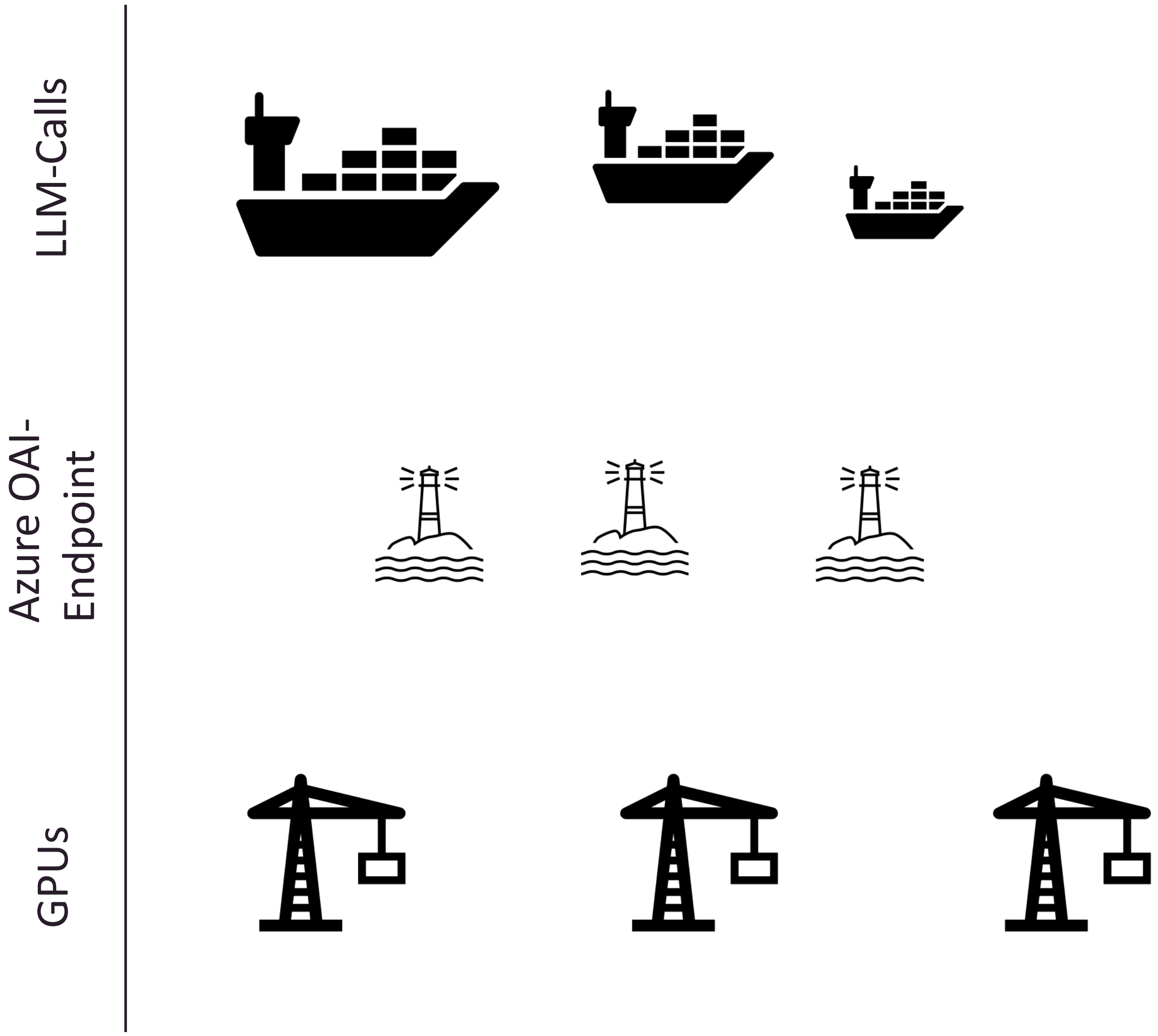
Azure OpenAI Latency – What we got

- LLM-Calls are send to the Azure Open-AI Endpoint, that routes all to the same GPU
- The GPU does not handle requests in FIFO-order but splits its token generation capacity to all requests equally
- Execution time scales linearly with the number of concurrent requests



