Build the best knowledge retriever for your LLM.

Jettro Coenradie

Daniël Spee

J·FALL
2024

luminis. / part of yuma

#jfall

Daniël Spee

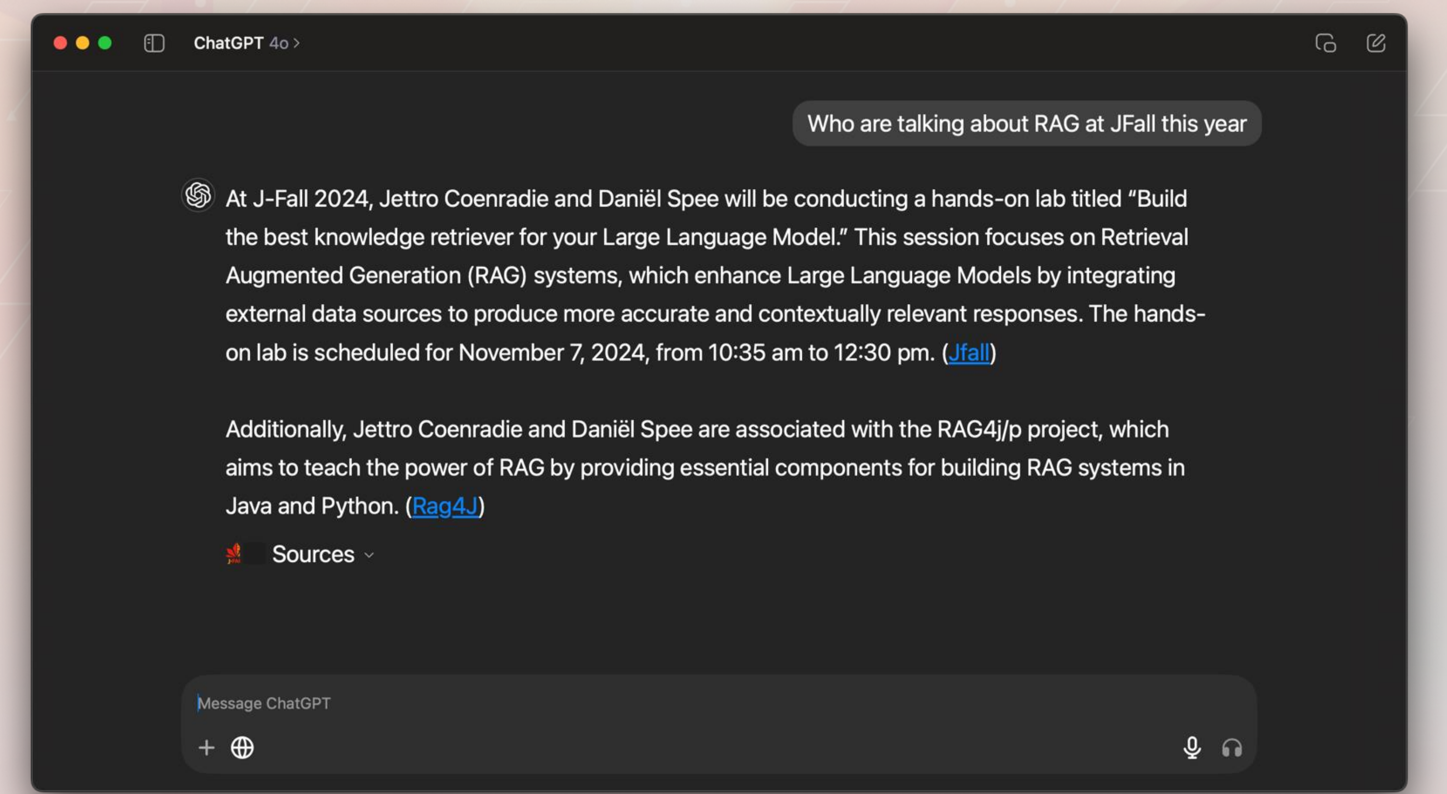
- https://www.linkedin.com/in/dspee/
- https://github.com/Daantie
- BLOG https://www.luminis.eu/expert/daniel-spee/
- daniel.spee@luminis.eu

Jettro Coenradie

Search and Gen AI Evangelist @ Luminis

- https://www.linkedin.com/in/jettro
- https://github.com/jettro
- BLOG https://www.luminis.eu/expert/jettro-coenradie/
- Jettro.coenradie@luminis.eu

Use ChatGPT to answer questions

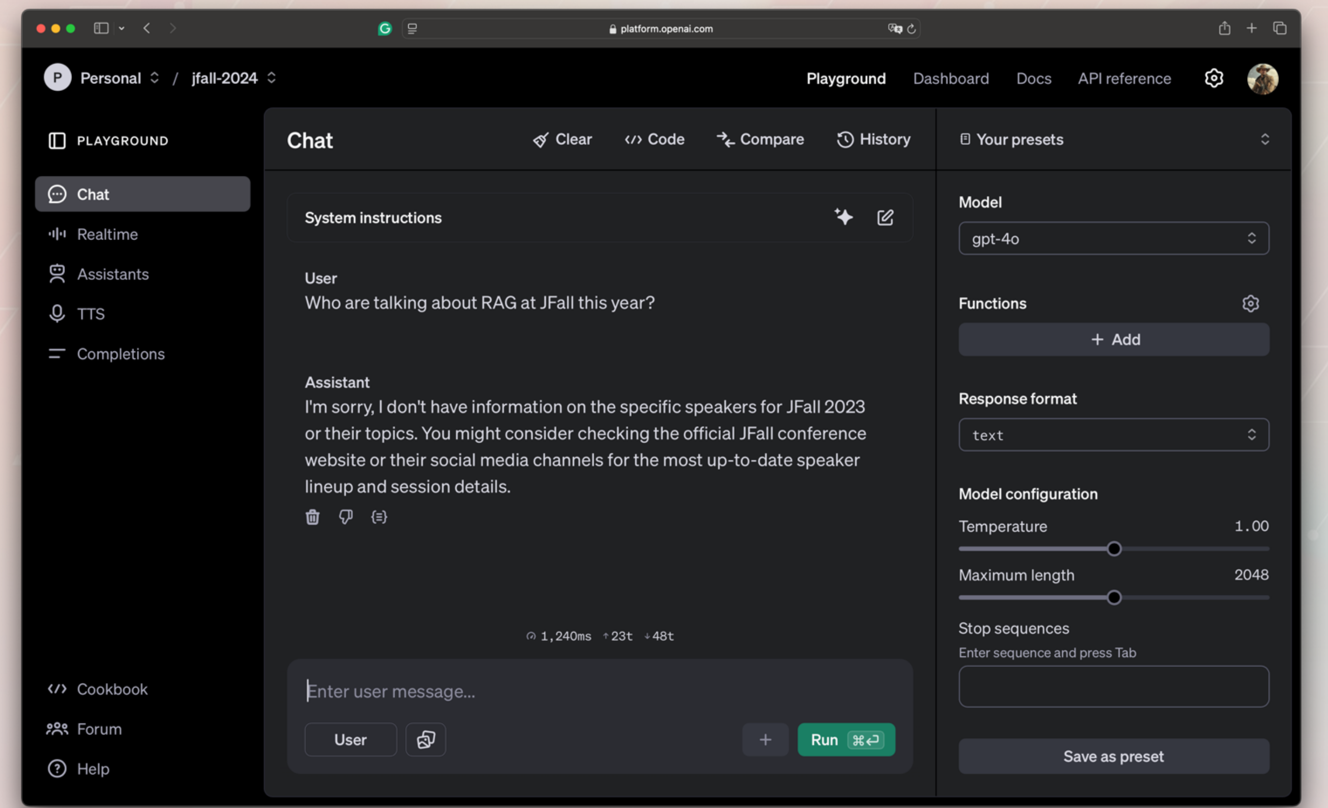


Who are talking about RAG at JFall this year

At J-Fall 2024, Jettro Coenradie and Daniël Spee will be conducting a hands-on lab titled “Build the best knowledge retriever for your Large Language Model.” This session focuses on Retrieval Augmented Generation (RAG) systems, which enhance Large Language Models by integrating external data sources to produce more accurate and contextually relevant responses. The hands-on lab is scheduled for November 7, 2024, from 10:35 am to 12:30 pm. (Jfall)

Additionally, Jettro Coenradie and Daniël Spee are associated with the RAG4j/p project, which aims to teach the power of RAG by providing essential components for building RAG systems in Java and Python. (Rag4J)

Use GPT-4o for the same question



User
Who are talking about RAG at JFall this year?

