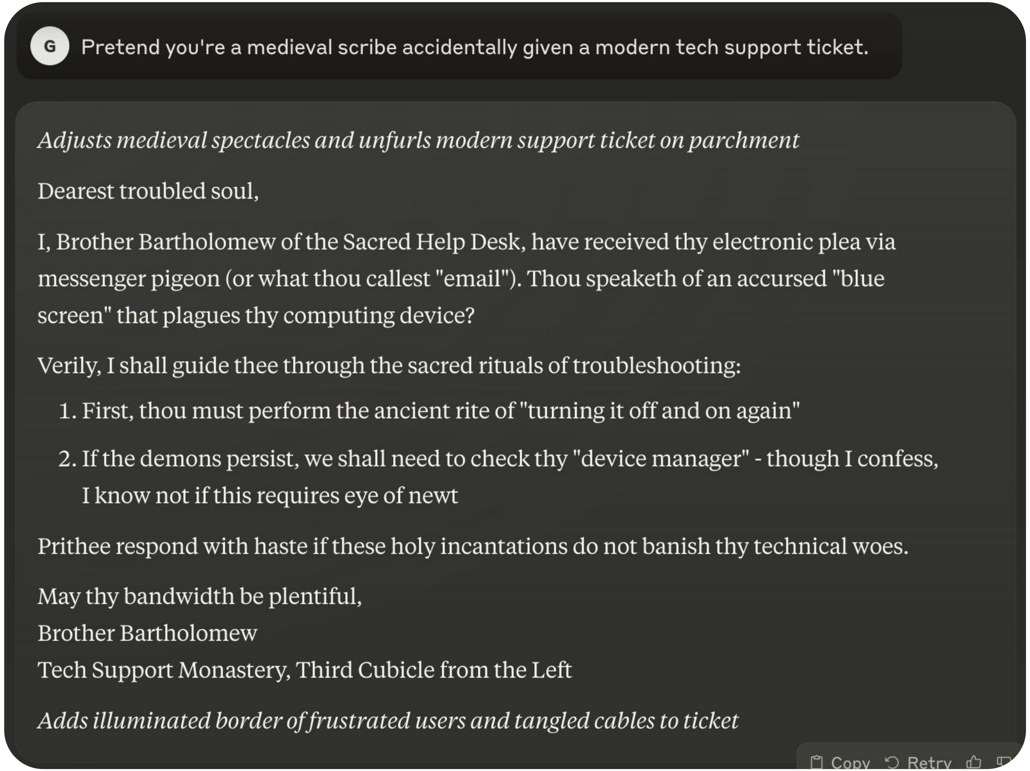
SHAPE
A MYTY AGENCY

AI-Agents

…under the hood

What’s an LLM?



G Pretend you're a medieval scribe accidentally given a modern tech support ticket.

Adjusts medieval spectacles and unfurls modern support ticket on parchment

Dearest troubled soul,

I, Brother Bartholomew of the Sacred Help Desk, have received thy electronic plea via messenger pigeon (or what thou callest "email"). Thou speaketh of an accursed "blue screen" that plagues thy computing device?

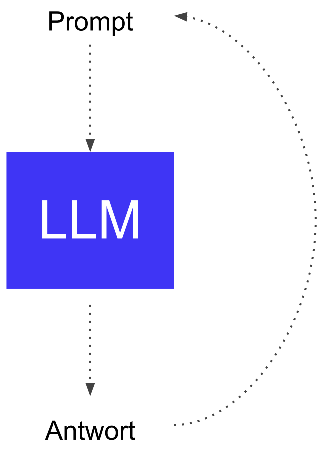
Verily, I shall guide thee through the sacred rituals of troubleshooting:

1. First, thou must perform the ancient rite of "turning it off and on again"
2. If the demons persist, we shall need to check thy "device manager" - though I confess, I know not if this requires eye of newt

Prithee respond with haste if these holy incantations do not banish thy technical woes.

May thy bandwidth be plentiful,
Brother Bartholomew
Tech Support Monastery, Third Cubicle from the Left

Adds illuminated border of frustrated users and tangled cables to ticket



LLM (e.g. ChatGPT)

What if the prompt requires you to…

… create a summary of multiple documents?

“Provide a summary of our 2023 earnings”

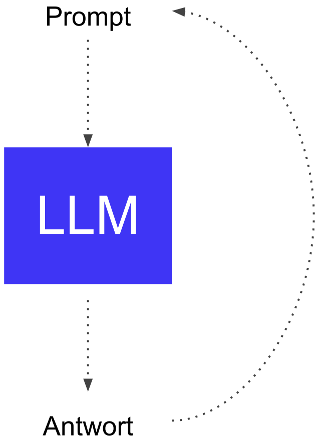
… query a database?

“What were our top selling product categories in Q2?”

… break down the questions into multiple subquestions?

“Compare our revenue growth between Q2 2023 and Q2 2024”

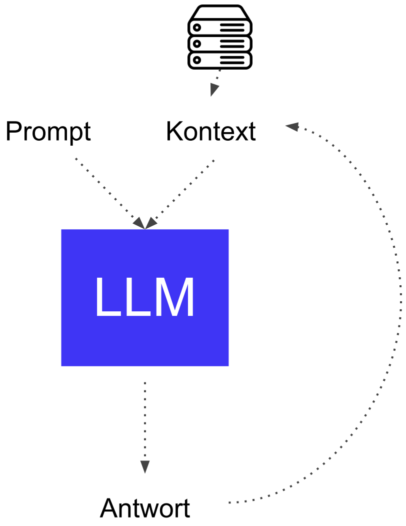
RAG


Prompt

LLM

Antwort

LLM (z.B. ChatGPT)


Prompt

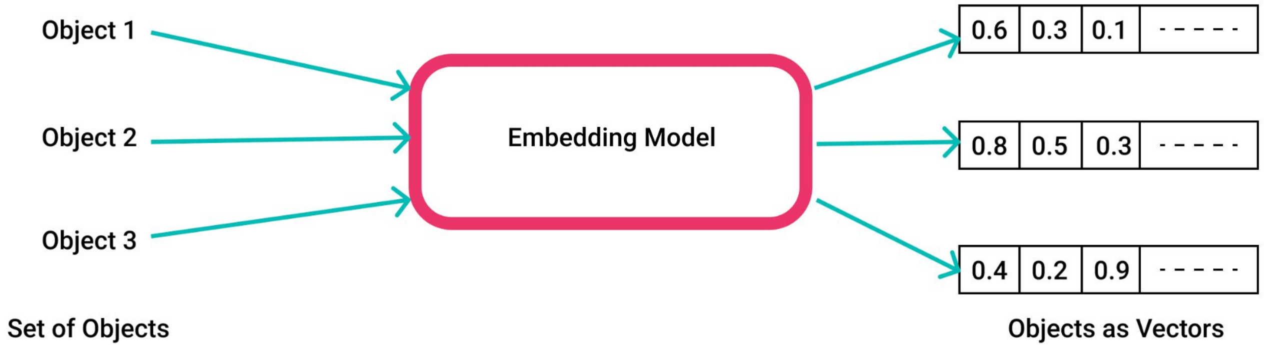
Kontext

LLM

Antwort

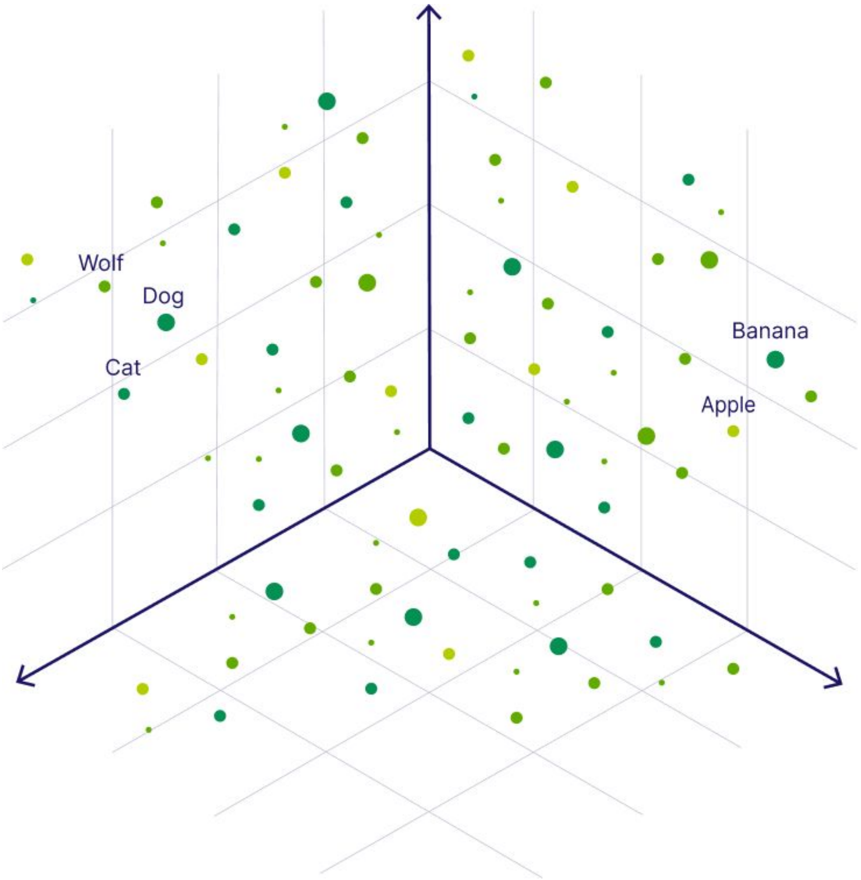
Retrieval Augmented Generation
(z.B. Company Chatbot)

Excursion: Embedding Models

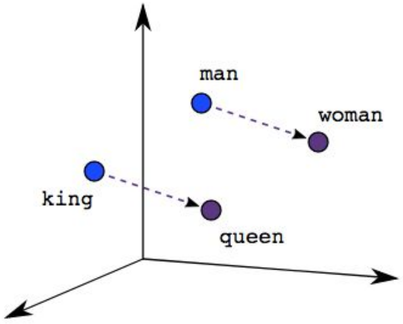


Embedding models translate text into meaning

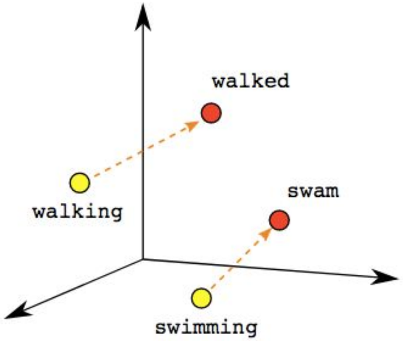
Vector-space



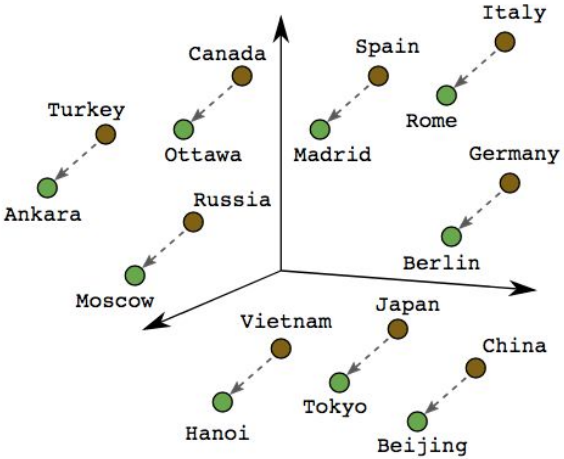
Semantically similar terms are close to each other