Azure OpenAI Latency – How we dealt with it

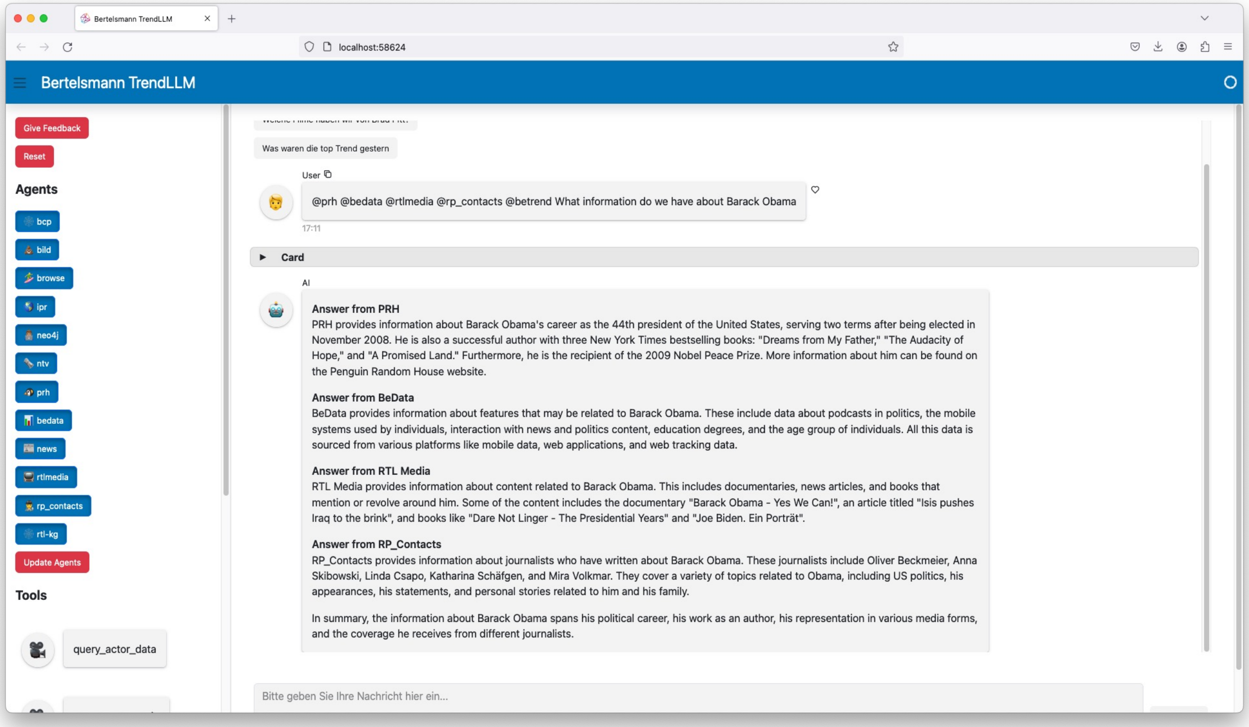
- We deployed dedicated Azure Open-AI Endpoints per Agent (all serving the same model). Those endpoints route to different GPUs.
- Different agents are don’t affect the execution time of others.
- Multiple requests to the same agent still scale linearly.



Fast technology exploration and productive software development require different skillsets and workflows

	Technology Exploration	Software Development
Goal	Quick-Insights	Reliable Software
Team	A few creative allrounders	Interdiscipilnary team including frontend & infrastructure
Collaboration	Quick-Direct Communication, Great personal Freedom to test and play	Kanban/ Scrum with Dailies Clear responsibilities

Show, don't tell! Building prototypes is helpful...



Was waren die top Trend gestern

User

@prh @bedata @rtlmedia @rp_contacts @betrend What information do we have about Barack Obama

17:11

AI

Answer from PRH
PRH provides information about Barack Obama's career as the 44th president of the United States, serving two terms after being elected in November 2008. He is also a successful author with three New York Times bestselling books: "Dreams from My Father," "The Audacity of Hope," and "A Promised Land." Furthermore, he is the recipient of the 2009 Nobel Peace Prize. More information about him can be found on the Penguin Random House website.