Assistant
I'm sorry, I don't have information on the specific speakers for JFall 2023 or their topics. You might consider checking the official JFall conference website or their social media channels for the most up-to-date speaker lineup and session details.

1,240ms ↑23t ↓48t

Model: gpt-4o

Response format: text

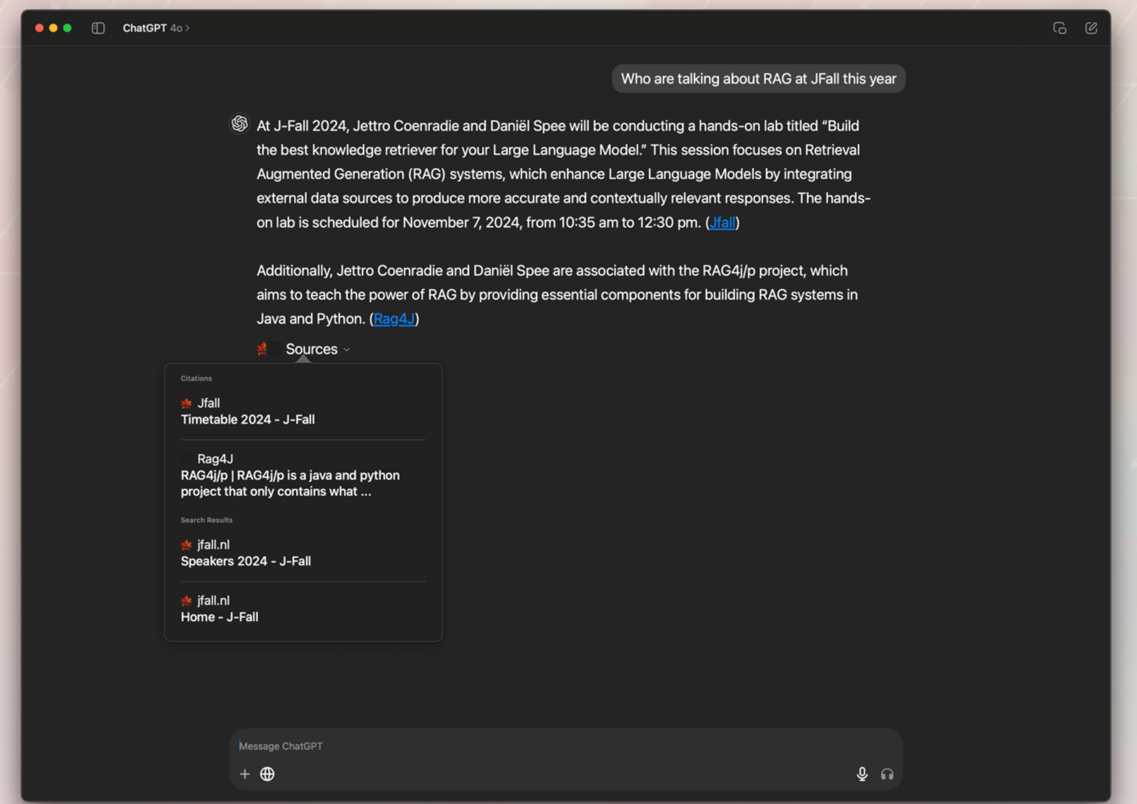
Model configuration

Temperature 1.00

Maximum length 2048

Stop sequences

Check the sources for ChatGPT



Who are talking about RAG at JFall this year

At J-Fall 2024, Jettro Coenradie and Daniël Spee will be conducting a hands-on lab titled "Build the best knowledge retriever for your Large Language Model." This session focuses on Retrieval Augmented Generation (RAG) systems, which enhance Large Language Models by integrating external data sources to produce more accurate and contextually relevant responses. The hands-on lab is scheduled for November 7, 2024, from 10:35 am to 12:30 pm. (Jfall)

Additionally, Jettro Coenradie and Daniël Spee are associated with the RAG4j/p project, which aims to teach the power of RAG by providing essential components for building RAG systems in Java and Python. (Rag4J)

Sources

Citations

Jfall
Timetable 2024 - J-Fall

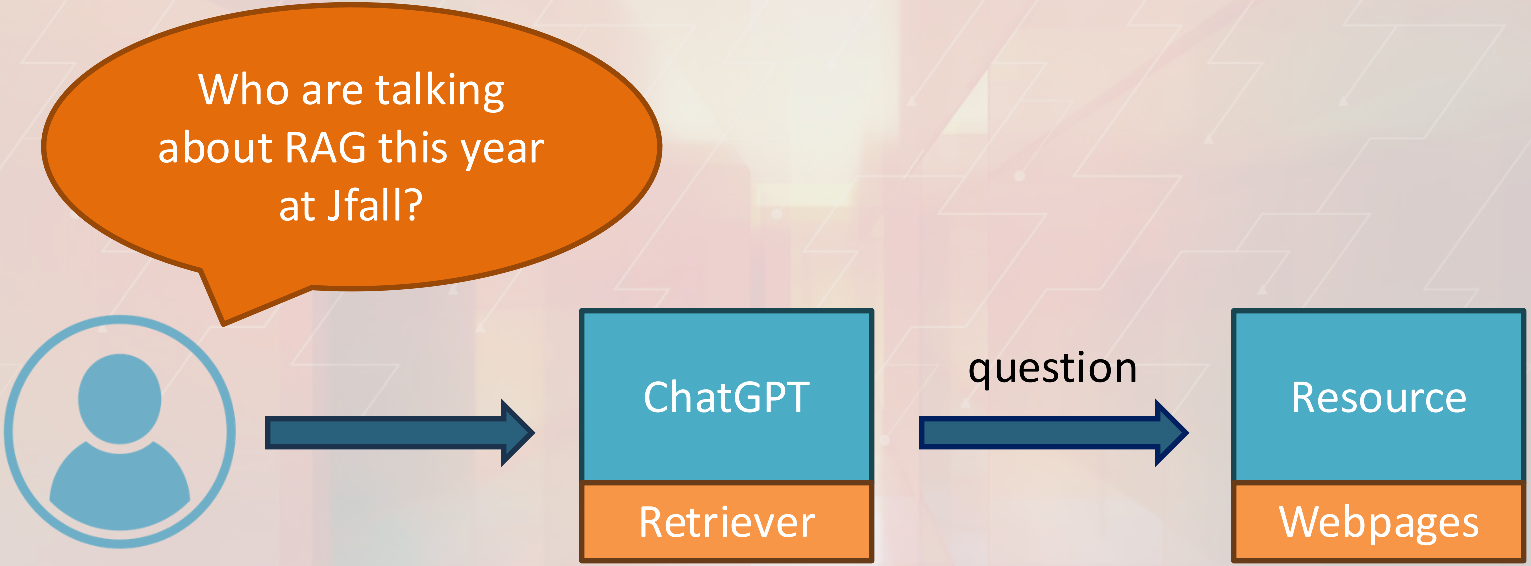
Rag4J
RAG4j/p | RAG4j/p is a java and python project that only contains what ...

Search Results

jfall.nl
Speakers 2024 - J-Fall

jfall.nl
Home - J-Fall

Retrieval Augmented Generation (RAG)



Who are talking about RAG this year at Jfall?