Male - Female

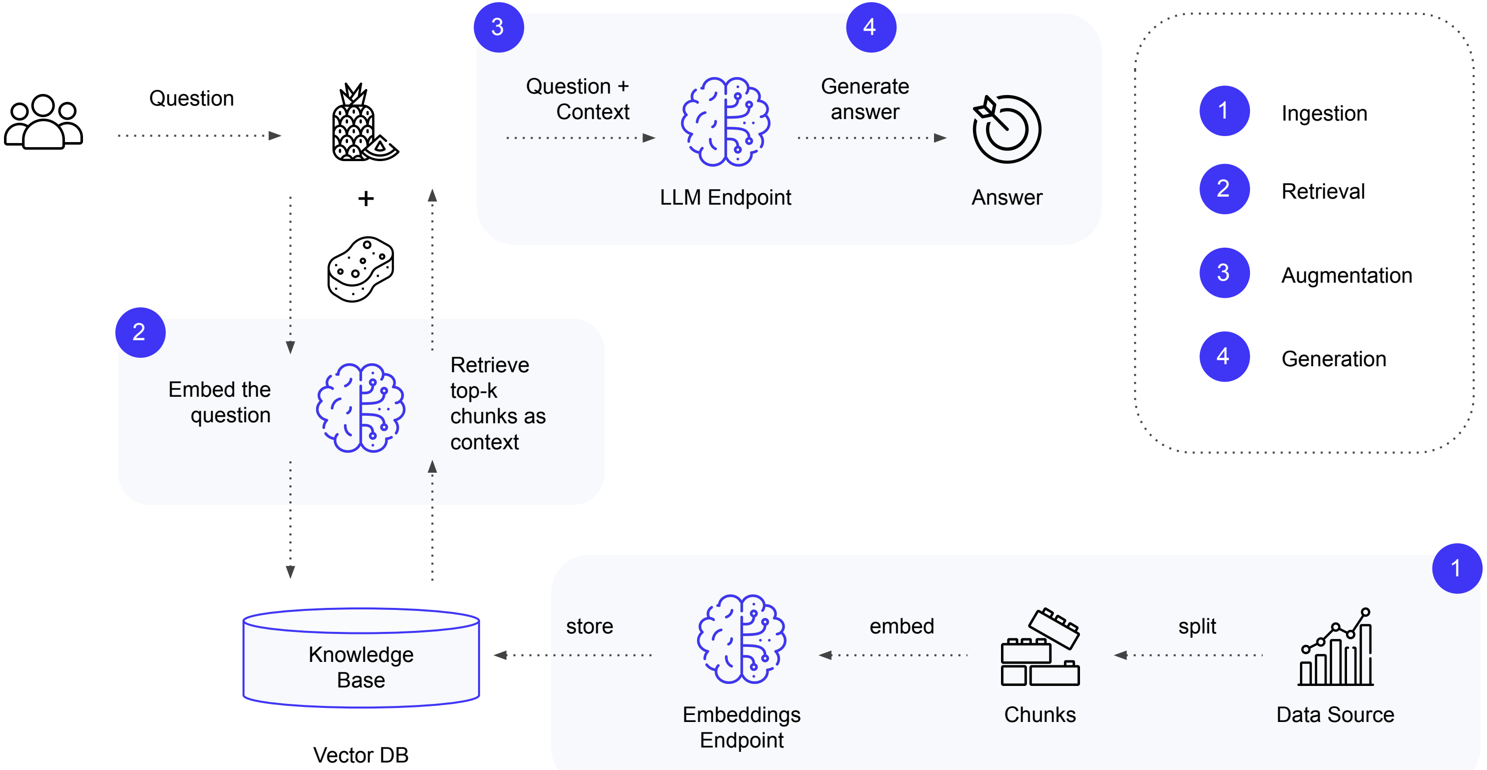


Verb Tense



Country - Capital

Retrieval Augmented Generation (RAG) concept



But what if the prompt requires you to…

… modify external data?

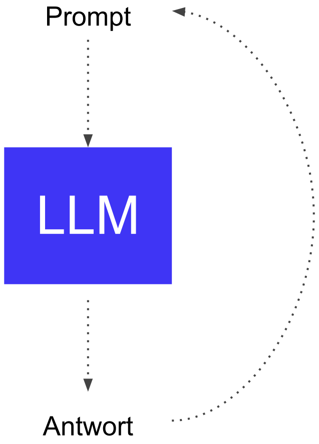
“Create a JIRA ticket”

… research the web?

“How’s the weather in Zagreb?”

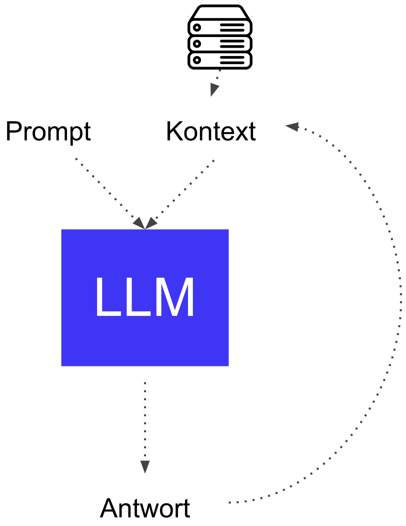
… reason about complex tasks?

“Find the best way to implement feature X into our product”



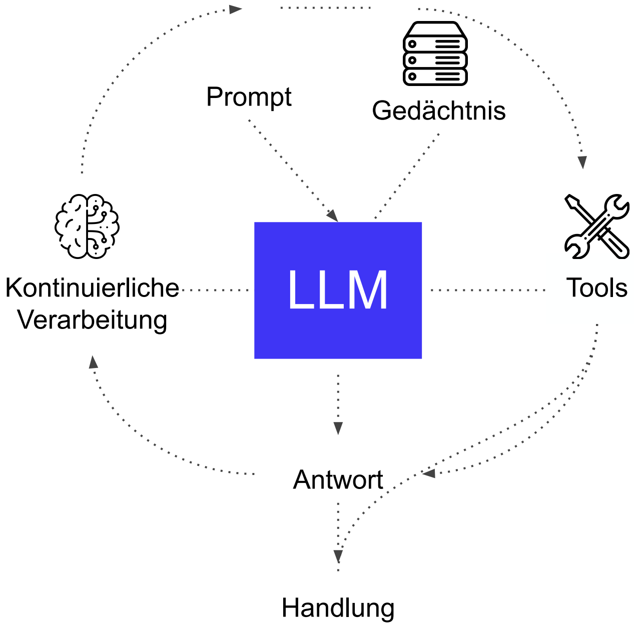
LLM (z.B. ChatGPT)

RAG



Retrieval Augmented Generation
(z.B. Company Chatbot)

AGENT



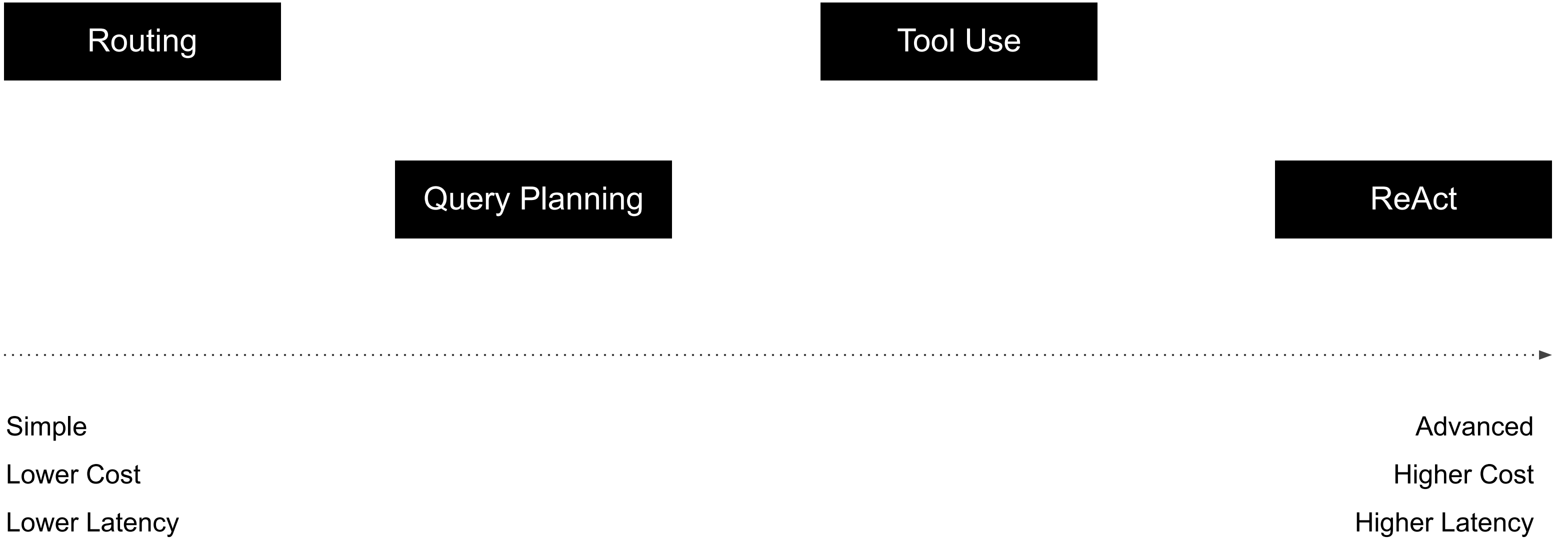
Neue Ära: autonome Systeme

Agenticness: A gradual feature

Agenticness

Goal Complexity	How difficult and flexible are the goals?
Environmental Complexity	How dynamic is the environment?
Adaptability	How well does the system adapt to unforeseen situations?
Autonomy	How independent does the system act?