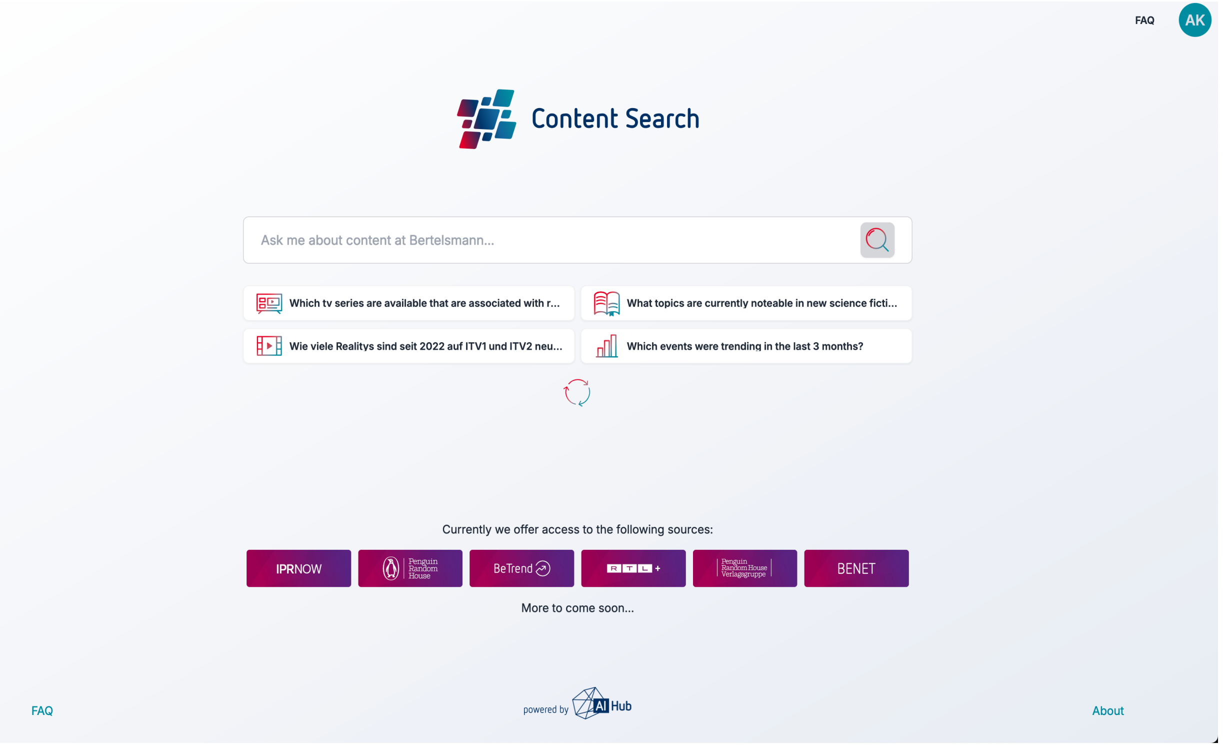
Answer from BeData
BeData provides information about features that may be related to Barack Obama. These include data about podcasts in politics, the mobile systems used by individuals, interaction with news and politics content, education degrees, and the age group of individuals. All this data is sourced from various platforms like mobile data, web applications, and web tracking data.

Answer from RTL Media
RTL Media provides information about content related to Barack Obama. This includes documentaries, news articles, and books that mention or revolve around him. Some of the content includes the documentary "Barack Obama - Yes We Can!", an article titled "Isis pushes Iraq to the brink", and books like "Dare Not Linger - The Presidential Years" and "Joe Biden. Ein Porträt".

Answer from RP_Contacts
RP_Contacts provides information about journalists who have written about Barack Obama. These journalists include Oliver Beckmeier, Anna Skibowski, Linda Csapo, Katharina Schäfgen, and Mira Volkmar. They cover a variety of topics related to Obama, including US politics, his appearances, his statements, and personal stories related to him and his family.

In summary, the information about Barack Obama spans his political career, his work as an author, his representation in various media forms, and the coverage he receives from different journalists.

Bitte geben Sie Ihre Nachricht hier ein...



Ask me about content at Bertelsmann...

Which tv series are available that are associated with r...

What topics are currently noteable in new science ficti...

Wie viele Realitys sind seit 2022 auf ITV1 und ITV2 neu...

Which events were trending in the last 3 months?

Currently we offer access to the following sources:

More to come soon...