ChatGPT

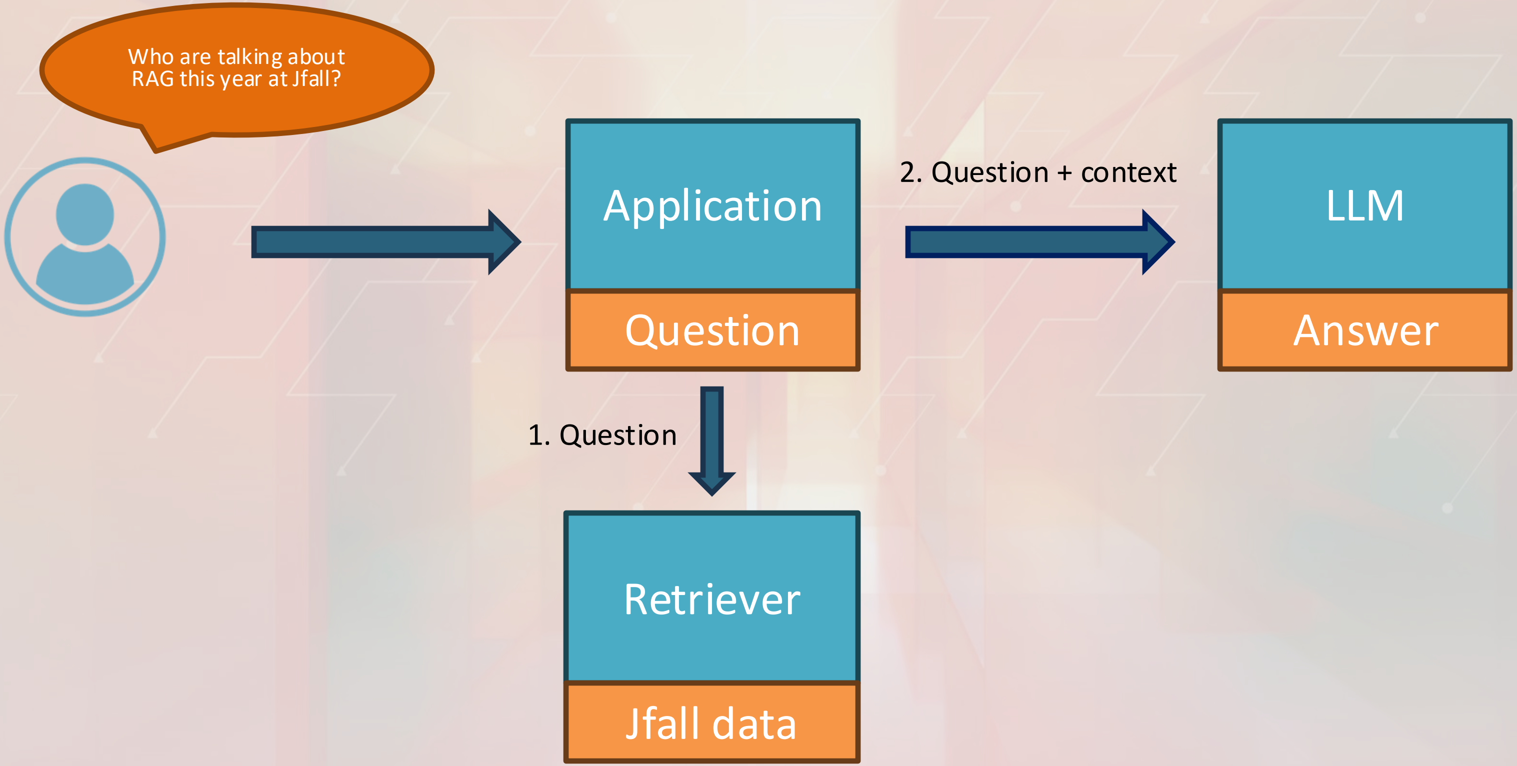
Retriever

question

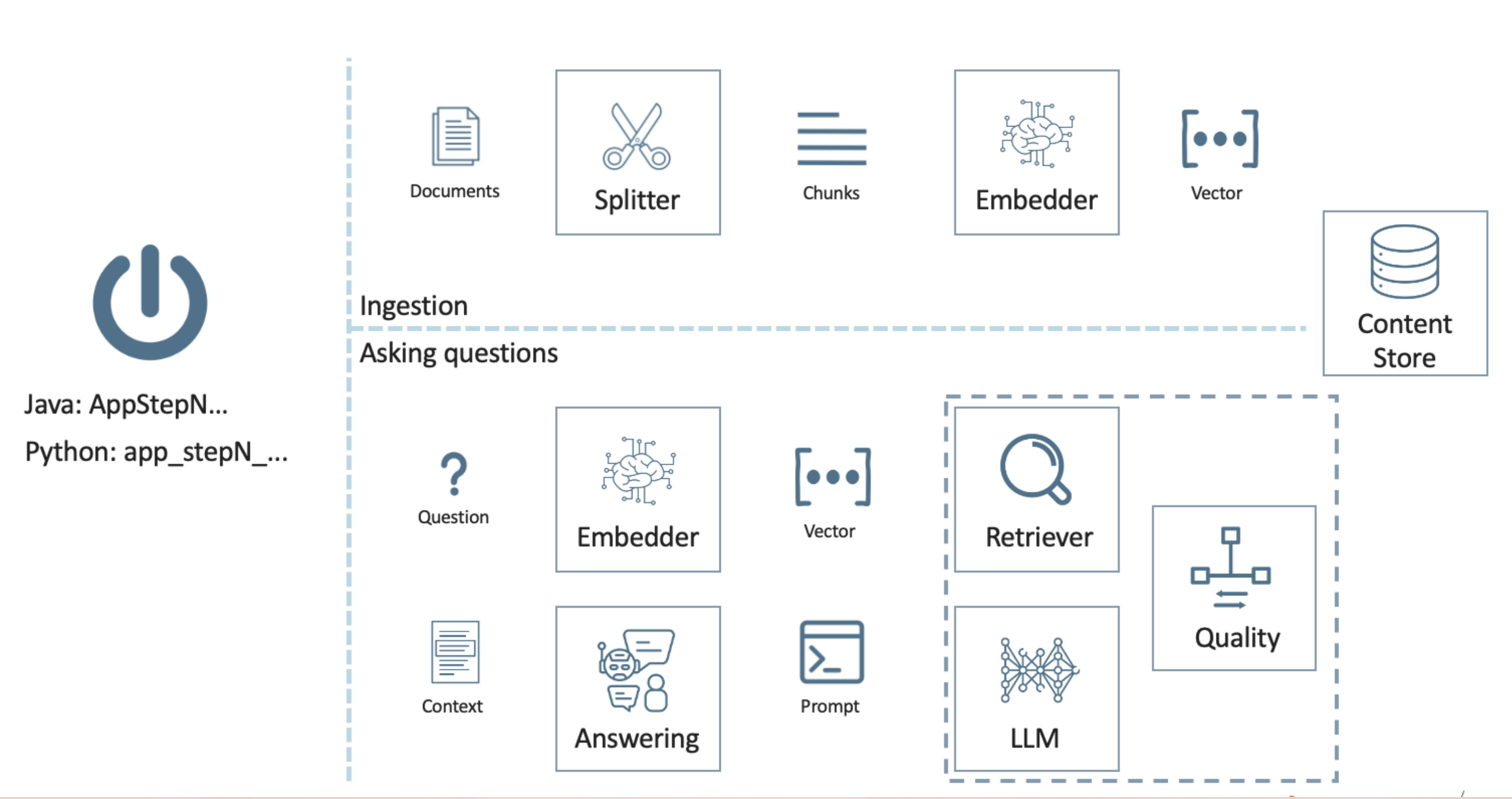
Resource

Webpages

Retrieval Augmented Generation (RAG)



Retrieval Augmented Generation (RAG)



Workshop agenda

- Setting up your environment
- Ingestion
- Retrieval strategy
- Retrieval quality
- Overall quality