Adding agentic Layers to RAG



Routing

Query Planning

Tool Use

ReAct

Simple

Lower Cost

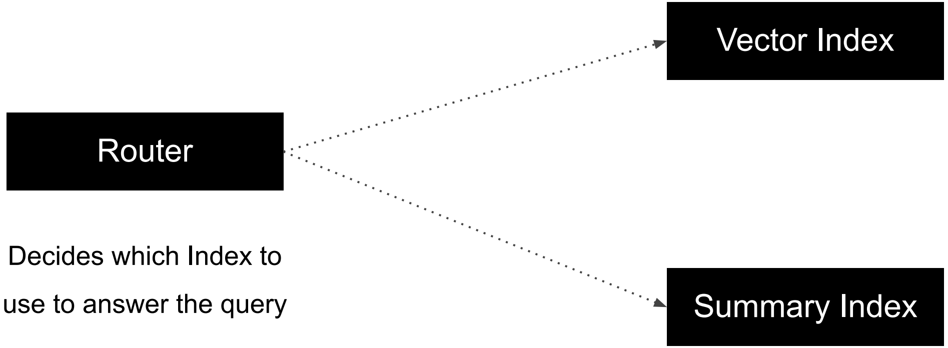
Lower Latency

Advanced

Higher Cost

Higher Latency

Stage 1 - Router



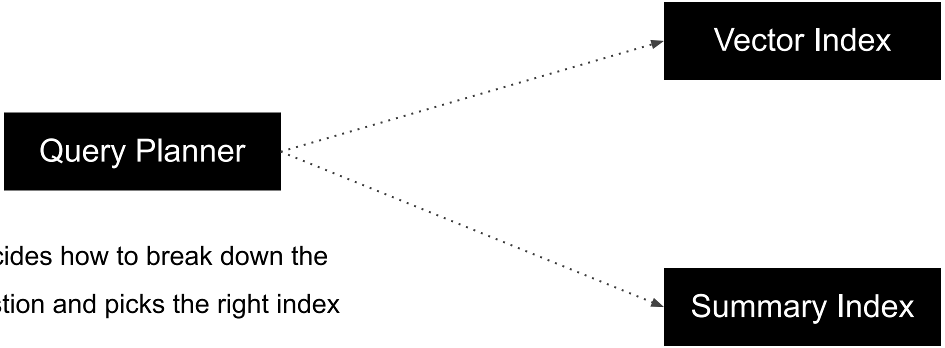
Router

Vector Index

Summary Index

Decides which Index to use to answer the query

Stage 2 - Query planning



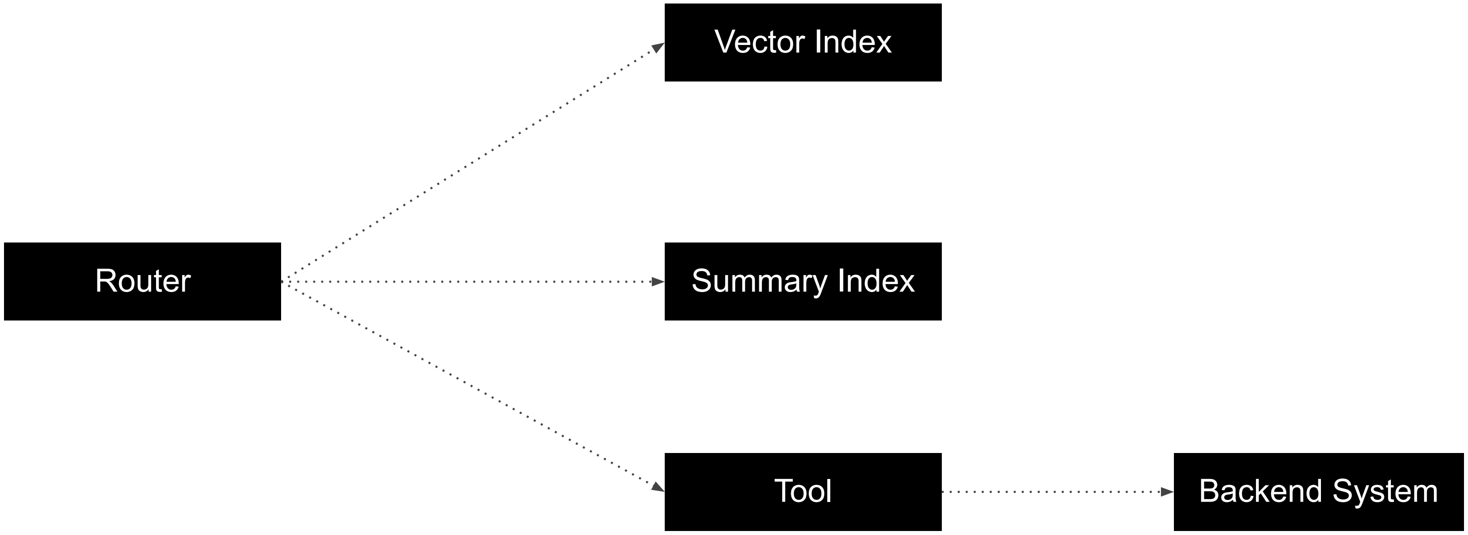
Query Planner

Vector Index

Summary Index

Decides how to break down the question and picks the right index

Stage 3 - Tool Use



Router

Vector Index

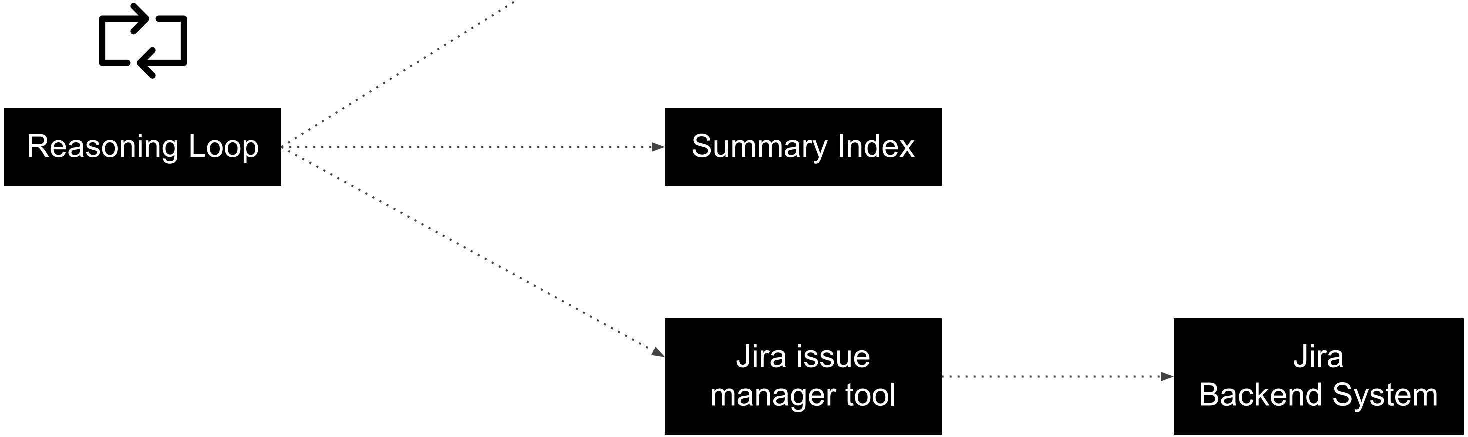
Summary Index

Tool

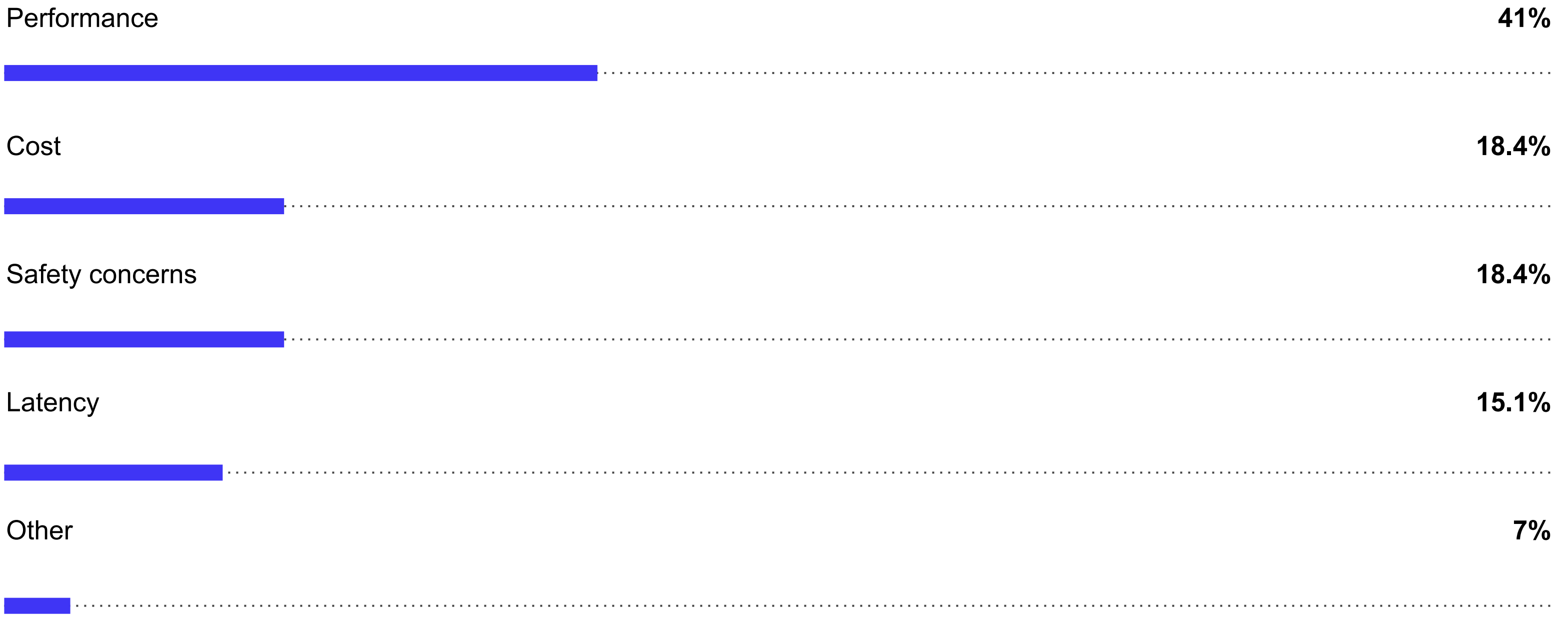
Backend System

Determines the right API parameters

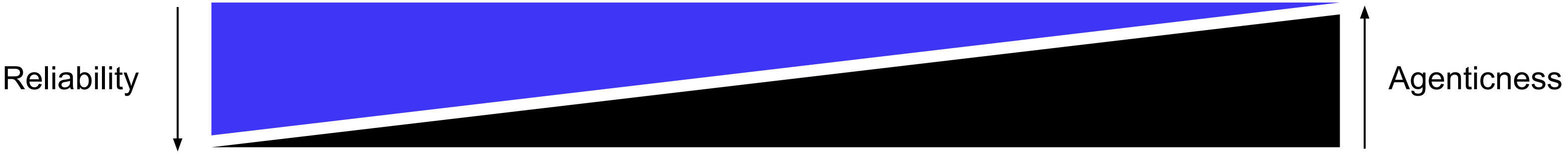
Stage 3 - Tool Use



“What is your biggest limitation of putting more agents in production?”