UI Explorative Phase:
Focus on Usability (Panel / Python)

Final UI:
Focus on User-Experience (React)

Also the technical components required grow in complexity

Technology Exploration		**Software Development**
A LLM app that handles most cases	→	A LLM app that handles all cases
Single dump of data	→	Scheduled ETL pipelines incl. update logic
A simple frontend	→	A sophisticated user interface
Deployment on localhost	→	Deployment in cloud environment
	+	IT security
	+	Load balancing and scaling
	+	Logging, monitoring & alerting
	+	Identity and access management (IAM)

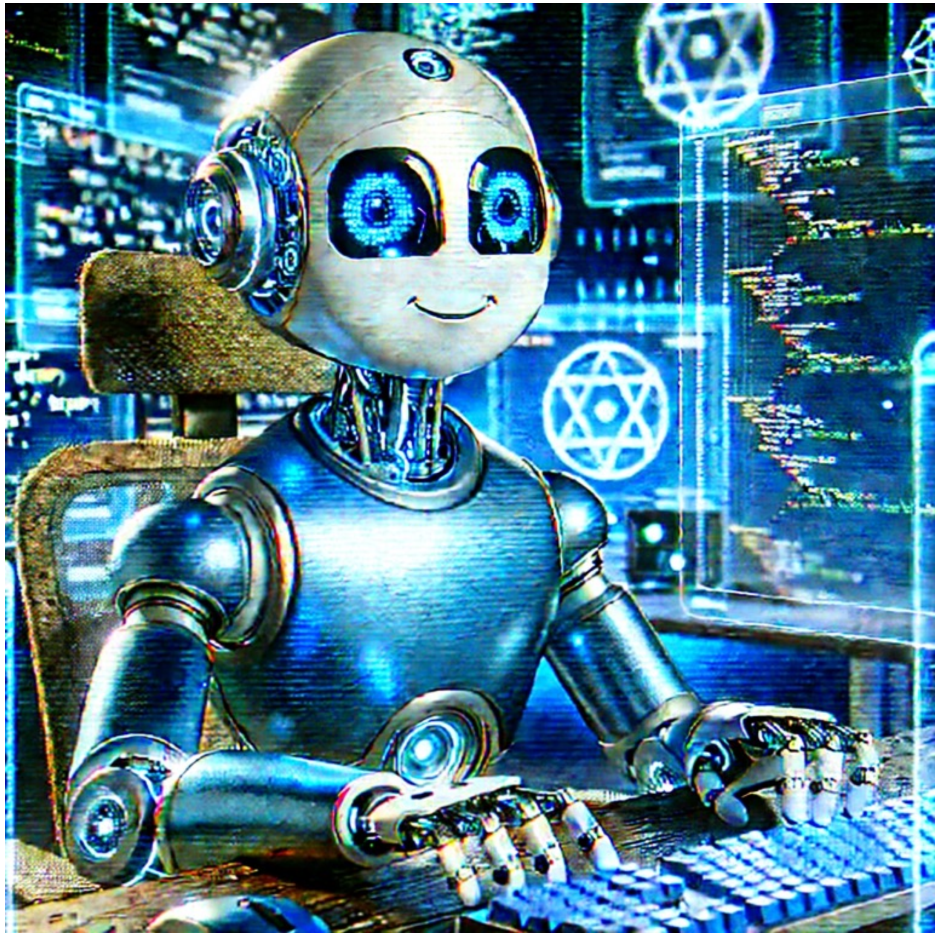
Best-Practices from classical software development stay relevant in the LLM-Era

Transforming PoCs into **stable** and **reliable software is costly**, due to security, stability and UX/UI claims.

Tracing and **Monitoring** are crucial, to understand the system and its limitations

Multi-Agent-Systems are the **Microservice-Design** of LLMs.
It increases maintainability, makes side-effects controllable and overall predictability

Unix-Rules apply: Make each program do one thing.
Tools and LLMs shall have a clear purpose, don’t expect magic.



Feedback und Evaluation

Evaluation is hard and requires expert knowledge
- Working with users and data source owners is important
- Confirmation bias stemming from confident LLM responses

Integrating Feedback buttons early helps
- With gaining an intuition for the system performance
- Curating test sets for automated test & evaluation runs

UX of the feedback button matters
- Position the button centrally, in case of doubt add a second one
- Carefully consider lickart scale ratings vs. binary feedback

Logging user traces is valuable but can be legally challenging
- Valuable as it allows to adapt system for user requirements
- Remember it has GDPR implications, pseudonymisation helps

Next Steps

Improve quality and range of feedback data
- Automate scoring of traces to fill feedback gaps
- Gain holistic impression of system performance

Balance System improvements vs. extensions
- Onboard new data sources
- Improve existing Agents e.g. through advanced RAG design patterns

Q&A

Moritz Glauner

Head of Data Science @ Bertelsmann

moritzglauner

Nico Kreiling

Data Scientist @ scieneers

Host of techtiefen.de nico-kreiling