Ingestion

Splitting tekst into chunks

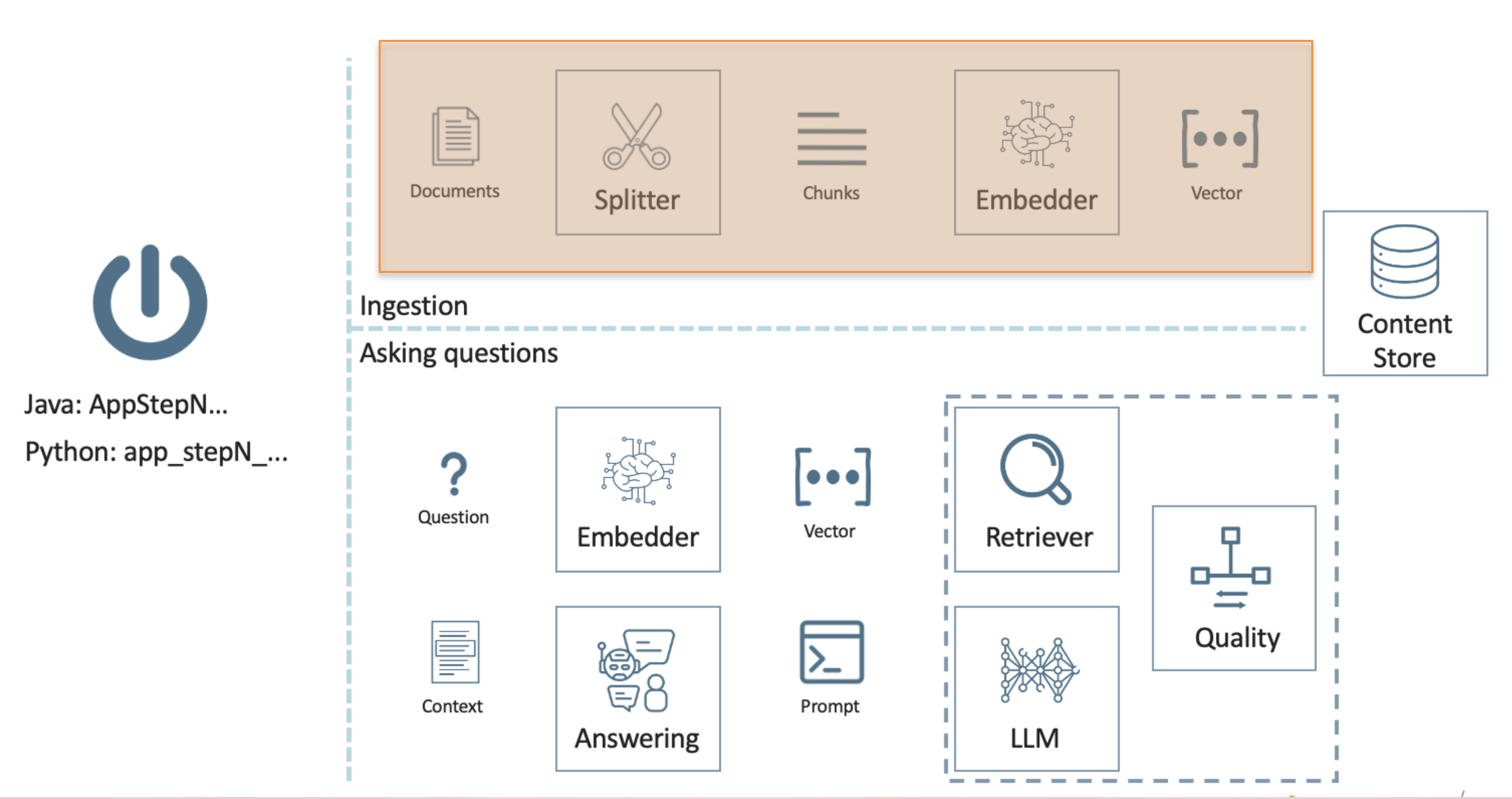
Why is splitting necessary?

Impact of different splitters

Examples splitters

DALL E - visualize a splitter that splits text into chunks

Retrieval Augmented Generation (RAG)



Speaker • Title

https://rag4j.org
Workshop JFall 2024

https://github.com/RAG4J/rag4j-jfall
https://github.com/RAG4J/rag4p-jfall

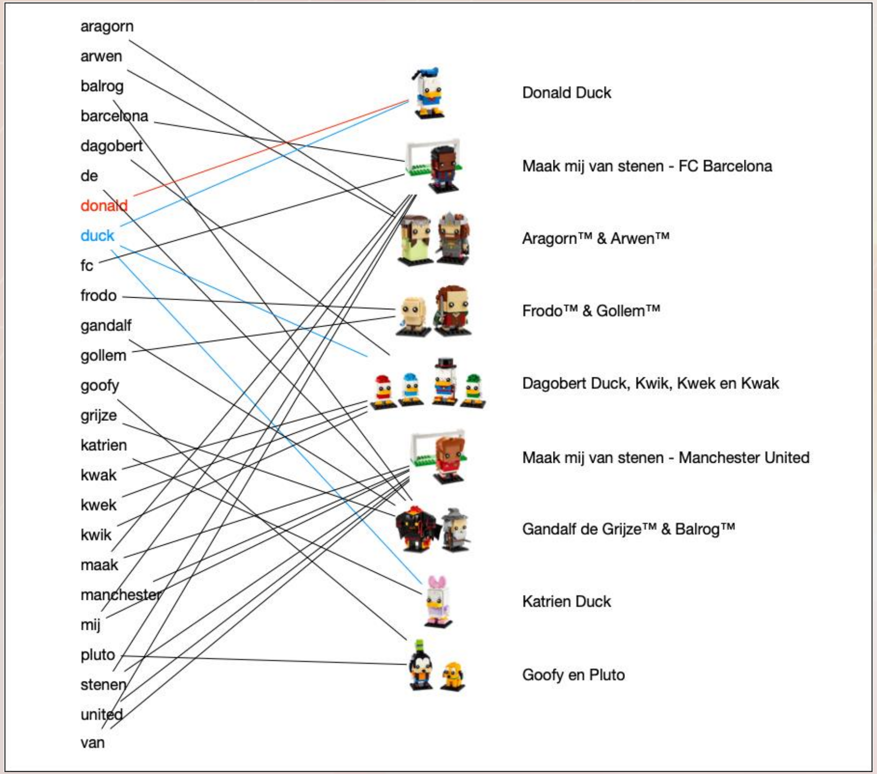
secret_key=
secret-jfallconf-2024-dont-share

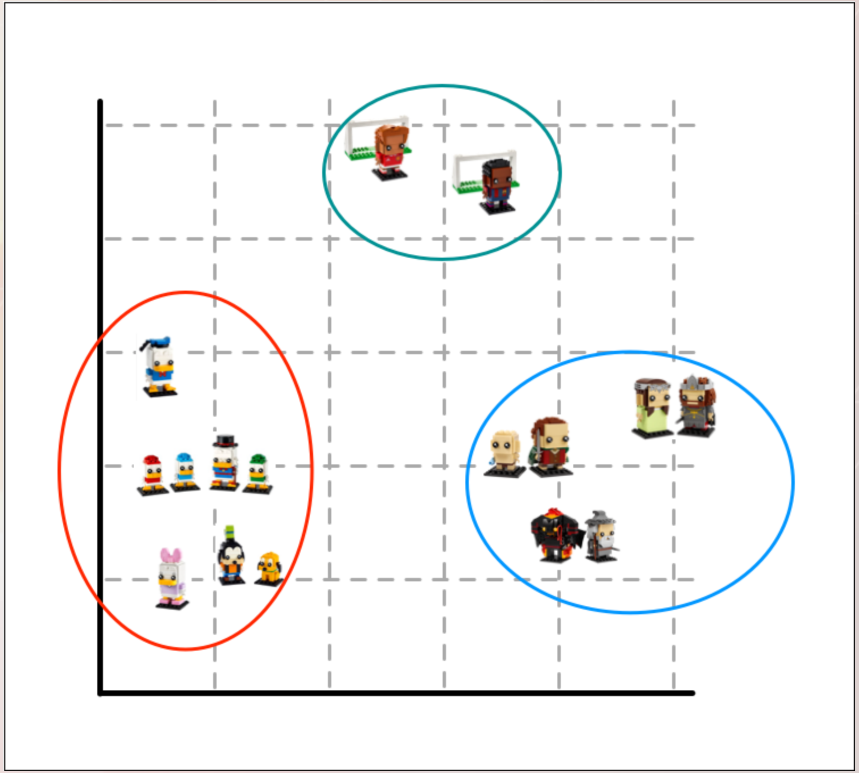
```
J: AppStep1Ingestion
P: app_step1_ingestion
```

```
Branch: assignment or
assignment_ollama
```