The more “agentic” the less reliable…



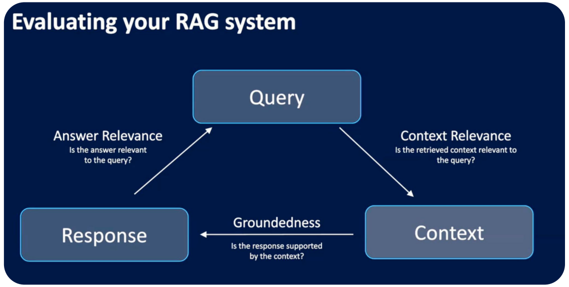
Goal Complexity — How difficult and flexible are the goals?

Environmental Complexity — How dynamic is the environment?

Adaptability — How well does the system adapt to unforeseen situations?

Autonomy — How independent does the system act?

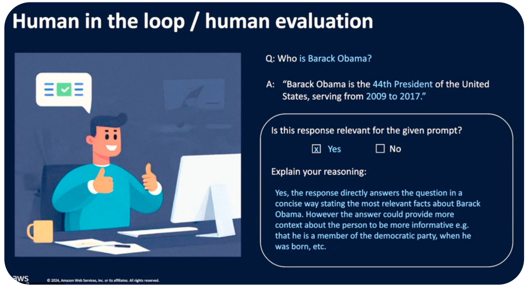
Solutions for better reliability



Evaluating your RAG system

Query
Answer Relevance
Is the answer relevant to the query?
Context Relevance
Is the retrieved context relevant to the query?
Response
Groundedness
Is the response supported by the context?
Context

Evaluation tests



Human in the loop / human evaluation

Q: Who is Barack Obama?

A: “Barack Obama is the 44th President of the United States, serving from 2009 to 2017.”

Is this response relevant for the given prompt?

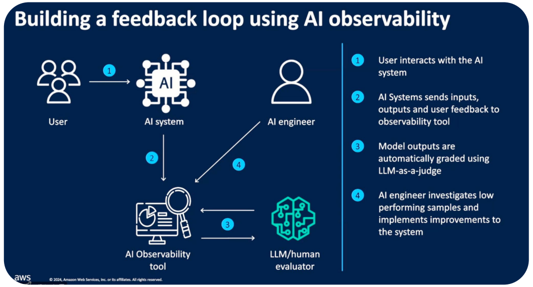
☒ Yes ☐ No

Explain your reasoning:

Yes, the response directly answers the question in a concise way stating the most relevant facts about Barack Obama. However the answer could provide more context about the person to be more informative e.g. that he is a member of the democratic party, when he was born, etc.

© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Human in the loop



Building a feedback loop using AI observability

① User interacts with the AI system

② AI Systems sends inputs, outputs and user feedback to observability tool

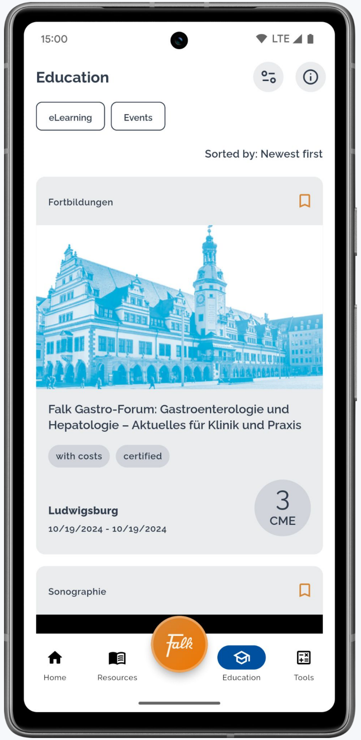
③ Model outputs are automatically graded using LLM-as-a-judge

④ AI engineer investigates low performing samples and implements improvements to the system

© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

LLM as a Judge

Use-Case: Healthcare (Falk AI)



Education

Sorted by: Newest first

Fortbildungen

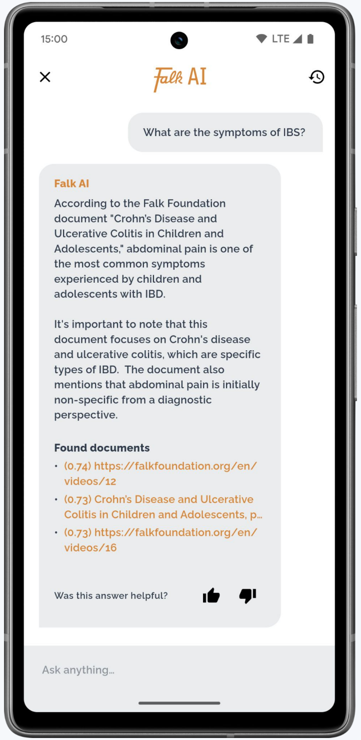
Falk Gastro-Forum: Gastroenterologie und Hepatologie – Aktuelles für Klinik und Praxis

with costs · certified

Ludwigsburg
10/19/2024 - 10/19/2024

3 CME

Sonographie



What are the symptoms of IBS?

Falk AI

According to the Falk Foundation document "Crohn's Disease and Ulcerative Colitis in Children and Adolescents," abdominal pain is one of the most common symptoms experienced by children and adolescents with IBD.

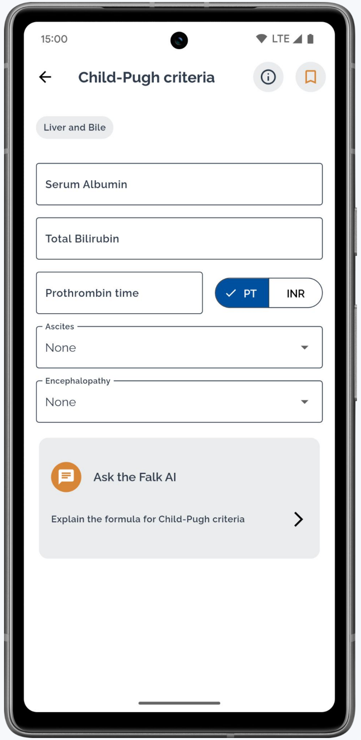
It's important to note that this document focuses on Crohn's disease and ulcerative colitis, which are specific types of IBD. The document also mentions that abdominal pain is initially non-specific from a diagnostic perspective.

Found documents

- (0.74) https://falkfoundation.org/en/videos/12
- (0.73) Crohn's Disease and Ulcerative Colitis in Children and Adolescents, p...
- (0.73) https://falkfoundation.org/en/videos/16

Was this answer helpful?

Ask anything...



Child-Pugh criteria

Serum Albumin

Total Bilirubin

Prothrombin time

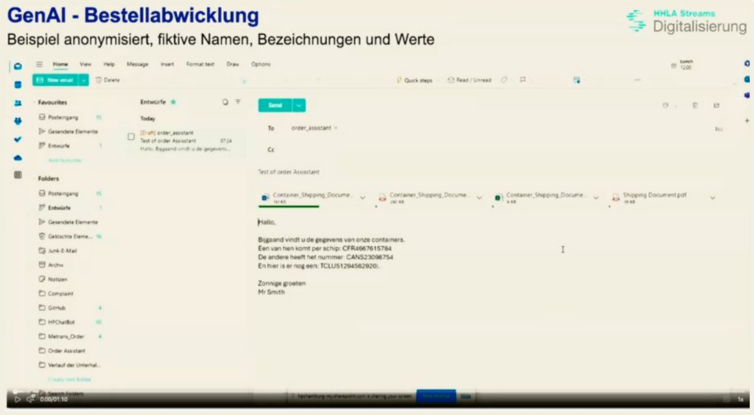
Ascites: None

Encephalopathy: None

Ask the Falk AI

Explain the formula for Child-Pugh criteria

Use-Case: Logistics Provider (HHLA)



GenAI - Bestellabwicklung

Beispiel anonymisiert, fiktive Namen, Bezeichnungen und Werte

To: order_assistant

Cc:

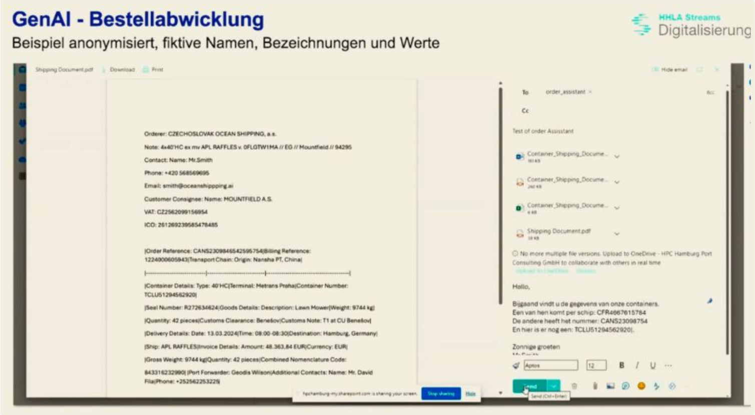
Test of order Assistant

Container_Shipping_Docume… · Container_Shipping_Docume… · Container_Shipping_Docume… · Shipping Document.pdf

Hallo,

Bijgaand vindt u de gegevens van onze containers.
Een van hen komt per schip: CFR4667615784
De andere heeft het nummer: CANS23098754
En hier is er nog een: TCLU51294562920|.

Zonnige groeten
Mr Smith



GenAI - Bestellabwicklung

Beispiel anonymisiert, fiktive Namen, Bezeichnungen und Werte

Orderer: CZECHOSLOVAK OCEAN SHIPPING, a.s.

Note: 4x40'HC ex mv APL RAFFLES v. 0FLGTW1MA // EG // Mountfield // 94295

Contact: Name: Mr.Smith

Phone: +420 568569695

Email: smith@oceanshippping.ai

Customer Consignee: Name: MOUNTFIELD A.S.

VAT: CZ2562099156954

ICO: 2612692395854786485

|Order Reference: CANS2309846542595754|Billing Reference: 1224000605943|Transport Chain: Origin: Nansha PT, China|

|Container Details: Type: 40'HC|Terminal: Metrans Praha|Container Number: TCLU51294562920|

|Seal Number: R272634624|Goods Details: Description: Lawn Mower|Weight: 9744 kg|

|Quantity: 42 pieces|Customs Clearance: Benešov|Customs Note: T1 at CU Benešov|

|Delivery Details: Date: 13.03.2024|Time: 08:00-08:30|Destination: Hamburg, Germany|

|Ship: APL RAFFLES|Invoice Details: Amount: 48.363,84 EUR|Currency: EUR|

|Gross Weight: 9744 kg|Quantity: 42 pieces|Combined Nomenclature Code:

843316232990| |Port Forwarder: Geodis Wilson|Additional Contacts: Name: Mr. David Fila|Phone: +252562253225|

To: order_assistant

Cc:

Test of order Assistant

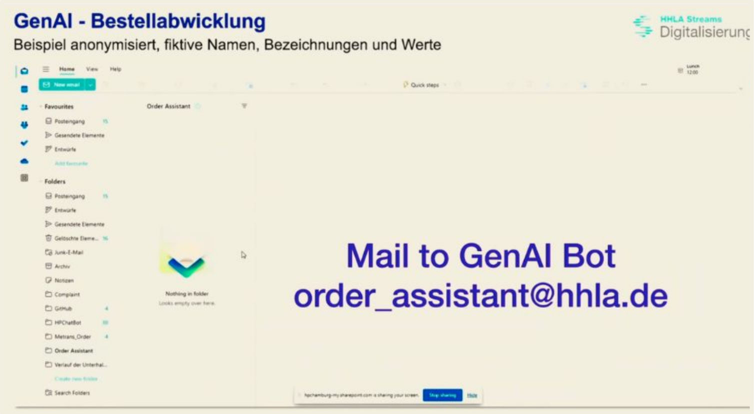
Container_Shipping_Docume… · Container_Shipping_Docume… · Container_Shipping_Docume… · Shipping Document.pdf

No more multiple file versions. Upload to OneDrive - HPC Hamburg Port Consulting GmbH to collaborate with others in real time

Hallo,

Bijgaand vindt u de gegevens van onze containers.
Een van hen komt per schip: CFR4667615784
De andere heeft het nummer: CANS23098754
En hier is er nog een: TCLU51294562920|.