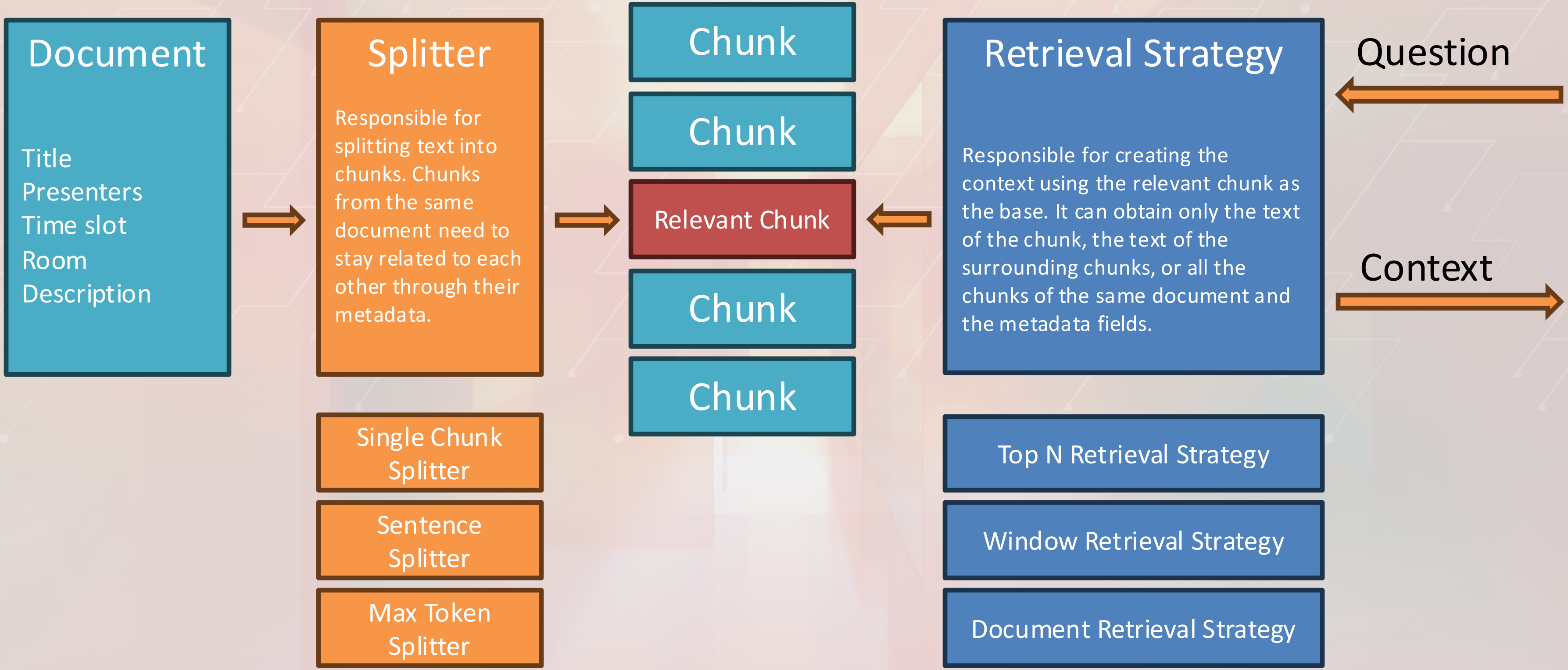
Retrieving

Lexical versus Semantic Search





Retrieving the context



Document

Title
Presenters
Time slot
Room
Description

Splitter

Responsible for splitting text into chunks. Chunks from the same document need to stay related to each other through their metadata.

Single Chunk Splitter

Sentence Splitter

Max Token Splitter

Chunk

Chunk

Relevant Chunk

Chunk

Chunk

Retrieval Strategy

Responsible for creating the context using the relevant chunk as the base. It can obtain only the text of the chunk, the text of the surrounding chunks, or all the chunks of the same document and the metadata fields.

Top N Retrieval Strategy

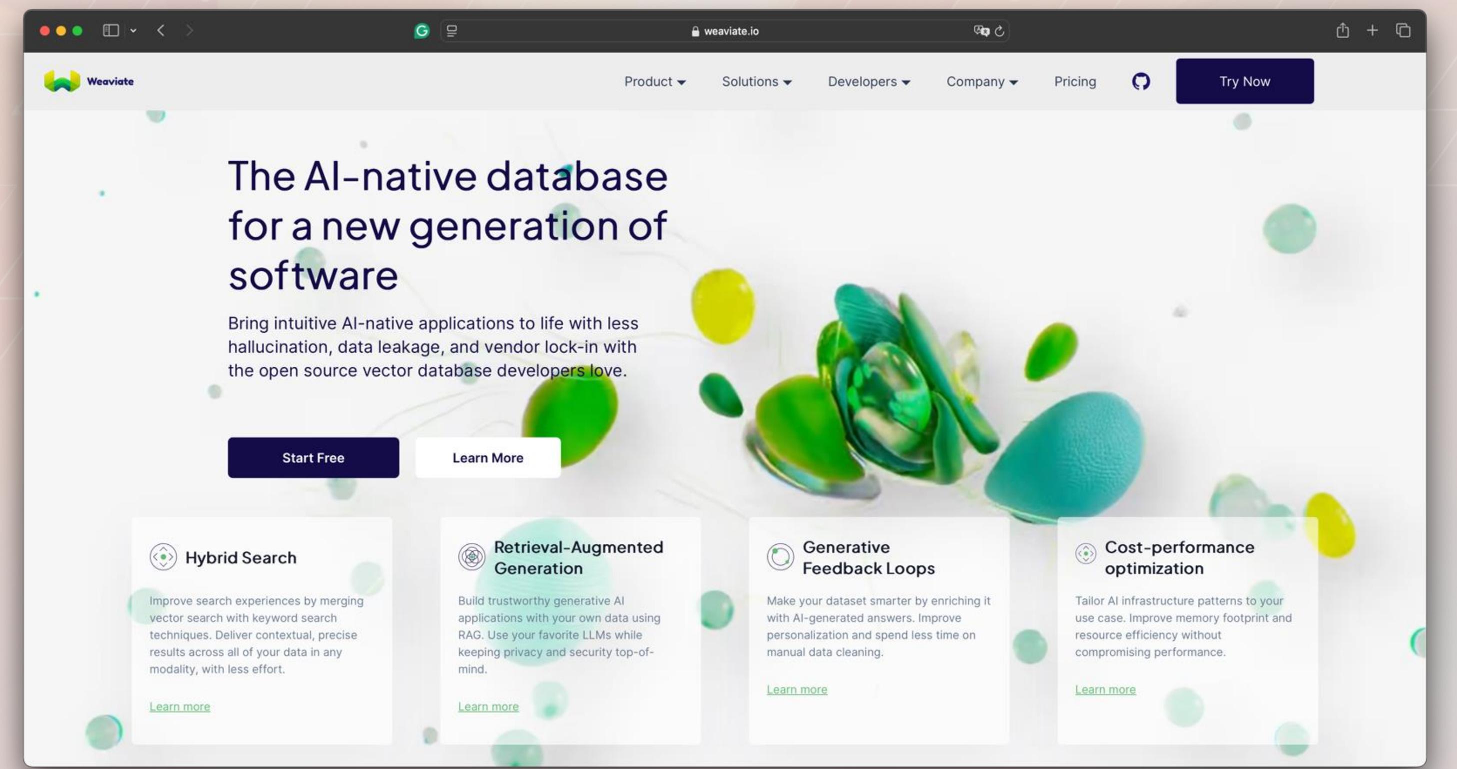
Window Retrieval Strategy

Document Retrieval Strategy

Question

Context

Weaviate ~ Vector store++



The AI-native database for a new generation of software

Bring intuitive AI-native applications to life with less hallucination, data leakage, and vendor lock-in with the open source vector database developers love.

Hybrid Search

Improve search experiences by merging vector search with keyword search techniques. Deliver contextual, precise results across all of your data in any modality, with less effort.

Learn more

Retrieval-Augmented Generation

Build trustworthy generative AI applications with your own data using RAG. Use your favorite LLMs while keeping privacy and security top-of-mind.

Learn more

Generative Feedback Loops

Make your dataset smarter by enriching it with AI-generated answers. Improve personalization and spend less time on manual data cleaning.

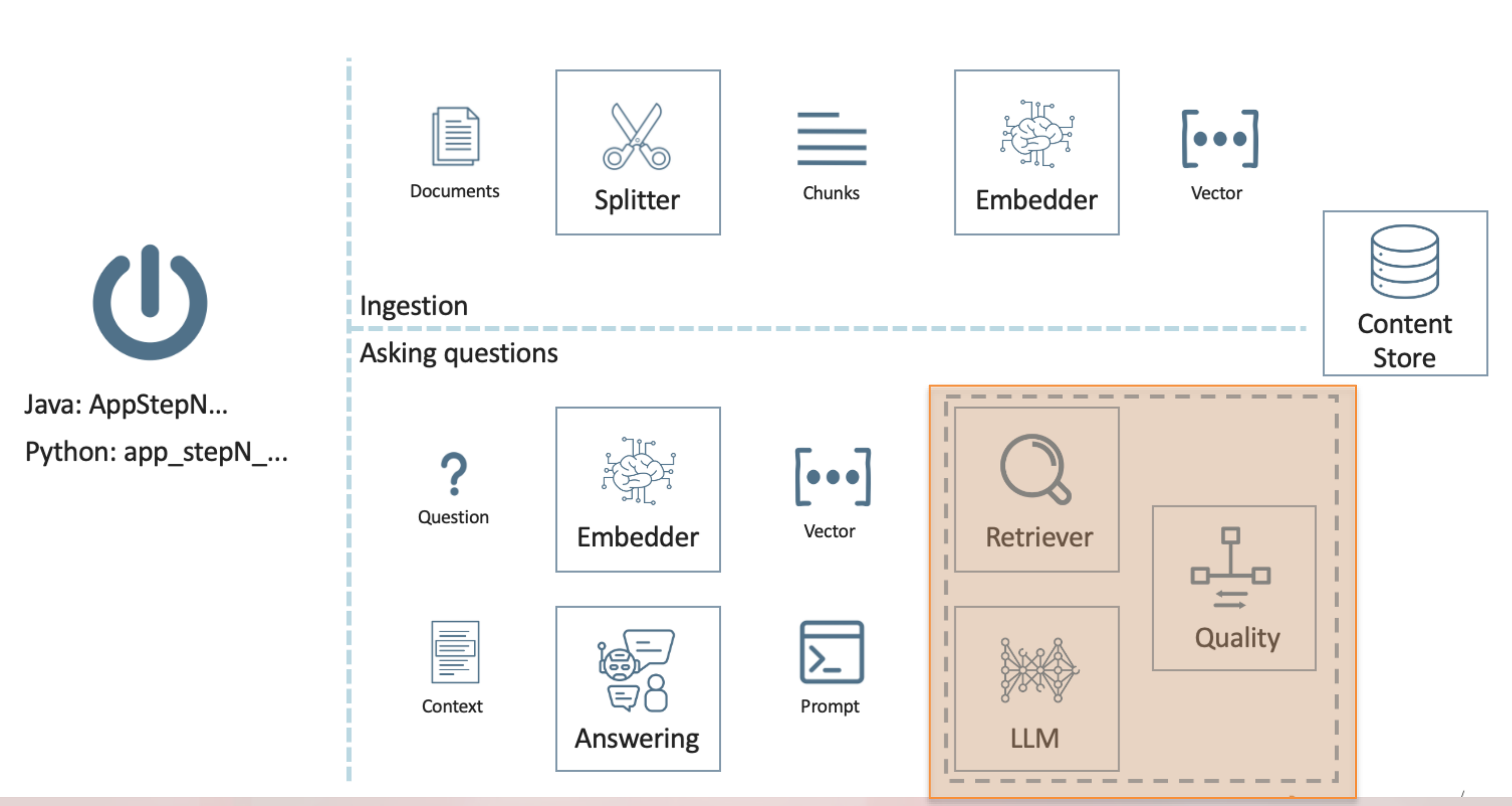
Learn more

Cost-performance optimization

Tailor AI infrastructure patterns to your use case. Improve memory footprint and resource efficiency without compromising performance.

Learn more

Retrieval Augmented Generation (RAG)