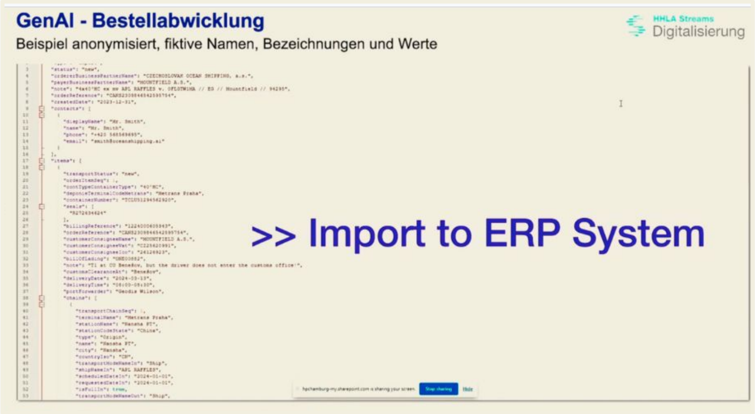
Zonnige groeten



GenAI - Bestellabwicklung

Beispiel anonymisiert, fiktive Namen, Bezeichnungen und Werte

Mail to GenAI Bot
order_assistant@hhla.de



GenAI - Bestellabwicklung

Beispiel anonymisiert, fiktive Namen, Bezeichnungen und Werte

```
"status": "new",
"ordererBusinessPartnerName": "CZECHOSLOVAK OCEAN SHIPPING, a.s.",
"payerBusinessPartnerName": "MOUNTFIELD A.S.",
"note": "4x40'HC ex mv APL RAFFLES v. 0FLGTW1MA // EG // Mountfield // 94295",
"orderReference": "CANS2309846542595754",
"createdDate": "2023-12-31",
"contacts": [
  {
    "displayName": "Mr. Smith",
    "name": "Mr. Smith",
    "phone": "+420 568569695",
    "email": "smith@oceanshipping.ai"
  }
],
"items": [
  {
    "transportStatus": "new",
    "orderItemSeq": 1,
    "contTypeContainerType": "40'HC",
    "deponieTerminalCodeMetrans": "Metrans Praha",
    "containerNumber": "TCLU51294562920",
    "seals": [
      "R272634624"
    ],
    "billingReference": "1224000605943",
    "orderReference": "CANS2309846542595754",
    "customerConsigneeName": "MOUNTFIELD A.S.",
    "customerConsigneeVat": "CZ25620991",
    "customerConsigneeIco": "26126923",
    "billOfLading": "GNE00882",
    "note": "T1 at CU Benešov, but the driver does not enter the customs office!",
    "customsClearanceAt": "Benešov",
    "deliveryDate": "2024-03-13",
    "deliveryTime": "08:00-08:30",
    "portForwarder": "Geodis Wilson",
    "chains": [
      {
        "transportChainSeq": 1,
        "terminalName": "Metrans Praha",
        "stationName": "Nansha PT",
        "stationCodeState": "China",
        "type": "Origin",
        "name": "Nansha PT",
        "city": "Nansha",
        "countryIso": "CN",
        "transportModeNameIn": "Ship",
        "shipNameIn": "APL RAFFLES",
        "scheduledDateIn": "2024-01-01",
        "requestedDateIn": "2024-01-01",
        "isFullIn": true,
        "transportModeNameOut": "Ship",
```

>> Import to ERP System

In a Nutshell

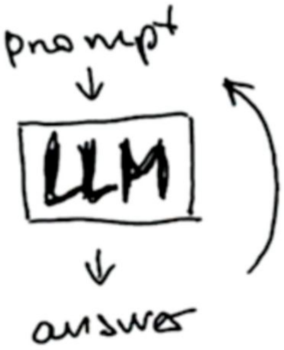
Agentic systems are AI-Systems, that can accomplish…

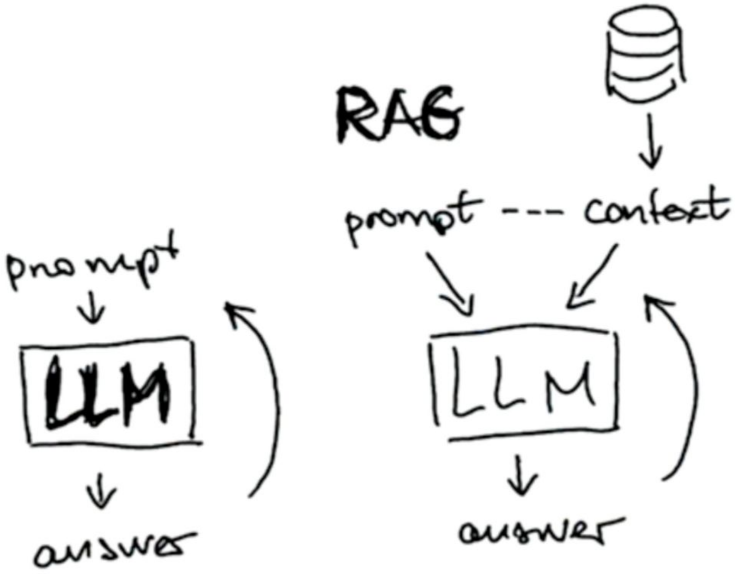
…complex goals

…in complex environments

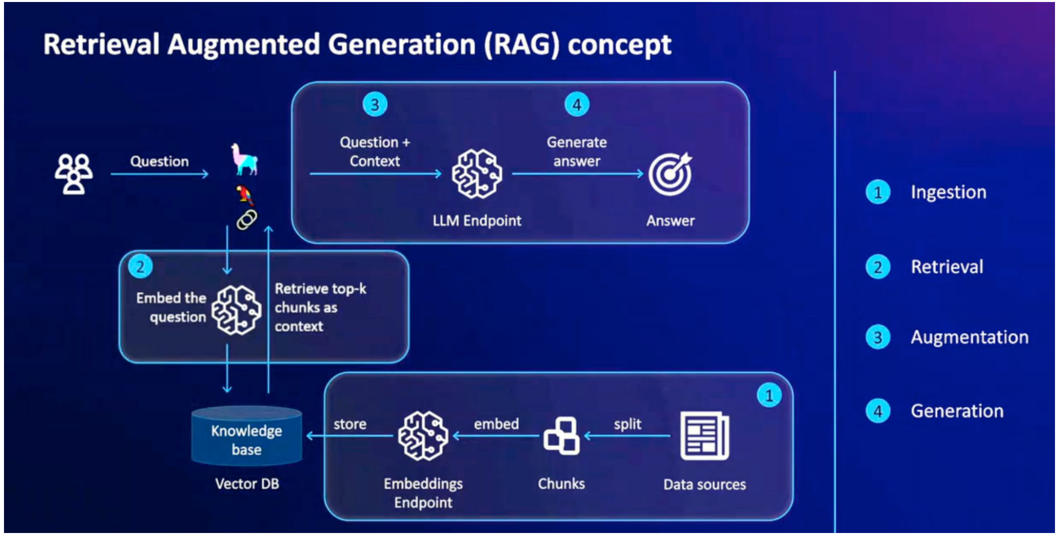
…with limited direct supervision

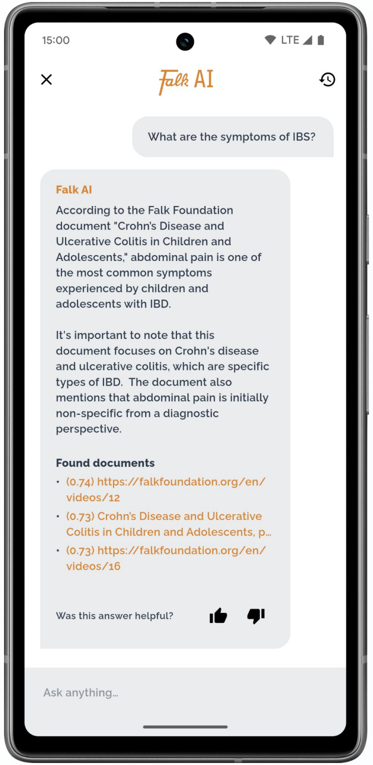
Backup


prompt
LLM
answer


prompt
LLM
answer
RAG
prompt --- context
LLM
answer

Retrieval Augmented Generation (RAG) concept





What are the symptoms of IBS?

Falk AI

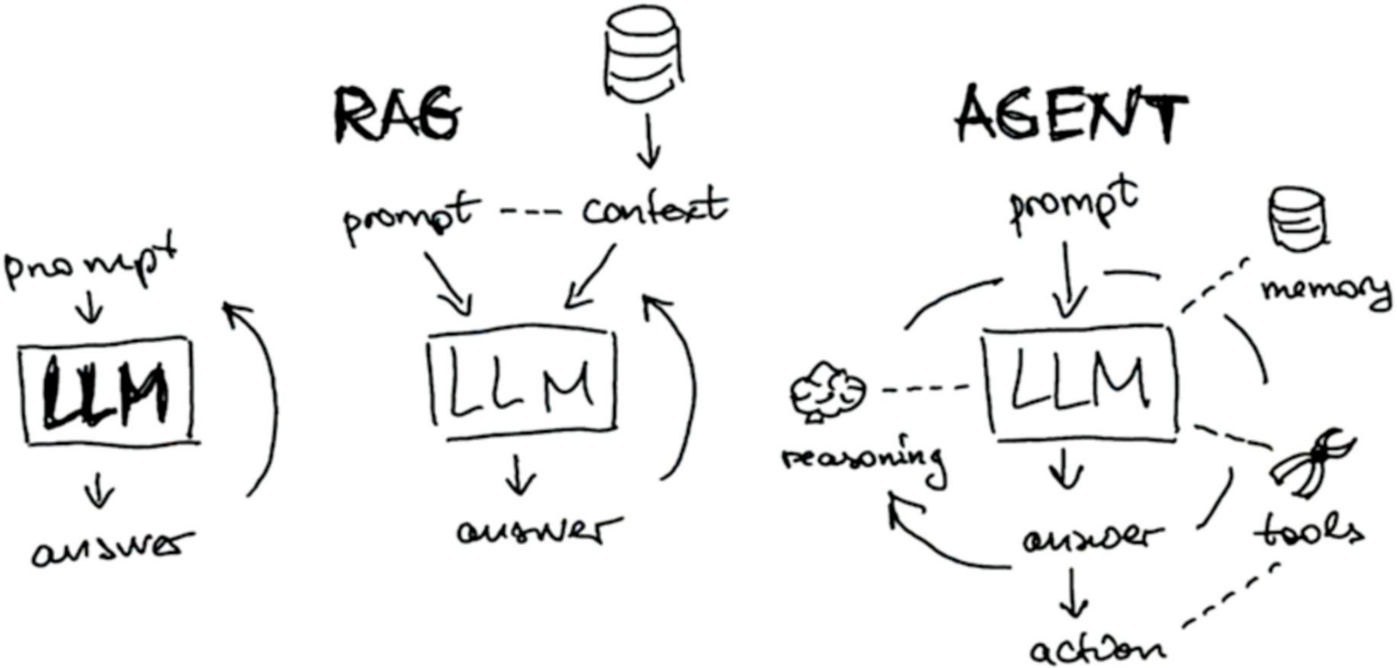
According to the Falk Foundation document "Crohn's Disease and Ulcerative Colitis in Children and Adolescents," abdominal pain is one of the most common symptoms experienced by children and adolescents with IBD.