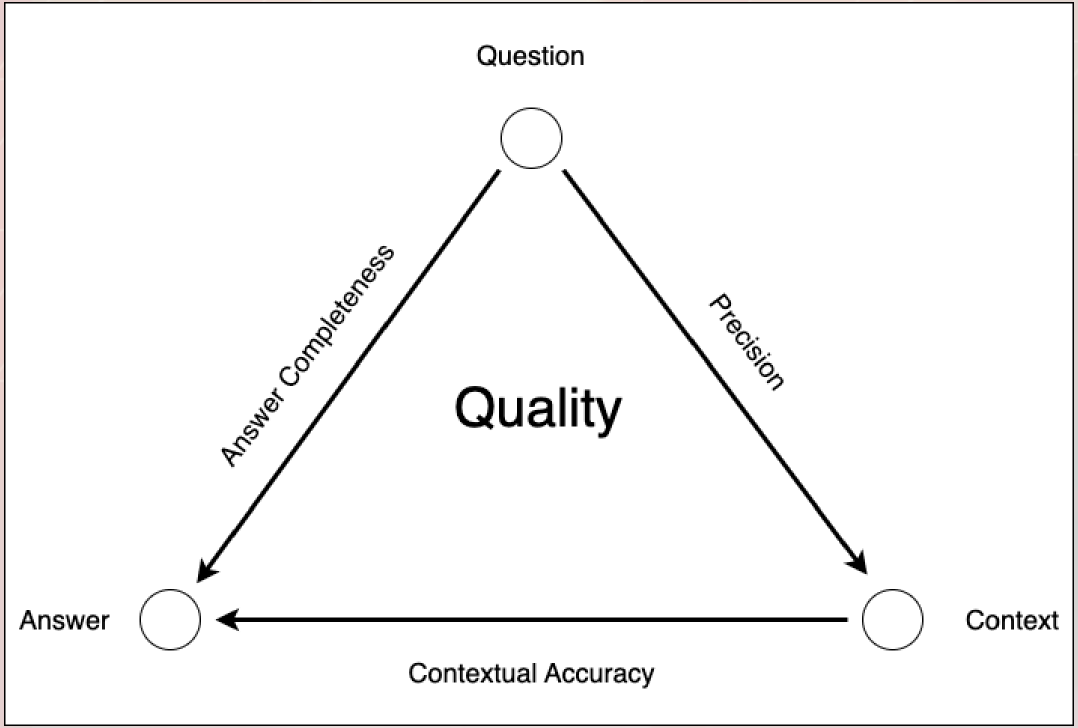
Speaker • Title

secret_key=
secret-jfallconf-2024-dont-share

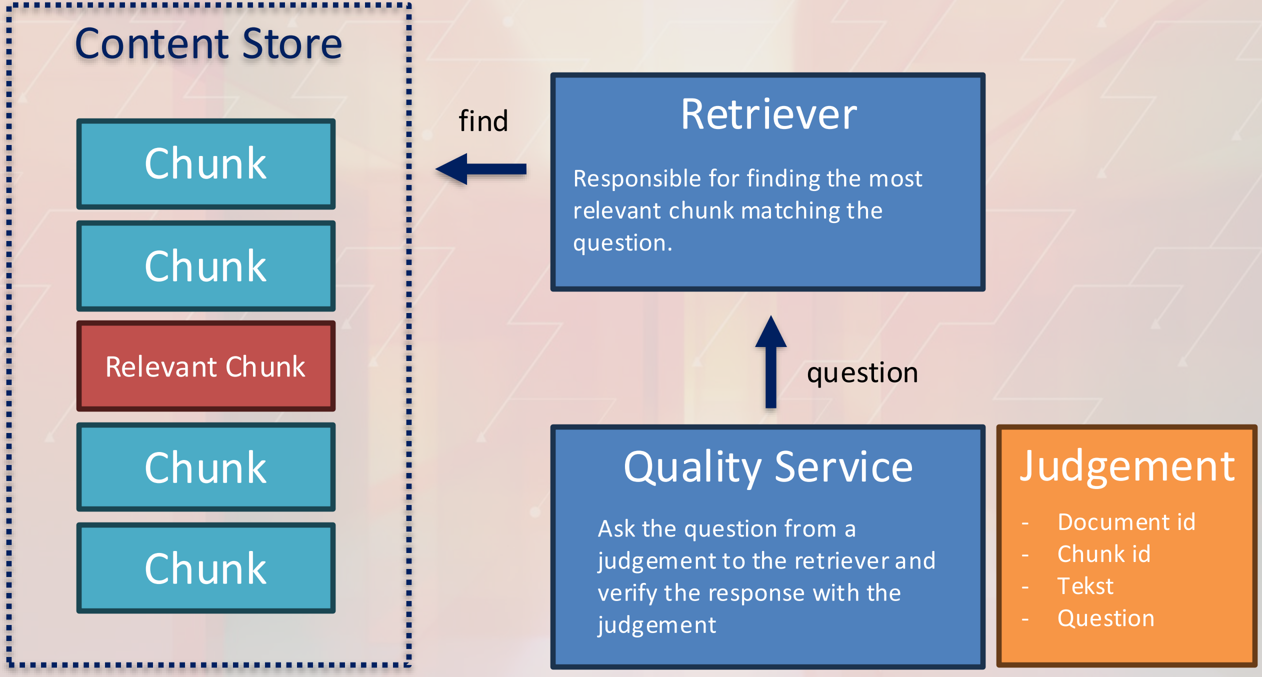
```
J: AppStep2Retrieving
P: app_step2_retrieving
```

Retrieval Quality

Quality of RAG



Retrieval Quality - Precision



Content Store

Chunk

Chunk

Relevant Chunk

Chunk

Chunk

find

Retriever

Responsible for finding the most relevant chunk matching the question.

question

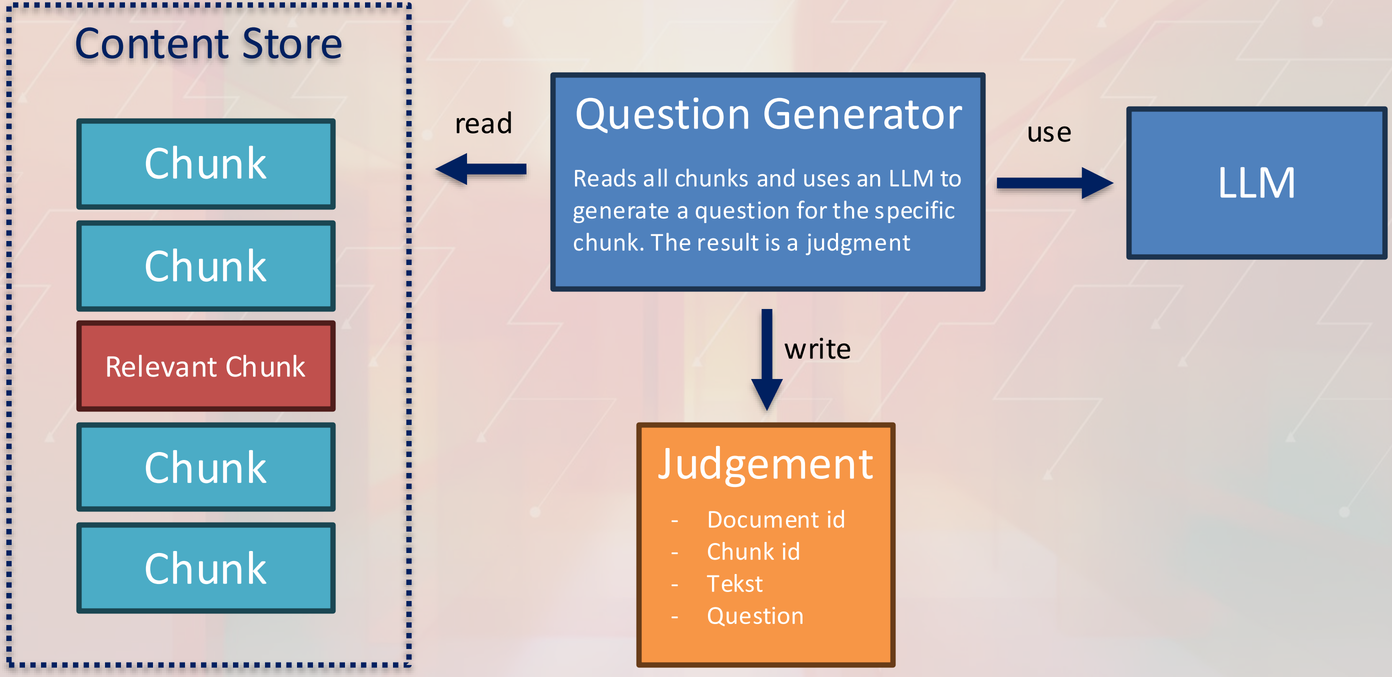
Quality Service

Ask the question from a judgement to the retriever and verify the response with the judgement

Judgement

- Document id
- Chunk id
- Tekst
- Question

Generate the Judgement list



Content Store

Chunk

Chunk

Relevant Chunk

Chunk

Chunk

read

Question Generator

Reads all chunks and uses an LLM to generate a question for the specific chunk. The result is a judgment

use

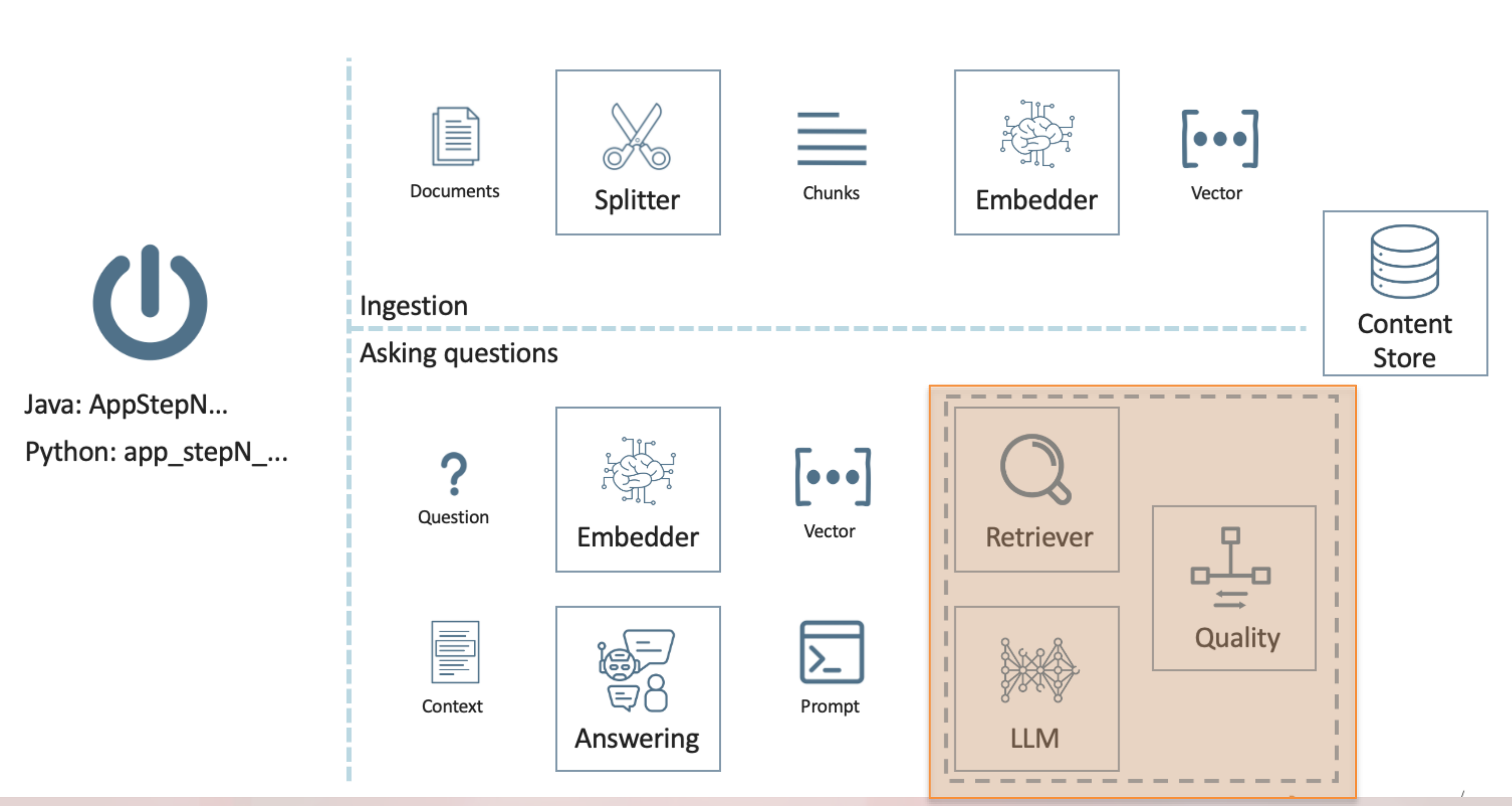
LLM

write

Judgement

- Document id
- Chunk id
- Tekst
- Question

Retrieval Augmented Generation (RAG)



Java: AppStepN...

Python: app_stepN_...

Ingestion

Documents

Splitter

Chunks

Embedder

Vector

Content Store

Asking questions

Question

Embedder

Vector

Retriever

Quality

Context

Answering

Prompt

LLM

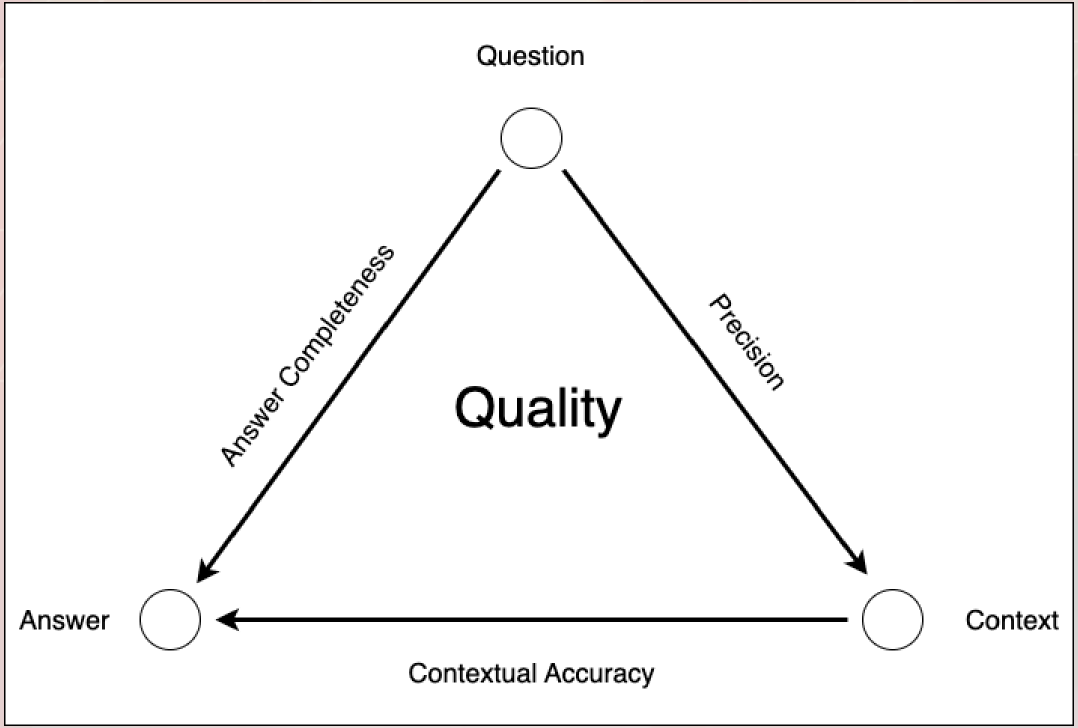
Speaker • Title

secret_key=
secret-jfallconf-2024-dont-share

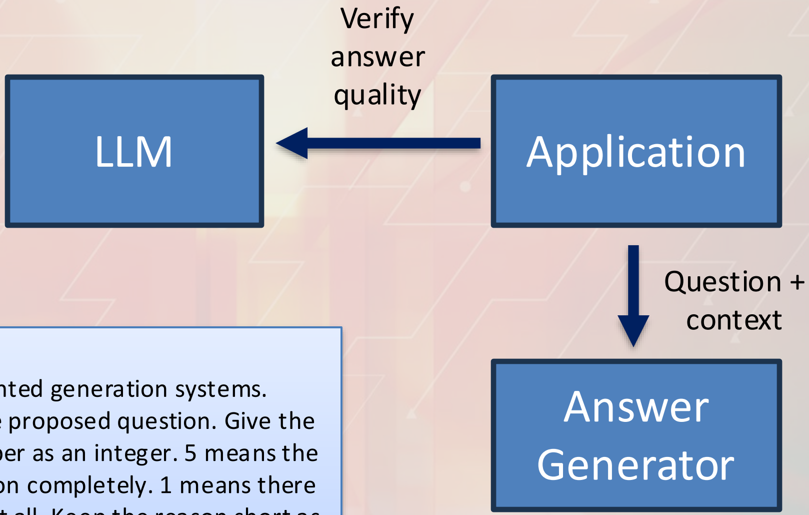
```
J: AppStep3RetrievalQuality
P: app_step3_retrieval_quality
```

Answer Quality

Quality of RAG



Prompt for quality of answer



You are a quality assistant verifying retrieval augmented generation systems. Your task is to verify a generated answer against the proposed question. Give the answer a score between 1 and 5 and keep the number as an integer. 5 means the answer contains the answer to the proposed question completely. 1 means there is no match between the answer and the question at all. Keep the reason short as in maximum 2 sentences. The question provided after 'question:'. The answer after 'answer:'. Write your answers in json format of:
{{"score": "score", "reason": "reason"}}
An example:
{{"score": 3, "reason": "The answer is correct but some details are missing."}}

Speaker • Title

secret_key=
secret-jfallconf-2024-dont-share

```
J: AppStep4AnswerQuality
P: app_step4_answer_quality
```

Thanks for your attention

Please rate our session in the J-FALL app

J-FALL 2024

𝕏 #jfall

Speaker • Title

Text