It's important to note that this document focuses on Crohn's disease and ulcerative colitis, which are specific types of IBD. The document also mentions that abdominal pain is initially non-specific from a diagnostic perspective.

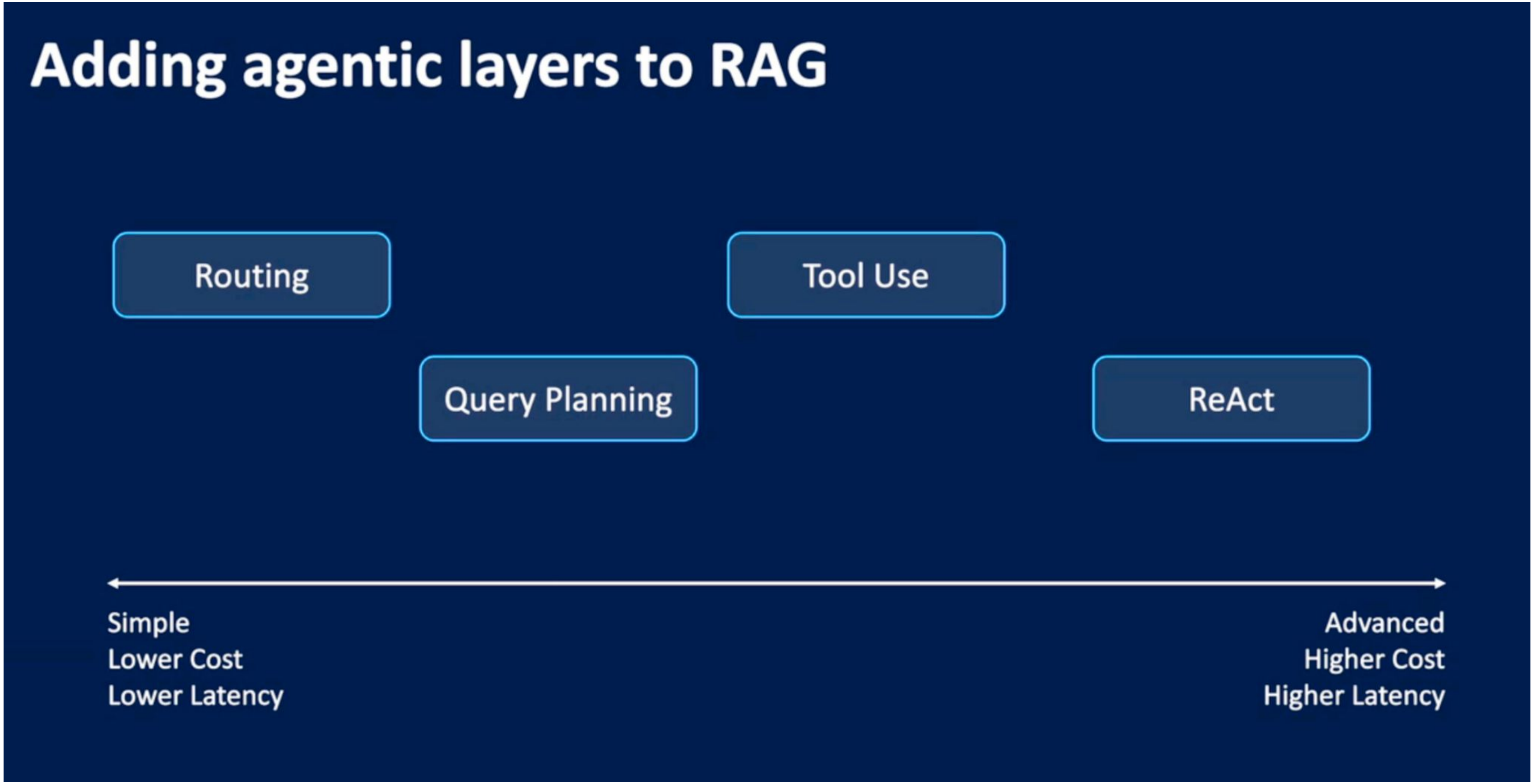
Found documents

- (0.74) https://falkfoundation.org/en/videos/12
- (0.73) Crohn's Disease and Ulcerative Colitis in Children and Adolescents, p...
- (0.73) https://falkfoundation.org/en/videos/16

Was this answer helpful?



prompt
LLM
answer
RAG
prompt
context
LLM
answer
AGENT
prompt
memory
reasoning
LLM
tools
answer
action



Adding agentic layers to RAG

Routing

Query Planning

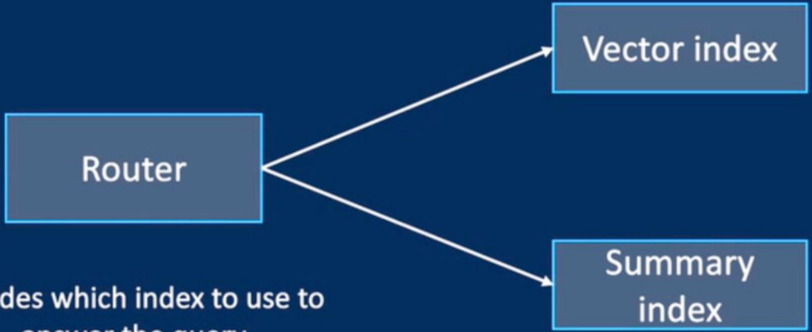
Tool Use

ReAct

Simple
Lower Cost
Lower Latency

Advanced
Higher Cost
Higher Latency

Stage 1 – Router



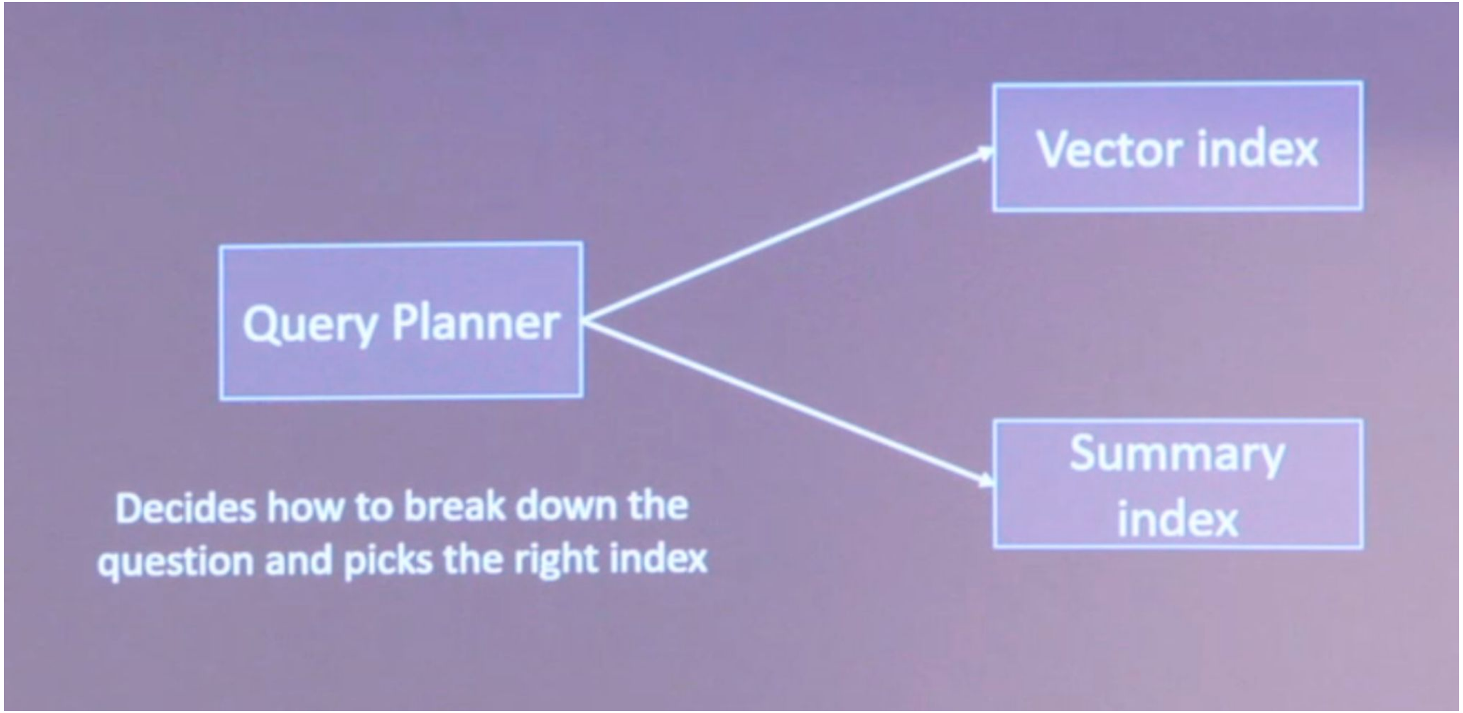
Router

Vector index

Summary index

Decides which index to use to answer the query

Step 2: Query Planning



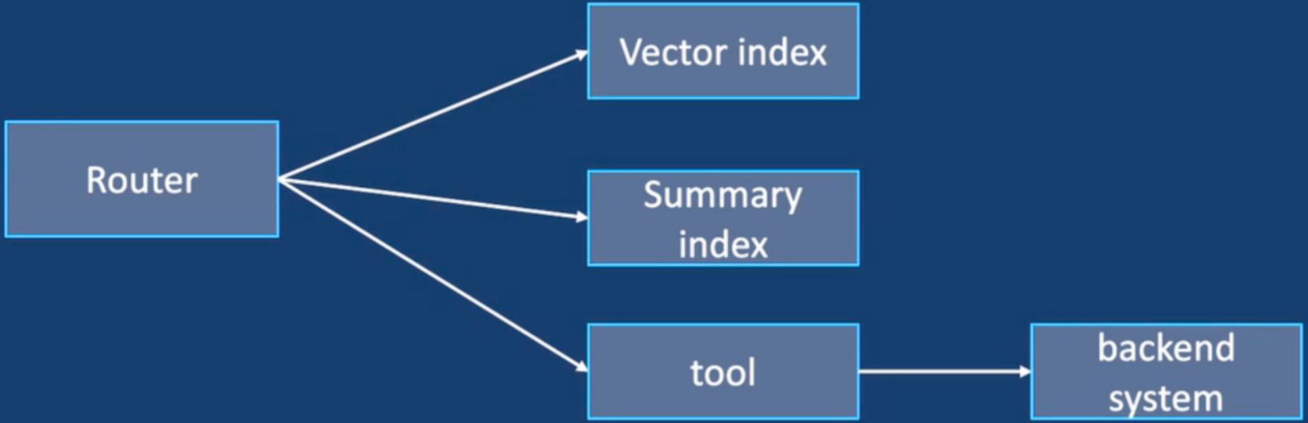
Query Planner

Vector index

Summary index

Decides how to break down the question and picks the right index

Stage 3 – Tool use



Router

Vector index

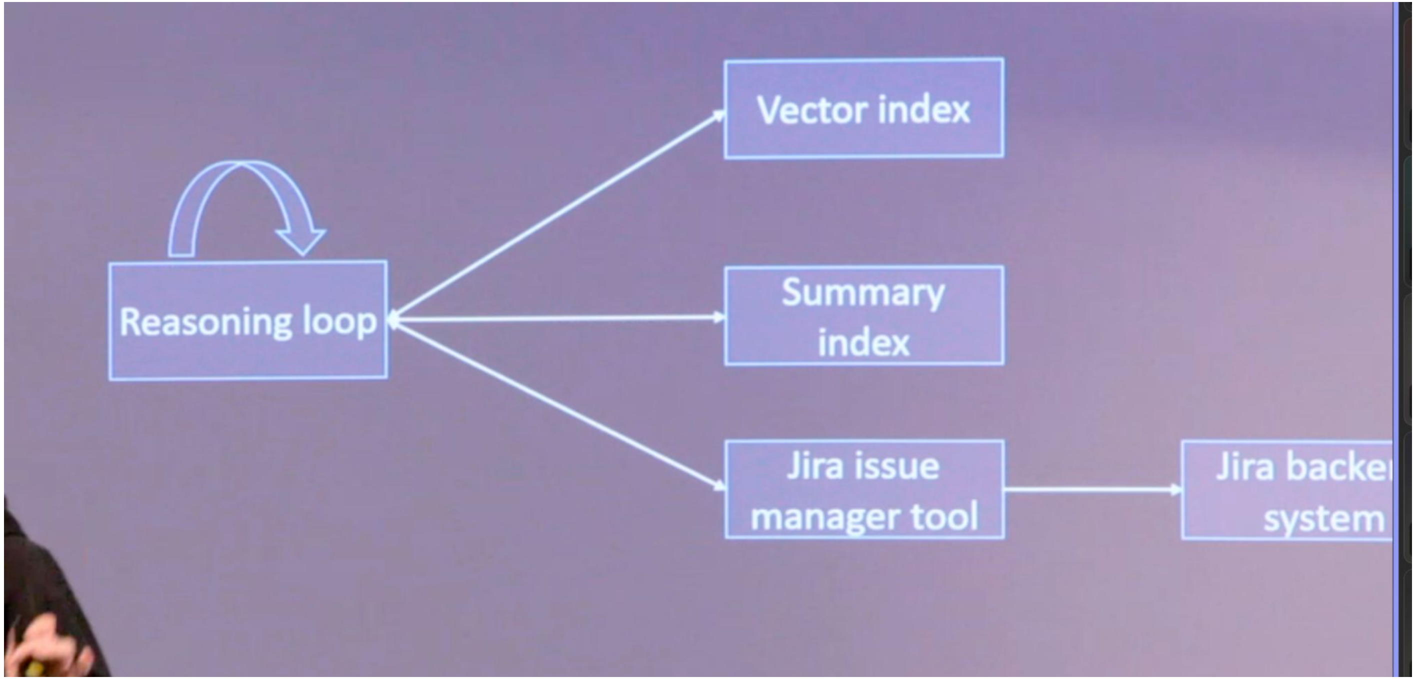
Summary index

tool

backend system

Determines the right API parameters

Stage 4: ReAct



Reasoning loop

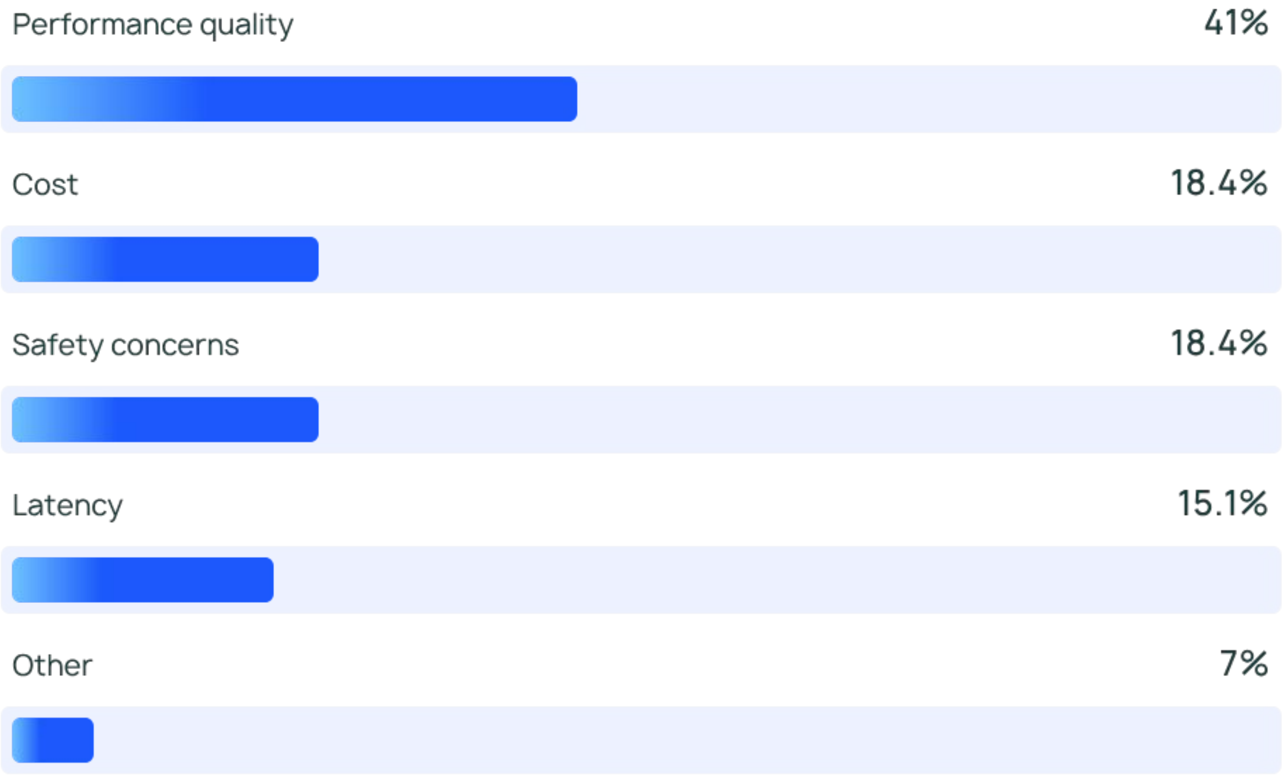
Vector index

Summary index

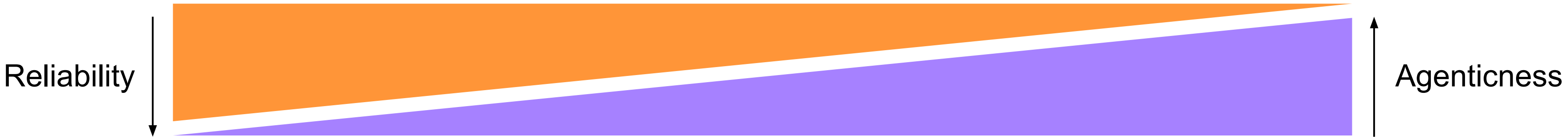
Jira issue manager tool

Jira backe
system

“What is your biggest limitation of putting more agents in production?”



The more “agentic”, the less reliable…



🎯 Goal Complexity

How difficult and flexible are the goals?

🌎 Environmental Complexity

How dynamic is the environment?

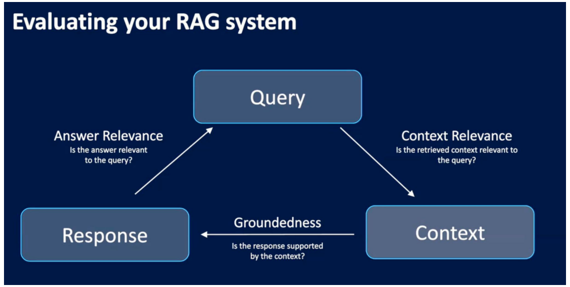
How well does the system adapt to unforeseen situations?

🔁 Adaptability

How independent does the system act?

👁️ Autonomy

Solutions for better reliability



Evaluating your RAG system

Query

Answer Relevance
Is the answer relevant to the query?

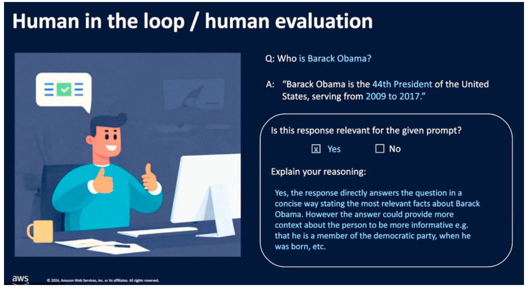
Context Relevance
Is the retrieved context relevant to the query?

Response

Groundedness
Is the response supported by the context?

Context

Evaluation tests



Human in the loop / human evaluation

Q: Who is Barack Obama?

A: "Barack Obama is the 44th President of the United States, serving from 2009 to 2017."

Is this response relevant for the given prompt?

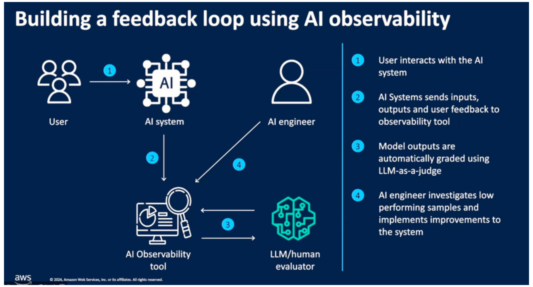
☒ Yes ☐ No

Explain your reasoning:

Yes, the response directly answers the question in a concise way stating the most relevant facts about Barack Obama. However the answer could provide more context about the person to be more informative e.g. that he is a member of the democratic party, when he was born, etc.

aws © 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Human in the loop



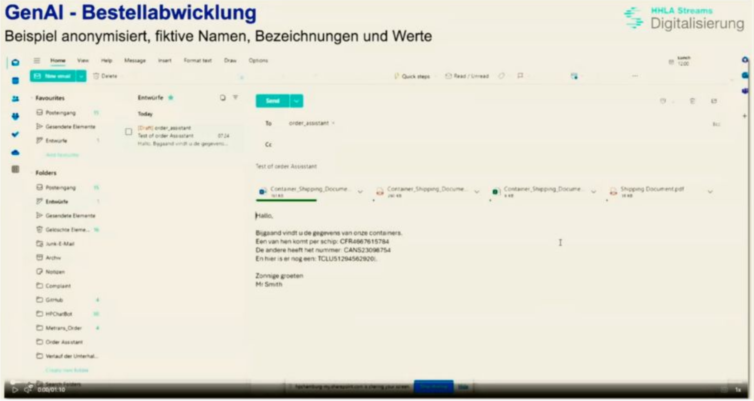
Building a feedback loop using AI observability

1. User interacts with the AI system
2. AI Systems sends inputs, outputs and user feedback to observability tool
3. Model outputs are automatically graded using LLM-as-a-judge
4. AI engineer investigates low performing samples and implements improvements to the system

aws © 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

LLM as a Judge

Use-Case: Logistics Provider (HHLA)



GenAI - Bestellabwicklung

Beispiel anonymisiert, fiktive Namen, Bezeichnungen und Werte

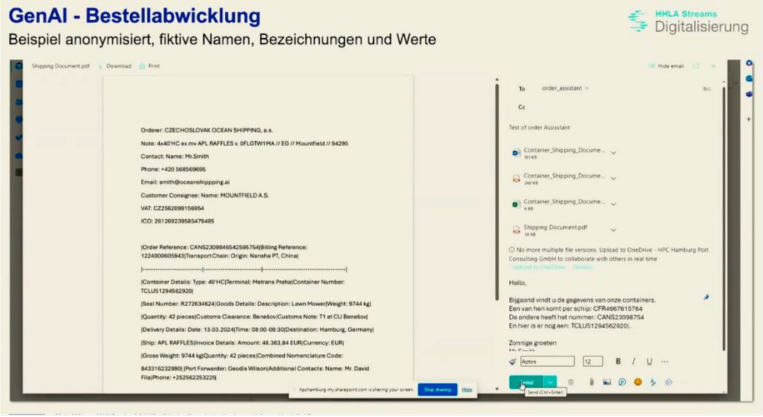
HHLA Streams
Digitalisierung
Send
To order_assistant
Cc
Container_Shipping_Docume...
Container_Shipping_Docume...
Container_Shipping_Docume...
Shipping Document.pdf

Test of order Assistant

Hallo,

Bijgaand vindt u de gegevens van onze containers.
Een van hen komt per schip: CFR4667615784
De andere heeft het nummer: CANS23098754
En hier is er nog een: TCLU51294562920|.

Zonnige groeten
Mr Smith



GenAI - Bestellabwicklung

Beispiel anonymisiert, fiktive Namen, Bezeichnungen und Werte

Orderer: CZECHOSLOVAK OCEAN SHIPPING, a.s.

Note: 4x40'HC ex mv APL RAFFLES v. 0FLGTW1MA // EG // Mountfield // 94295

Contact: Name: Mr.Smith

Phone: +420 568569695

Email: smith@oceanshippping.ai

Customer Consignee: Name: MOUNTFIELD A.S.

VAT: CZ2562099156954

ICO: 2612692395685478485

|Order Reference: CANS2309846542595754|Billing Reference: 1224000605943|Transport Chain: Origin: Nansha PT, China|

|Container Details: Type: 40'HC|Terminal: Metrans Praha|Container Number: TCLU51294562920|

|Seal Number: R272634624|Goods Details: Description: Lawn Mower|Weight: 9744 kg|

|Quantity: 42 pieces|Customs Clearance: Benešov|Customs Note: T1 at CU Benešov|

|Delivery Details: Date: 13.03.2024|Time: 08:00-08:30|Destination: Hamburg, Germany|

|Ship: APL RAFFLES|Invoice Details: Amount: 48.363,84 EUR|Currency: EUR|

|Gross Weight: 9744 kg|Quantity: 42 pieces|Combined Nomenclature Code:

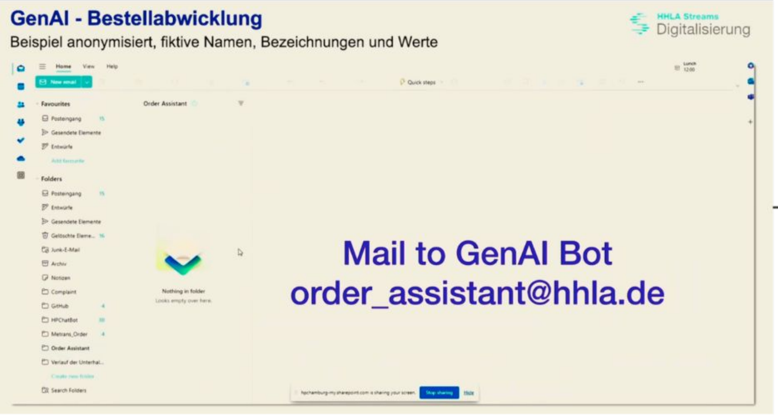
843316232990| |Port Forwarder: Geodis Wilson|Additional Contacts: Name: Mr. David Fila|Phone: +252562253225|

Test of order Assistant

Hallo,

Bijgaand vindt u de gegevens van onze containers.
Een van hen komt per schip: CFR4667615784
De andere heeft het nummer: CANS23098754
En hier is er nog een: TCLU51294562920|.

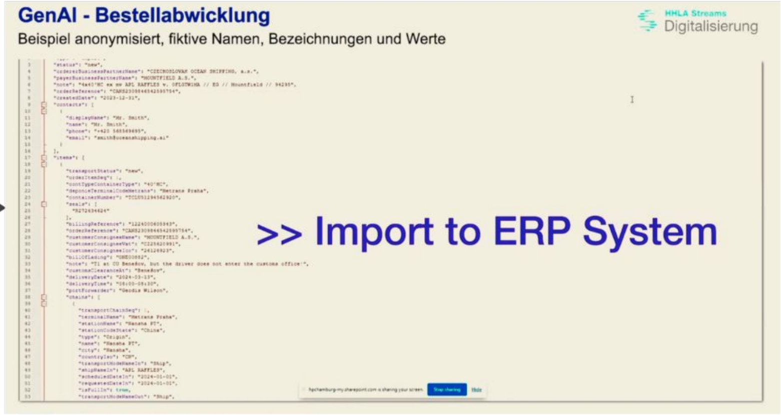
Zonnige groeten



GenAI - Bestellabwicklung

Beispiel anonymisiert, fiktive Namen, Bezeichnungen und Werte

Mail to GenAI Bot
order_assistant@hhla.de



GenAI - Bestellabwicklung

Beispiel anonymisiert, fiktive Namen, Bezeichnungen und Werte

```
  "status": "new",
  "ordererBusinessPartnerName": "CZECHOSLOVAK OCEAN SHIPPING, a.s.",
  "payerBusinessPartnerName": "MOUNTFIELD A.S.",
  "note": "4x40'HC ex mv APL RAFFLES v. 0FLGTW1MA // EG // Mountfield // 94295",
  "orderReference": "CANS2309846542595754",
  "createdDate": "2023-12-31",
  "contacts": [
    {
      "displayName": "Mr. Smith",
      "name": "Mr. Smith",
      "phone": "+420 568569695",
      "email": "smith@oceanshipping.ai"
    }
  ],
  "items": [
    {
      "transportStatus": "new",
      "orderItemSeq": 1,
      "contTypeContainerType": "40'HC",
      "deponieTerminalCodeMetrans": "Metrans Praha",
      "containerNumber": "TCLU51294562920",
      "seals": [
        "R272634624"
      ],
      "billingReference": "1224000605943",
      "orderReference": "CANS2309846542595754",
      "customerConsigneeName": "MOUNTFIELD A.S.",
      "customerConsigneeVat": "CZ25620991",
      "customerConsigneeIco": "26126923",
      "billOfLading": "GNE03882",
      "note": "T1 at CU Benešov, but the driver does not enter the customs office!",
      "customsClearanceAt": "Benešov",
      "deliveryDate": "2024-03-13",
      "deliveryTime": "08:00-08:30",
      "portForwarder": "Geodis Wilson",
      "chains": [
        {
          "transportChainSeq": 1,
          "terminalName": "Metrans Praha",
          "stationName": "Nansha PT",
          "stationCodeState": "China",
          "type": "Origin",
          "name": "Nansha PT",
          "city": "Nansha",
          "countryIso": "CN",
          "transportModeNameIn": "Ship",
          "shipNameIn": "APL RAFFLES",
          "scheduledDateIn": "2024-01-01",
          "requestedDateIn": "2024-01-01",
          "isFullIn": true,
          "transportModeNameOut": "Ship",
```

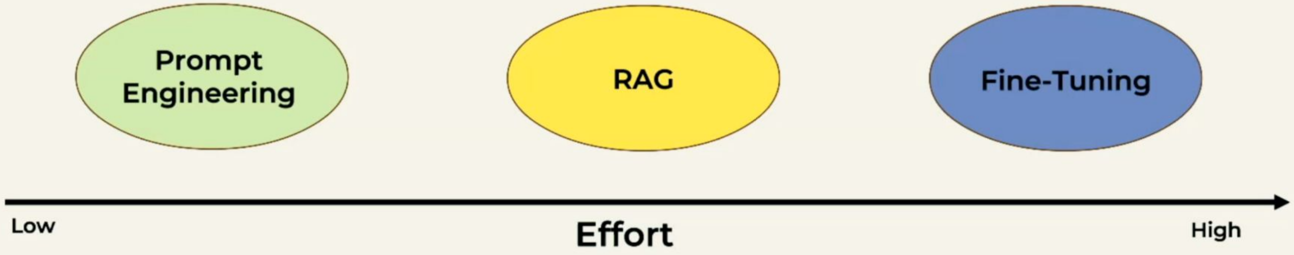
>> Import to ERP System

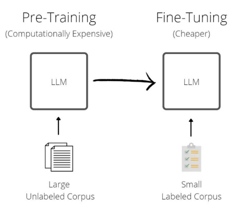
In a Nutshell 🥜

Agentic systems are AI-Systems, that can accomplish

- complex goals 🎯
- in complex environments 🌏
- with limited direct supervision 👀

Improving LLMs 🤖


Prompt Engineering
RAG
Fine-Tuning
Low
Effort
High


Pre-Training
(Computationally Expensive)
Fine-Tuning
(Cheaper)
LLM
LLM
Large Unlabeled Corpus
Small Labeled Corpus